

Empirical Perspectives on Learning at Work

by

by **Jan Meyer**

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

in Electrical Engineering

Specialisation in Quality Engineering

Approved by the Thesis Committee:

Supervisor:	Prof. Dr. Werner Bergholz	Professor of Electrical Engineering Jacobs University, Bremen
Examiner:	Prof. Dr. Christian Roßnagel	Professor of Organizational Behavior Jacobs University, Bremen
Examiner:	Prof. Dr. Sven Voelpel	Professor of Business Administration Jacobs University, Bremen
Examiner:	Prof. Dr. Adalbert FX Wilhelm	Professor of Statistics Jacobs University, Bremen
Examiner:	Prof. Dr. Utz Schäffer	Professor of Business Administration Otto Beisheim School of Management (WHU), Vallendar

Date of Defense: May 28th, 2010

School of Engineering and Science

Empirical Perspectives on Learning at Work

by Jan Meyer
May 25, 2011, Final Publication Version

Abstract

In today's knowledge economy, many organizations are trying to manage their knowledge base in order to gain and maintain a competitive edge. Yet knowledge in peoples' heads can hardly be managed *directly*, since knowledge transfer involves an active and individual learning effort. Nevertheless, organizations can support learning *indirectly*. Therefore this study approaches the challenge of knowledge-intensive work from another perspective: How can organizations support informal and individual on-the-job learning?

To obtain a ranking of the most important organizational factors that support learning, a fully structured and dynamic survey at a German shipyard was used to measure the intensity of learning and potentially relevant driving or inhibiting factors in different working environments. Due to the stochastic nature of the learning process and the large number of variables, it was necessary to develop a new algorithm, called BOGER, for the statistical analysis, which features automatic and robust non-linear model selection.

Interdisciplinary insights from a wide field of literature as well as the empirical ranking were condensed into the newly developed PIA-model. "PIA" is the acronym for: 'perspective taking, integration, action' – see figure 2.1 on page 31. Some of the illustrated mechanisms explain the already effective application of state-of-the-art industrial practice models such as the Toyota Production System or the EFQM model, both of which emphasize organizational learning. Thus, this study provides a deepened understanding of the most important organizational levers influencing learning for their application and adaptation to new industrial contexts.

For a 4-page summary see section 1.1 on page 13.

Subject Terms / Keywords: Workplace Learning, On-The-Job Learning, Knowledge Management, PIA-Model, Perspective Taking, Organizational Learning, BOGER Algorithm, Algorithmic Statistical Modelling, Learning Index

Acknowledgments

First and foremost I want to thank my advisor Prof. Werner Bergholz. He once told me that a doctoral advisor should be like a sparring partner to the Ph.D. student – and he truly was to me. His ability to fuse his experience from science and industry, provided me with valuable feedback that allowed me to improve my work in numerous ways. I appreciate all his contributions of time and ideas to make my Ph.D. experience productive and stimulating. At the same time, he gave me enough latitude to leave my own mark on my work and follow my purpose.

My special thanks also belong to the thesis committee members: Prof. Christian Roßnagel, Prof. Utz Schäffer, Prof. Sven Voelpel and Prof. Adalbert Wilhelm. With all of them I had very inspiring discussions and received good feedback that allowed me to draw valuable insights from all their different disciplines.

To integrate insights and methods from other disciplines, required me to learn about these disciplines first. My fellow Ph.D. students from Jacobs University, most notably Anette Eva Fasang, Andries Oeberst, David Richter, Polina Isichenko, Sara Geerdes, Siegmund Otto, Youlia Spivak and Zheng Han, in addition to my wife Elisabeth, helped me learn to approach my research in different ways and with different methods.

Last but not least, I thank my parents for their sustained support during this extensive project.

Contents

1. Introduction	13
1.1. Executive Summary	13
1.2. Introduction and Motivation	17
2. Theory of Learning	21
2.1. Summary of the Theory Chapter	22
2.2. Multiple Perspectives on Learning	23
2.3. Literature Findings Integrated into the PIA-Model	28
2.3.1. Active Learning	28
2.3.2. Perspective Taking	30
2.3.3. Integrating information	34
2.3.4. Prior / Background Knowledge and Perception	41
2.3.5. Language for Thought and Diversity in Discussions	43
2.3.6. Iterative Learning modelled with Feedback Loops	47
2.3.7. Searching for Information	51
2.3.8. Tacit Knowledge and Implicit Learning	52
2.4. Using the PIA-Model to Explain Industrial Practice Models	57
2.4.1. Commonalities of Industrial Practice Models	58
2.4.2. Project Management	58
2.4.3. The Toyota Production System	60
2.4.4. EFQM	62
2.4.5. Link to the PIA-Model	66
2.5. Alternative Perspectives Covered in Literature	68
2.5.1. Knowledge as Object rather than a Personal Skill	69
2.5.2. Identifying Application Relevant Knowledge in Maps	73
2.5.3. True Knowledge vs. Diversity of Perspectives	74
2.5.4. Valuation of Knowledge	75
2.5.5. Importance of Knowledge Definitions	76
2.5.6. Summary – Alternative Perspectives	76
2.6. Assumed Perspective & Definitions	77
2.6.1. Knowledge Definition	77
2.6.2. Organizational Learning Definition	78
2.6.3. Individual Learning Definition	79
2.7. Research Gap and Research Question	80
2.7.1. Research Gap	80
2.7.2. Research Question	80
3. Theory of Science & Methodology	83
3.1. Scientific Paradigma	84
3.1.1. Methodology as Quality Standard for Methods	84

3.1.2.	Science – a High-Quality Form of Investigation	84
3.1.3.	Basic Assumptions about Reality and the Value of Research	86
3.1.4.	General Value Proposition of this Research	91
3.1.5.	Perspective Validation and Empirical Results	92
3.1.6.	Iteration Improves Research Quality	93
3.1.7.	Methodological Approach	95
3.1.8.	Optimizing Methods – Cost vs. Benefit and Quality	96
3.2.	Choice of Methods for this Study	97
4.	Statistical Theory on Algorithmic Modelling	101
4.1.	The Theory regarding Model Selection and Fitting	103
4.1.1.	The Process of Statistical Inference	103
4.1.2.	Theory-based vs. Algorithmic Modeling Approaches	105
4.1.3.	Overall Aim of Statistical Inference	105
4.1.4.	Modeling Stochastic Processes	108
4.1.5.	Automation Limits of Statistical Analysis & Causality	109
4.1.6.	No Principle Difference between Model Selection and Fitting	111
4.1.7.	Model Selection Criteria: Model Fit vs. Model Error	114
4.1.8.	Variable Selection vs. Model Selection	116
4.2.	Practical Challenges with Algorithmic Model Selection	119
4.2.1.	Measures for Model Fit with Reality (R^2)	119
4.2.2.	Biased Model Fit Estimation and Overfitting	123
4.2.3.	Challenges in Model Selection	128
4.2.4.	Estimates for Predictive Error	129
4.2.5.	Examples of Robust Algorithms and their Properties	134
5.	Quantitative Stage - The Survey Instrument	137
5.1.	Overview	138
5.2.	Pre-Survey Qualitative Pilot Interviews	139
5.3.	Covered Constructs in the Survey	141
5.4.	Quantifying On-The-Job Learning – Learning Index	145
5.4.1.	Learning Index Survey Tool	146
5.4.2.	Learning Index Definition	150
5.5.	Survey Pilots	152
5.6.	Survey Design Goals	153
5.7.	Survey Algorithm	153
5.8.	Survey Conduction and Resulting Sample	158
5.9.	Actual Performance of the Interactive Survey	159
5.10.	Data Pre-Processing	161
5.11.	Validity Investigation of the Learning Index	162
5.12.	Properties of the Data Set	163
5.12.1.	Multi-Variate Relationships / Collinearity / Correlations	163
5.12.2.	Noise	166
5.12.3.	Non-Linear Relationships	168

6. Statistical Analysis with BOGER	171
6.1. Performance of Existing Algorithms	172
6.1.1. Choice of Existing Algorithms	172
6.1.2. Comparison Method	174
6.1.3. Comparison Results – Existing Algorithms	175
6.1.4. The Process of the Analysis	177
6.2. Design of the BOGER Algorithm	179
6.2.1. Design Goals	179
6.2.2. The Mathematical Model	180
6.2.3. Genetic (Non-linear) Function Fit Algorithm	183
6.2.4. Overview on Model Building and Robustness Testing	185
6.2.5. The Fully Automatic Screening Stage	187
6.2.6. The Final Stage of Model Selection	192
6.2.7. Robust Model Fit Estimation in BOGER	195
6.2.8. Implementation in \mathbb{R}	197
6.3. Design Requirement Validation & Performance of BOGER	198
6.3.1. Model Selection Progress	198
6.3.2. Model Fit & Predictive Power Estimate	199
6.3.3. Design Requirement Validation Summary	202
7. Statistical Results and their Interpretation	205
7.1. Result Interpretation Procedures	206
7.1.1. Interpretation Criterion – Variable Importance	206
7.1.2. Interpreting Interactions	212
7.1.3. Model Shape Graphics by Variable	213
7.2. Overall Statistical Results	217
7.2.1. Survey Questions of the Model Variables	217
7.2.2. Variable Rankings and fitted Model Parameters	219
7.2.3. BOGER Model Parameter Results	224
7.3. By Variable Results and Interpretation	226
7.3.1. Learning Strategy Profile	226
7.3.2. Leadership Effect	231
7.3.3. Personal Interest	238
7.3.4. Personal Working History Variable Group	240
7.3.5. Learning Barriers Variable Group	243
7.3.6. Epistemological Beliefs about Learning	247
7.3.7. Task Type Variable Group	250
7.3.8. Description Detail-Level of Procedures	252
7.3.9. Number of Seminars	254
7.3.10. Job Closure	255
7.3.11. Openness to New Experiences (Big Five)	257
7.3.12. Task Difficulty	257
7.3.13. Fault Culture	258
7.3.14. Surprisingly insignificant Factors (Non-Factors)	260

8. Summary, Implications, Limitations and Future Research	263
8.1. Summary of Research Findings	263
8.1.1. Principal Insights from Literature	263
8.1.2. Results Overview	267
8.2. Practical Implications	270
8.3. Relevance of the Results to Literature	275
8.3.1. Relevance to Organizational Learning	275
8.3.2. Relevance to Industrial Practice Models	276
8.3.3. Relevance to Sense Making, Problem Solving and Knowing	277
8.3.4. Relevance to Statistics	277
8.3.5. Relevance to Knowledge Management	278
8.3.6. Selection of a Few from Many Plausible Explanations	279
8.4. Limitations	280
8.5. Areas for Future Research	282
A. Appendix	287
A.1. Writing Style and Conventions	287
A.2. Searching in Literature	288
A.3. Company Profile – Meyer Werft	289
A.4. Validity Investigation of the Learning Index in Detail	290
A.4.1. Inspection of the Input Data to the Learning Index	291
A.4.2. Distribution of the Learning Index	295
A.4.3. Cross-Validation of the Learning Index with Related Questions	297
A.5. Data Pre-Processing Details	302
A.5.1. Filtering and Outlier Removal	302
A.5.2. Imputation and Missing Value Filtering	303
A.6. Details on BOGER	305
A.6.1. Generating Data Frequency Equalized Bootstrapping Samples	305
A.6.2. Implementation Details of BOGER in \mathbb{R}	307
A.6.3. Flexible Model Fitting - an Interesting Accident	310
A.6.4. Residuals	311
A.6.5. Empirical Robustness of Model Fit Measures	314
References	316
Index	339

List of Figures

1.1. PIA-model	15
2.1. PIA-model	31
2.2. Brunswik's Lens Model	33
2.3. PIA-model with Visualization	36
2.4. Butler and Winne's Self-Regulated Learning (SRL) Model	48
2.5. Gantt Chart Example	59
2.6. The EFQM Excellence Model	63
4.1. Overfitting - A Graphical Example	125
4.2. Overfitting Depending on Data Density and Model Flexibility	127
5.1. Learning Index Tool Flow Chart	147
5.2. Survey Algorithm Flow Chart	156
5.3. Reduction Level Frequencies	160
5.4. Survey Duration Distribution	161
5.5. Learning Index vs. Education Level	163
5.6. Task Branch vs. Education Level	164
5.7. Learning Index by Task Branch	165
5.8. A Noisy Relationship	167
5.9. A Non-linear Relationship	169
6.1. BOGER Algorithm Flow Chart	186
6.2. A Graphical Impression of Model Fit	200
7.1. Interpretation of Interactions	212
7.2. Annotated Model Shape Graphics	215
7.3. Variable Ranking Results	221
7.4. Variable Importance Distribution for the Complete Data	222
7.5. Variable Importance Distribution for the Test Data	223
7.6. BOGER Parameter Values and Parameter Instabilities	225
7.7. Learning Strategy Profile Dimensions Compared	227
7.8. Reading & Discussion Learning Strategy – Model Shape Graphic	229
7.9. Leadership Profile Dimensions Compared	232
7.10. Learning Supportive Leadership Profile – Model Shape Graphic	234
7.11. Personal Interest – Model Shape Graphic	238
7.12. Job History – Model Shape Graphic	241
7.13. Learning Barrier Profiles Compared	244
7.14. Approach Not Clear & Expert Not Found or Available – Model Shape Graphic	245
7.15. Epistemological Beliefs about Learning – Model Shape Graphic	249
7.16. Task Type – Model Shape Graphic	251

List of Figures

7.17. Procedural Description Detail – Model Shape Graphic	253
7.18. Learning Index vs. Routine Level	254
7.19. Number of Seminars – Model Shape Graphic	255
7.20. Job Closure – Model Shape Graphic	256
7.21. Personality – Openness to New Experiences – Model Shape Graphic	258
7.22. Task Difficulty – Model Shape Graphic	259
7.23. Fault Culture – Model Shape Graphic	259
8.1. PIA-model repeated for convenience	264
8.2. Implications Summarized in a Mindmap	270
A.1. Frequencies of Learning Situations Levels	291
A.2. Frequencies of Learning Importance Levels by Workstep	292
A.3. Frequencies of Learning Importance Levels by Learning Situation	293
A.4. Frequencies of Learning Usefulness Levels by Workstep	294
A.5. Frequencies of Learning Usefulness Levels by Learning Situation	295
A.6. Relationship between Mean Learning Usefulness and Learning Importance	296
A.7. Distribution of the Learning Index	296
A.8. Distribution of Learning Index vs. General Learning Impression	298
A.9. Distribution of Learning Index vs. Post-Task Self-Efficacy for each Person and Workstep	299
A.10. Distribution of Learning Index vs. Summed Post-Task Self-Efficacy	300
A.11. Distribution of Learning Index vs. Post-Task Self-Efficacy for each Person and Workstep	301
A.12. Dependence of Learning Index on Mean of Learning Strategy Items	302
A.13. Distribution of the Fraction of Missing Values per Data Row (Participant)	304
A.14. Distribution of the Fraction of Missing Values per Variable	305
A.15. BOGER \mathbb{R} Code in the Editor Emacs	309
A.16. Residual Scatterplots for each Independent Variable (in the final model)	312
A.17. Residual Distributions for each Independent Variable (in the final model)	313
A.18. Distributions of the different R^2 and R_{abs} estimates – for all fitted individual models	314
A.19. Distributions of the different R^2 and R_{abs} estimates – only for the individual filtered (“good”) models in the bag	316

List of Tables

4.1. Principle Approaches to Model Selection	106
4.2. Meaning of Different Values of R^2	123
5.1. Survey Constructs	141
5.1. Survey Constructs	142
5.1. Survey Constructs	143
5.1. Survey Constructs	144
5.1. Survey Constructs	145
5.2. Usage of the 3 Different Task Branches	159
5.3. Reduction Level Usage	160
6.1. Simple Cross-Validation Data Sets A, B, C	175
6.2. Model Fit Comparison of Existing Algorithms	176
6.3. Parameter Labels	183
6.4. Model Fit Summary – Solid Results	200
7.1. Variable Shortname and Question Overview	219
7.2. Statistical Results – Personal History	240
A.1. Usage of Worksteps	291
A.2. Accident Model Fit Summary	310

previous SVN revision 658

1. Introduction

Chapter Contents

1.1. Executive Summary	13
1.2. Introduction and Motivation	17

1.1. Executive Summary

In competitive business environments, the effective creation, transfer and use of knowledge by organizations has only recently been widely recognized as a principle factor that creates value for stakeholders – by driving innovation as well as creating and maintaining competitive advantage (see section 1.2 on page 17). Therefore the aim of this study is to provide management guidance to support organizations with this ‘knowledge management’¹ challenge.

As will be detailed in section 2.3.1 on page 28, supporting learning is the primary challenge in knowledge management, since also **knowledge transfer involves**, on the receiving side, a person, who needs to acquire, i.e. *learn*, the transferred knowledge. Therefore this study focuses on **informal on-the-job learning** that occurs during normal problem solving activities at work and during innovation projects. This type of *individual* learning effect may eventually lead to group learning and the implementation of organizational improvements – i.e. *organizational learning* (see section 2.6 on page 77 for the detailed definitions of learning and knowledge). Whereas the focus on the individual person as unit of analysis may be unusual for the research stream on knowledge management, I argue that, instead, focusing on the individual level allows for more concrete managerial guidance.

Another reason for choosing **learning as the primary perspective** is because learning requires an active involvement on behalf of the learner. Hence the learner needs to be motivated and able to integrate the new knowledge into the prior knowledge he or she

¹In the field of business administration, this topic is referred to as ‘knowledge management’, however – as will be detailed later in this section – depending on the definition for ‘knowledge’, the question arises, if knowledge can be ‘managed’ directly at all.

already has on the topic (section 2.3.3 on page 34). Therefore learning is not a process that can be managed by direct intervention (in contrast to managing cash flows or material – see section 2.5.1 on page 69). Yet organizations can support learning *indirectly* by creating learning supportive working conditions.

In the literature on knowledge management and related fields, many features of organizations have been observed to affect learning. Rarely have these studies quantified the effect strength of these features on learning. However, because constraints such as limited time and people (and thus resources), **organizations** cannot practically optimize all aspects of their organization. Thus, it is imperative to **prioritize** their actions and concentrate on a few selected features with the greatest leverage effect on learning first.

Thus, the aim of this study is to give managerial guidance towards **which features of an organization can support on-the-job learning most strongly**.

To obtain a **ranking** of the most relevant organizational factors for learning, a fully structured survey was fielded in many different departments at Meyer Werft – a ship yard in northern Germany. In order to cover a broad range of organizational factors within a limited time frame, a dynamic and interactive online survey was implemented. (see chapter 5 on page 137). Given the lack of standardized, relevant and adequate standard constructs, on-the-job learning intensity at the individual level was measured by a newly developed and validated construct. This construct, referred to as the **learning index**, asks the participant to evaluate concrete learning episodes in a way that directly relates to the actual work of the respective participant (section 5.4.1 on page 146).

The data sample collected with the survey has an effective sample size of $n = 292$ with **293 variables** in total (section 6.3 on page 198). This large number of variables compared to the sample size in combination with a substantial amount of noise in the data, nonlinearities and collinearities (section 5.12 on page 163) posed a formidable challenge for the statistical analysis. Given these data specific challenges, it was necessary to **develop and validate the new statistical analysis algorithm BOGER**. BOGER features systematic and robust creation of a statistical model with a much smaller number of variables – which is composed of only the most important factors on learning (chapter 6 on page 171).

The result of the statistical analysis is a ranking of the most important factors affecting learning as well as an indication of the robustness and accuracy of this ranking (section 7.2 on page 217). Additionally, the effect of each organizational factor on learning is visualized with newly developed *model shape graphics* (section 7.1.3 on page 213). In these model shape graphics the factor-dependent distribution of the raw data is compared side-by-side with the effect as approximated by model. Compared to using a table with a few scalars per variable, this type of **visualization** allows for a much more direct and rich inspection of the results.

Despite the robust and relevant statistical results that emerged, this type of statistical

analysis only provides information about the association of a factor with learning rather than insights that directly address the causal effects of a factor on learning (section 4.1.5 on page 109). Therefore the statistical results for each important factor were analyzed in conjunction with insights from theory in section 7.3 on page 226.

The creation of the survey involved a detailed literature search, however, the statistical results inspired another wave of literature research in an even broader range of research streams (section 2.2 on page 23) with even further refined search terms. This lead to an overall improvement in the quality of the research. More details on this **iterative approach**, scientific theory and methodology can be found in the chapter on 3 on page 83.

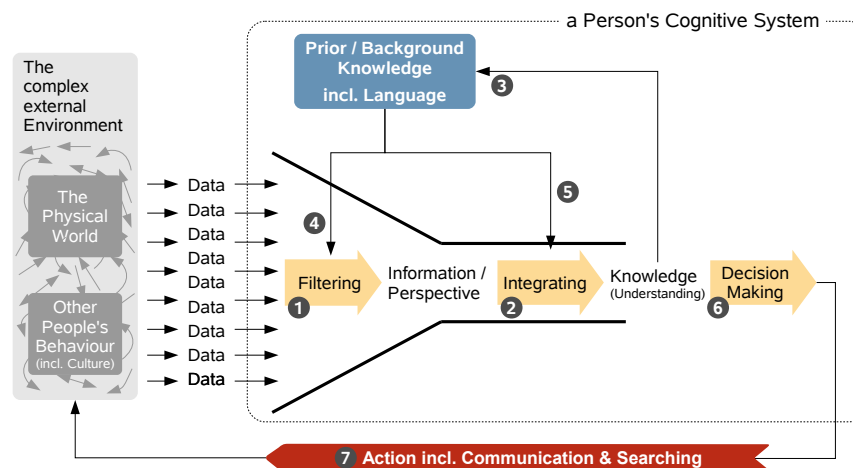


Figure 1.1.: Condensed Literature Insights: the PIA-Model (Source: Author)

Fused from literature and empirical results, the insights for this study are condensed in the newly developed PIA²-model (figure 1.1). The PIA-model illustrates the number of principle aspects of the learning process, which is in-line with the empirical results :

- As humans, we have developed the skill that allows us to **filter** (step 1) vast amounts of data as we receive it through all our senses. This skill allows us to effectively make sense of the complex situations that we face, despite our limited cognitive resources. The resulting information, which has usually been refined to a more relevant set of filtered information, is then either used in step 2 for creating new knowledge (i.e. learning) or directly for decision making. It is because of this process that we can behave and work effectively in a complex world – even if we do not and cannot understand its full complexity. Following Orr (1996), this filtering process is referred to as **perspective taking** (section 2.3.2 on page 30). Decision making (step 6) is

²PIA-model stands for 'Perspective Taking / Integration / Action Model'.

here a direct **link to action** and hence the **utilization of knowledge**.

- **Prior knowledge** plays an important role in filtering information and consequently also affects learning and decision making (links 4 and 5). Given that prior knowledge is shaped by personal histories and the socially constructed corporate culture, personal history and corporate culture affect learning and decision making (section 2.3.4 on page 41).
- As illustrated by the PIA-model and confirmed by the empirical results, **learning** is driven by two **feedback loops**, which require **personal motivation**, sufficient external data and effective filtering to support learning and effective decision making in a virtuous cycle (section 2.3.6 on page 47 and 7.3.5 on page 243).

While managers cannot directly ‘manage’ their employees’ cognitive processes, organizations have three principal levers for supporting individual learning:

- A climate for **controversial yet constructive discussions**, that allows for a comparison of a diversity of perspectives on a particular problem, can help employees refine their own perspectives on a problem and thus support decision making (section 2.3.5 on page 45 and 7.3.1 on page 226). Furthermore, employing a **systematic perspective** on a problem within an organization or its status as a whole can be offered to the involved organizational actors. Such a systematic analysis method frequently involves suitable **visualizations**. A systematic perspective on an organization can be created with the analysis and visualization methods inherent in the **EFQM** model, Kaplan’s balanced score card (**BSC**)³ or the **Toyota Production System** in the form of process models, project management plans or key performance indicators (section 2.4 on page 57). Alternatively, a suitable systematic perspective can be provided with interactive analysis and visualization tools in the form of **software**. In either case, the systematic perspective increases transparency by using a pre-filtering step to analyzing the available data. This allows the involved organizational actors to better filter the essence of a problem and thus improve decision making. The greatest benefit of using this approach is that it fuses a systematic but unintelligent perspective of the involved actors with the subjective but intelligent perspectives (section 2.3.3 on page 35).
- Support the **learning feedback loop**, by e.g. offering inspiring and motivating tasks, establishing a **suitable fault culture** (for constructively dealing with mistakes) or allowing for multiple iterations during problem solving.
- Provide **more raw data** by e.g. better access to information, better search tools, more experiments or more IT systems that capture information about the organi-

³see Kaplan and Norton (2007) or Lamotte and Carter (2000)

zation's performance. This lever, however, only becomes effective if the other two levers are already effective at a good level – i.e. if the involved actors already filter effectively and learn iteratively.

More details on the implications can be found in chapter 8 on page 263.

1.2. Introduction and Motivation

In an environment of globally available capital and production equipment, the shared knowledge embedded in a skilled workforce remains the only production factor, which is difficult to imitate by the competition (Teece, 2000). Thus in contrast to more traditional models of competitiveness, which are based e.g. on market dominance and efficiencies of scale, organizations in today's commercial environment should be viewed as dynamic systems of distributed knowledge with the purpose of coordinating complex activities and enabling technical as well as organizational innovations (Malik, 2008; Spender, 1996; Tsoukas, 2005b).

A corresponding shift of the perceived value of firms from a basis of tangible towards intangible factors of production, can also be observed in the stock market – as a kind of consensus amongst all investors: As Davenport et al. (2005) argue, the fraction of tangible assets to the total market capitalization of Dow Jones Index listed firms has dropped from almost 100% to only 20% between 1980 and 2005. Given that most analytic firm valuation methods are based on a more or less sophisticated analysis of cash flows⁴ (Luenberger, 1998), this observation implies that on average the community of market participants expects (discounted) cash flows that in sum⁵ far exceed the current value of the firm's assets. Hence investors, based on hopes, speculation but also based on the track record of firms, expect high returns from the firm's commercial activities due to largely intangible production factors such as brand recognition, customer connections – and knowledge (Davenport et al., 2005; Spender, 1996). Since balance sheets only list tangible assets, there are a number of initiatives⁶ to establish new and standardized valuation methods to quantify (i.e. estimate) the intangible value components of corporations (Andriessen, 2004; North, 2002).

Hence for corporations with highly trained employees and commonly also with high labor costs, business success critically depends on how well they acquire, protect, combine and utilize the knowledge embedded in their employees, the organization itself and

⁴A simple indicator for the relative value of a firm is the widely used future-earnings-per-share also known as the forward EPS ratio (in Germany the share price-to-current-profit (KGV) ratio is more popular). More sophisticated methods are based on the expected net present value of cash flows that are subject to random processes, which are modeled by event trees and estimated probabilities (Luenberger, 1998).

⁵Hence the net present value of the cash flows far exceeds the current total value of all tangible assets.

⁶In Germany the knowledge balance sheet initiative is can be found under 'Arbeitskreis Wissensbilanz' (<http://www.akwissensbilanz.org/>).

partner organizations in the value chain in order to create and/or maintain a competitive advantage (Spender, 1996; Teece, 2000). Schreyögg and Geiger (2007) add to this claim that the body of knowledge and its use needs to be superior compared to body and use of knowledge by competitors in order to provide a true competitive advantage.

While there is broad consensus about the importance of knowledge for organizations, there is little agreement on the definition of knowledge. In addition most definitions in the field of knowledge management are very vague⁷ (Schreyögg and Geiger, 2007). This disagreement and ambiguity causes severe limitations for application and research: Without a concrete definition of knowledge, which is distinguishable from data and information, it is difficult to operationalize knowledge, compare the results of different studies and design specific features of the organization in ways that support the creation, transfer and use of knowledge (section 2.5.5 on page 76).

Moreover there are various challenges with designing a useful definition of knowledge that is concrete enough to be useful in practical application – as will be detailed in section 2.5.6 on page 76. Therefore a number of authors suggest to focus on the activities connected to knowledge-intensive value creation processes within organizations (section 2.5.5 on page 76) instead of focusing on knowledge with object-like characteristics (section 2.5.1 on page 69). Therefore *individual learning* during problem solving at work – as an important part of the knowledge management challenge – was chosen as the principal perspective of this study (section 2.6 on page 77).

Focusing on knowledge intensive activities at work, furthermore allows to consider a much wider spectrum of research streams with many valuable insights (section 2.2 on page 23). Thus this study approaches knowledge intensive work in a interdisciplinary manner.

Given the abstract nature of the topic and the challenges with operationalizing knowledge for research, there is a lot of theoretical knowledge management literature. Hence there is no shortage of theories, frameworks and models on knowledge management. There is however a shortage of high quality empirical studies on knowledge. Since systematic empirical observation (qualitative as well as quantitative) is important to critically evaluate the large existing variety of theories and frameworks on knowledge, the literature research, that is a principle part of this study, is complemented with an empirical part: a fully structured survey (chapter 5 on page 137).

In summary, effectively creating, sharing and using knowledge is a principle factor for corporate competitiveness in today's global economy. Thus the aim of this thesis is to provide managerial guidance for the challenge of supporting knowledge intensive work with organizational means. To provide this guidance, an *engineering research approach* is employed, while leveraging theory and constructs from other fields such as psychology,

⁷Frequently non-knowledge is not defined (Schreyögg and Geiger, 2007). See also the knowledge definition of this study in section 2.6.1 on page 77.

management and the educational sciences.

The overall structure of this thesis is conventional with a theory chapter ([2 on page 21](#)), various chapters on data collection and methods ([3](#), [4](#), [5.2](#), [5](#) and [6](#)), a chapter on the analysis and interpretation of the empirical results ([7 on page 205](#)) followed by an implications chapter ([8 on page 263](#)) – including area for future research. A number of details, which are non-essential for an overall understanding of the arguments, are covered in the appendix ([A on page 287](#)).

Each chapter begins with an introduction to the contents of the respective chapter. For the writing style conventions used for this thesis, see also appendix section [A.1 on page 287](#). In the electronic PDF version of this document, all literature and page references are highlighted in blue or light green in order to mark hyperlinks, which allow the reader to directly jump to the reference target with a mouse click.

2. Theory of Learning

Chapter Contents

2.1. Summary of the Theory Chapter	22
2.2. Multiple Perspectives on Learning	23
2.3. Literature Findings Integrated into the PIA-Model	28
2.3.1. Active Learning	28
2.3.2. Perspective Taking	30
2.3.3. Integrating information	34
2.3.4. Prior / Background Knowledge and Perception	41
2.3.5. Language for Thought and Diversity in Discussions	43
2.3.6. Iterative Learning modelled with Feedback Loops	47
2.3.7. Searching for Information	51
2.3.8. Tacit Knowledge and Implicit Learning	52
2.4. Using the PIA-Model to Explain Industrial Practice Models	57
2.4.1. Commonalities of Industrial Practice Models	58
2.4.2. Project Management	58
2.4.3. The Toyota Production System	60
2.4.4. EFQM	62
2.4.5. Link to the PIA-Model	66
2.5. Alternative Perspectives Covered in Literature	68
2.5.1. Knowledge as Object rather than a Personal Skill	69
2.5.2. Identifying Application Relevant Knowledge in Maps	73
2.5.3. True Knowledge vs. Diversity of Perspectives	74
2.5.4. Valuation of Knowledge	75
2.5.5. Importance of Knowledge Definitions	76
2.5.6. Summary – Alternative Perspectives	76
2.6. Assumed Perspective & Definitions	77
2.6.1. Knowledge Definition	77
2.6.2. Organizational Learning Definition	78
2.6.3. Individual Learning Definition	79
2.7. Research Gap and Research Question	80
2.7.1. Research Gap	80
2.7.2. Research Question	80

2.1. Summary of the Theory Chapter

A large amount of literature is concerned with knowledge or learning. To increase this study's practical usefulness as managerial guidance, first this broad spectrum of literature is presented in section 2.2 on the next page. The aim is to integrate general insights on learning from several different disciplines, an approach that has yielded superior results with increased practical usefulness in other circumstances (Gittelman and Kogut, 2003).

Section 2.3 on page 28 discusses in greater depth select literature that theoretically supports the newly created PIA-model. The selection is in part due to the literature research for this thesis as well as the empirical results of this study – as a result of an iterative engineering research approach (see next section and section 3.1.6 on page 93 for further details). Other major contributions to the literature that are in contradiction with the PIA-model are presented in section 2.5 on page 68.

The theory chapter concludes with section 2.6 on page 77 on definitions used for the study and section 2.7 on page 80 on the research gap and question.

The PIA-model (figure 2.1 on page 31) was developed to illustrate and integrate the most important theoretical insights relevant to this study (see sections 1.2 on page 17 and 2.7.2 on page 80). The insights highlighted in the PIA-model can be summarized as follows:

- **Learning is an active endeavor**, primarily on behalf of the learner (section 2.3.1 on page 28).
- Humans live and work in complex environments and thus are constantly able to perceive a large stream of data. To make sense of all this data, the human brain **filters** it down to relevant bits of information. This process of **complexity reduction**, referred to as **perspective taking** (section 2.3.2 on page 30), has some special properties. For example, it is intelligent, adaptive and subjective, since it is **based on prior knowledge**, which can be socially constructed (section 2.3.4 on page 41) and thus itself subject to learning.
- Knowledge is created, i.e., learned, when the filtered information is **integrated** with previously acquired *prior background knowledge* into a coherent understanding of the situation or the problem (section 2.3.3 on page 34). This knowledge is then what drives **decision making** and action – i.e., the application of knowledge or *knowing* (section 2.5.5 on page 76).

By this definition, knowledge is only in the heads of people and directly connected with action (including communication and searching), which implies that knowledge cannot be managed directly like other corporate resources (such as money) – see section 2.5.1 on page 69. Instead the organizational environment can be designed to support learning –

which, however, requires different approaches than those that are popular in the knowledge management literature (e.g., knowledge databases – as described in section 2.5 on page 68). The approaches that support learning are discussed further at the very end of this thesis in the implication chapter 8 on page 263.

Finally, the definitions and perspectives used for this study are defined (section 2.6 on page 77), and the following research question is formulated:

What are the most important organizational features that support or hinder on-the-job learning?

(see section 2.7 on page 80)

2.2. Multiple Perspectives on Learning

As discussed in the introduction 1.2 on page 17 and given the focus on knowledge-intensive activities, a wide range of research fields hold valuable insights regarding the research question of this study.

Therefore, for this study, it is not sufficient to limit the search for relevant literature to a small set of technical terms, since technical terms are commonly specific to a single research field.

Ph.D. theses commonly begin with a literature section that covers at least a representative sample of the relevant literature. A couple of decades ago the search for literature was determined by the researcher's diligence and by the scope of the literature available from his or her library and inter-library loans. The latter was frequently the most severe limitation. Nowadays the widespread use of literature databases has dramatically enlarged the range of literature in which the researcher can search. Even though the search is still limited by the databases' size, that limit rarely poses a problem. Given the much larger search possibilities, searching strategies and in particular the choice of search keywords become the dominant challenge for finding relevant and high-quality literature.

When I¹ began my literature research in the field of knowledge management, I first searched for models of knowledge – probably driven by my engineering background. In engineering it is common and frequently effective to aim first at understanding the constitutive parts of any machine or process. However, instead of finding a single unambiguous and widely accepted model of knowledge, I found a large number of different perspectives on the same problem, highlighting various aspects of it. In addition, the categories used in these models are usually not clearly defined (examples are Argote et al. (2003); Elsbach et al. (2005); Hargadon and Fanelli (2002); North (2002); von Krogh and Venzin (1995)), making it difficult to accurately differentiate the categories, which creates

¹See appendix section A.1 on page 287 for the writing style conventions used for this thesis.

challenges for operationalization and practical application. (A more detailed discussion follows in section 2.5.5 on page 76.)

Thus, given the wide spectrum of perspectives on knowledge-intensive work and many different definitions and models of knowledge, it became necessary to scan the following range of related research streams:

- **Management Science & Organizational Science**

- **Knowledge Management** – focusing on how knowledge is created and transferred between organizational units or firms. The most prominent model of knowledge was popularized by Nonaka (1991) based on the explicit and tacit² distinction of knowledge by Polanyi (1966). Other authors focus more on the activity of using knowledge, referred to as *knowing* (Cook and Brown, 1999; Orlikowski, 2002; Tsoukas, 2005b).
- **Organizational Learning** – learning of employees within organizations, usually including implementation of the organizational changes (i.e. change management) (Argote, 1999; Argyris, 2002a; Brown and Duguid, 1991; Crossan et al., 1999) – see also section 2.6 on page 77.
- **Sense Making** – describes the process of understanding a situation – within a complex organizational setting – by a group or an individual. This understanding then serves as a “springboard into action” (Weick et al., 2005).
- **Situated Practice & Narratives** – This stream aims to describe and explain how practice evolves within organizations – frequently including a focus on narratives and their importance in informal knowledge transfer (Orr, 1996). Ethnographic research methods (Samra-Fredericks, 2000), involving the creation of thick descriptions³ (Elsbach et al., 2005; Weick, 1993) of the observed processes and actors, are very common in this research stream.
- **Expert Systems** In the 1980s a number of scholars and software engineers were trying to create so called *expert systems*, which – mostly by rule-based software – were aiming to capture expert knowledge and eliminate the need for expert attention for relatively simple tasks. Since these not truly intelligent systems were not able to imitate human judgment (Ackoff, 1989; Svenmarck and Dekker, 2003), work on this topic soon faded away. While this research

²Explicit knowledge in Nonaka’s model is knowledge that is easily verbalized and captured in documents, while tacit knowledge is personal and in the heads of people, requiring significantly more effort and different methods to convert it into the more transferable explicit knowledge. There is an extensive discussion about the precise features of tacit knowledge and whether it can be converted into explicit knowledge (D’Eredita and Barreto, 2006a; Glisby and Holden, 2003; Schreyögg and Geiger, 2005; Tsoukas, 2005b) – more details in section 2.5.1 on page 69.

³*Thick descriptions* are detailed story-like reports on the results of ethnographic research – e.g. collected while shadowing a developer team at their normal work.

stream is certainly not a current and up-to-date discussion, the stream's history from an early hype to negligence holds some insightful lessons – useful to judge the state of current research streams.

- **Psychology** In the field of psychology, useful insights regarding knowledge intensive work are clustered around the following terms and research paradigms⁴:

- **Cognitive Psychology** – focuses on the mechanisms of human information processing incl. perception, memory training and learning (Anderson, 1990). The first focus was on simple learning processes (such as vocabulary memorization) but later publications highlight that cognition also depends on the knowledge a person has acquired in the past (Sternberg, 2008) – this ability is termed *crystalized intelligence* (Cattell, 1971).
- **Psychology of Child Development** – investigates how children develop their cognitive abilities. The research stream is rooted in educational psychology (Piaget, 2003) and in cognitive psychology Siegler (2005).
- **Lifespan Psychology** – Drawing on the insights from cognitive psychology and the psychology of child development, lifespan psychology is concerned with the development of cognitive abilities over the entire lifespan – not just childhood (Baltes and Staudinger, 1999).
- **Educational Psychology** – focuses on the process of academic or school-based learning (Butler and Winne, 1995; Siegler, 2005) and how learning skills can be trained and improved (Roßnagel, 2008).
- The **Research on Epistemological Beliefs** – can be seen as a branch of educational psychology, which focus on a particular kind of prior knowledge: the beliefs about the structure and validity of knowledge. Various studies show that these beliefs have an important effect on learning (Hofer, 2001; Schommer, 1990).
- The **Psychology of Expertise** – investigates how outstanding expertise is developed, using world class chess or piano players as an example (Ericsson and Lehmann, 1996; Ericsson et al., 2007).
- The **Psychology of Problem Solving** – is concerned with the cognitive strategies that people use to solve complex problems using experiments with computer simulations (“micro-worlds”) – see Dörner et al. (1999) and Badke-Schaub and Frankenberger (2004).

⁴This categorization does not intend to be a formal and complete model of current psychological research but is meant only as a possibly incomplete overview of research in psychology relevant to knowledge intensive work. Noteworthy is particularly that these research paradigms evolve over time and thus that they can not be part of a consistent model encompassing all research efforts.

- **Psychology of Decision Making** – closely related to problem solving, the research on decision making focuses how decisions are made using prior knowledge, experience, mental models, discussion, analysis and more – individually or in teams (Badke-Schaub et al., 2007; Brehmer, 2005)
- **Educational Sciences** – investigates learning processes in school or higher education settings (Clark, 2005). In these settings the knowledge or skill is usually pre-defined and taught to an entire group of students – both aspects that make these formal teaching situations very amenable to research of formal learning.
- **Sociology**
 - **Knowledge and Theory of Action in Sociology** – is concerned with how societies and organizations shape the way peoples perceive their environment and how people with their knowledge and actions can shape the shared beliefs in societies and organizations (Habermas, 1989; O'Donnell et al., 2003).
- **Philosophy** – One of the focuses of philosophy is epistemology, i.e. the theory of knowledge. Aside from the classical thinkers such as Plato, Aristoteles, or the philosophers of the period of enlightenment such as Kant and Foucault, a number of modern day scholars use deduction in combination with insightful examples or cases for their arguments regarding the nature of knowledge. Examples for studies of this kind are Habermas (1989), Weick (1993) and Tsoukas (2005b).
- **Neurosciences** – deals with the mechanisms of the brain currently on a very low biological level of the neuron. With new sensory technology such as magnetic resonance imaging (MRI), recent findings confirm a number of insights from cognitive psychology in a more direct way (Jaeggi et al., 2007).
- **Industrial Practice / Engineering** Outside of the academic community in various industries, a few models featuring organizational learning have been developed as an essence of the experiences of many. The following models stood the test of application in practical settings:
 - The **Toyota Production System (TPS)** developed and used at car maker Toyota focuses on continuous organizational learning and has many methods that include features also suggested by academic research. It is a method that has proven to be very successful in practice – even outside the Japanese cultural context – see section 2.4.3 on page 60.
 - **TQM & EFQM** – related to the Toyota philosophy and stemming from a tradition of quality management, total quality management (TQM) methods have been further developed into an overarching framework for evaluating organizations using a combination of key performance indicators (KPIs) and qualitative

judgment. The way in which the EFQM ‘model criteria’ (groups of KPIs) are designed and causally connected as leading and lagging indicators, echos many of the findings from academic research. Further details in [section 2.4.4 on page 62](#).

All of these perspectives are closely related but not identical. The different perspectives are due to slightly different objectives – e.g.:

- the control (“management”) of knowledge in corporate contexts by management instead of
- the accumulation of knowledge and skill without direct connection to a value-adding work process in the context of schools or universities.

The different objectives lead to different foci expressed by the language and, in particular, by the metaphors used in these research streams – see ([Tsoukas, 2005b](#), Chpt. 10) and ([Andriessen, 2006](#)). Language is here both driven by the objectives and shaping our perspective and with it our way to think about the subject ([Tsoukas \(2005a\)](#), [Dörner \(2005, p. 120\)](#), [Woodward-Kron \(2008\)](#) and [Mertz \(2007\)](#)), which in turn may influence, refocus or sharpen our objectives (see also [section 2.3.5 on page 43](#)). The result is that whole research communities thrive around particular metaphors for a problem and may even re-explore aspects of the problem that have been already investigated and discussed in other research communities⁵. An example is the research on knowledge management, which only significantly unfolded in the last two decades ([Nonaka et al., 2006](#)) more or less from scratch with a special focus on non-formal learning within firms⁶, while many important theoretical foundations had already been laid in the educational sciences – e.g. ([Piaget, 2003](#); [Rasmussen, 2001](#)).

The different research streams, despite using different perspectives, frequently come to similar results. Sometimes however differences are maintained over longer periods of time. Research streams with hypothesis that cannot be maintained or with a perspective that proves itself as less useful compared to other perspectives, frequently just die out – as has happened to the research on expert systems ([Ackoff, 1989](#)).

In summary: There are many research streams or perspectives covering aspects of knowledge management from different angles. The challenge for this study is to integrate these perspectives into a few coherent general insights about knowledge management.

The aim of this literature search is therefore to present the most prominent perspectives on the problem and integrate the findings from those perspectives – leading to a few general insights that have received broad support. Based on these general insights, a particular

⁵Noteworthy is that some researchers work in multiple research communities and some of the research streams also overlap.

⁶see also the detailed discussion of the knowledge management research stream in [section 2.5 on page 68](#)

2.3. Literature Findings Integrated into the PIA-Model

perspective is chosen for the purpose of this study including definitions of the problem and of the key terms.

To improve quality, this research was conducted following an iterative approach (see section 3.1.6 on page 93): After a first literature search, a draft literature section was written, the survey designed, the data collected and analyzed. The statistical analysis of the data then triggered further literature research, of which the insights were used to refine this theory chapter and the findings were used to support the interpretations directly in the result interpretation chapter 7 on page 205. A more detailed discussion on the research approach will follow in section 3.2 on page 97.

2.3. Literature Findings Integrated into the PIA-Model

From the different research streams, mentioned in section 2.2 on page 23, a number of common insights emerge. A summary of the insights was already given in section 2.1 on page 22.

To provide a graphical overview, these insights have been integrated in a model called PIA(see figure 2.1 on page 31).

2.3.1. Active Learning

Around the 1970s the dominant view of learning shifted from a notion of a passive learner learning from an active teacher to learning theories, which stress the active engagement of the student. This shift occurred both in the educational sciences (Clark, 2005) and developmental psychology (Siegler, 2005, p. 770). In these models the teacher can 'only' support the student's learning efforts by creating a suitable context for example by giving the student a task, which requires the student to learn a new skill in order to complete the task. Hence the motivation of a student to actively engage in learning in order to solve the task plays an important role (see also the self-regulated learning model of Butler and Winne (1995) in figure 2.4 on page 48).

Some of these learning models are very specific to school or academic settings (Butler and Winne, 1995; Roßnagel, 2008), in which a predefined body of knowledge is taught with often small and targeted exercises. Other authors cover a broader and less specific spectrum of learning situations, which arise while solving everyday problems in child development (Piaget, 2003) and while solving predefined tasks posed to school children in various experiments (Siegler, 2005). Thus the latter kind of models, covers situations very similar to problem solving on-the-job.

In the field of knowledge management, most researchers emphasize the process of externalizing knowledge, i.e. getting the knowledge out of the head of employees and preparing it for the transfer to other organizational units by means of documents, discussions or

collaboration (Argote et al., 2003; Nonaka and Takeuchi, 1995; North, 2002). Hence the focus is on a *push* concept of knowledge while only few use a *pull* concept, which involves searching (see section 2.3.7 on page 51). This is also partly due to the choice of *knowledge as research paradigm*, which inherently suggests a quasi-material object-like nature of knowledge, which can be deliberately created, stored, hoarded⁷ (Abou-Zeid, 2002; De Long and Fahey, 2000; North, 2002) and transferred (Nonaka et al., 2000).

Thus, while the principal insight, that learning is primarily an active and personal endeavor on behalf of the learner (Butler and Winne, 1995; Jacobson and Prusak, 2006; Piaget, 2003; Siegler, 2005), appears trivial – it is frequently overlooked or at least deemphasized in much of knowledge management literature.

An important implication of an active learning perspective is furthermore that knowledge transfer based on individual learning processes makes *direct* management of knowledge impossible. Knowledge can thus only be management *indirectly* by creating a learning supportive context (Fahey and Prusak, 1998; O'Donnell et al., 2003), which can include a wide range of measures such as the supply of suitable information as well as a supportive leadership style.

Furthermore a number of scholars stress the importance of the task or the exercise that serves as objective and context for learning (Butler and Winne, 1995; Piaget, 2003; Siegler, 2005). In the before mentioned research stream 'Psychology of Problem Solving' Badke-Schaub and Strohschneider (1998); Brehmer (2005); Dörner et al. (1999) and others are designing tasks to be completed in computer simulated 'micro worlds' as learning contexts – in order to study the problem-solving and learning behavior of different participants with different cultural or experience backgrounds in experiments⁸. Brehmer (2005) for example observed, that tasks with delayed feedback impede learning while performing the task. Furthermore Salter and Gann (2003) in his study on drivers of innovation, concludes that difficult and open tasks lead to more learning and thus innovation. Thus the task, as an important aspect of the learning context, plays an important role in learning and hence features of the task affect learning.

Given the high level of attention that the task receives in the literature related to learning and the variability of tasks encountered in work settings, this study – with an *on-the-job* learning focus – needs to carefully take into account the nature of the task triggering learning experiences.

Consequently the properties of the task should be included in the empirical data acquisition. As will be presented in section 5.3 on page 141 in further detail, several aspects of the task – during which learning experiences occurred – have been surveyed, for example: the concrete example task, which is chosen by the survey participant, is classified into

⁷The knowledge as object perspective will be covered in further detail in section 2.5.1 on page 69.

⁸Unfortunately these studies with micro-worlds do not include the opposite (and difficult to realize) approach: studying variability of learning behavior given different tasks and different micro-worlds.

2.3. Literature Findings Integrated into the PIA-Model

three groups: routine tasks, more open tasks within longer term projects and even more open tasks within innovation projects.

This is why innovation is included in the survey data as a special task type: special innovation projects, for which it is clear *from the beginning* that previously applied solutions will not work without substantial modification and thus the employee is challenged to devise a novel and innovative solution. While innovations are frequently the *spontaneous* end-result of a learning process, innovations are included in this survey as an *innovation challenge* rather than innovation as an outcome variable. In order to ensure that only work projects that include an obvious challenge for innovation, are classified correctly in the data as innovation projects, i.e. in order to assure that the definition of ‘innovation project’ is understood in the intended way by the participants, multiple probing questions are used to check a number of criteria for innovative projects (see section 5.7 on page 153 for details).

Another aspect of active learning is the need to economize learning effort (e.g. invested time) (Boisot and Canals, 2004; Cohen and Levinthal, 1990). Davenport and Beck (2001) and Hansen and Haas (2001) go even further and point out that in a world with easily accessible information, those who supply information need to compete for the learners’ attention.

Summarizing, learning requires the *active* engagement of the learner and the learners effectiveness in economizing his or her resources (e.g. time) for learning. Furthermore the nature of the task during which an employee learns, is an important aspect of the learning context and sets the learning objectives and therefore should be considered in this study on on-the-job learning.

2.3.2. Perspective Taking

Perspective Taking is filtering Information out of all available Data Given that learning is an active and personal process, investigating the details of the process on an individual level yields useful insights regarding how to manage a learning supportive organizational context.

Knowledge workers continually have to make sense of the situation at work. Not only since the wide-spread introduction of the Internet, do knowledge workers have a lot of data and information accessible to them. In fact humans in general are constantly faced with a very large stream of data, which includes any visual, auditory or other sensory input and may or may not be relevant to the problem at hand. Hence, for the purpose of this study and following Ackoff (1989), Davenport et al. (2001) and Boisot and Canals (2004), *data* is defined as the unfiltered stream of sensory input available to a person or an organization⁹.

⁹In large parts of the knowledge management literature data, information and knowledge are neither

Given that most of this data is not relevant to the task at hand and cognitive resources are limited, humans (and most other living species) have developed filtering mechanisms in order to extract relevant information. All our sensory systems (especially the eye (Platek and Kemp, 2009)) are built in ways that allow for effective reduction¹⁰ of the data down to much fewer – yet relevant – features (D'Eredita and Barreto, 2006b; Tsoukas, 2005b). Thus *information* is defined here as a collection of relevant features, i.e. a filtered extract of the entire available data. Orr (1996) refers to this filtering and feature selection process as viewing a particular problem from a particular *perspective*. Other authors refer to this filtering process as ‘attention drawing’ to particular aspects of the problem (Davenport and Beck, 2001; Davenport et al., 2001; Weick et al., 2005).

Following this idea of filtering, it becomes evident that the more data is available, the better the filtered information can be – provided that sufficiently efficient filtering strategies are employed.

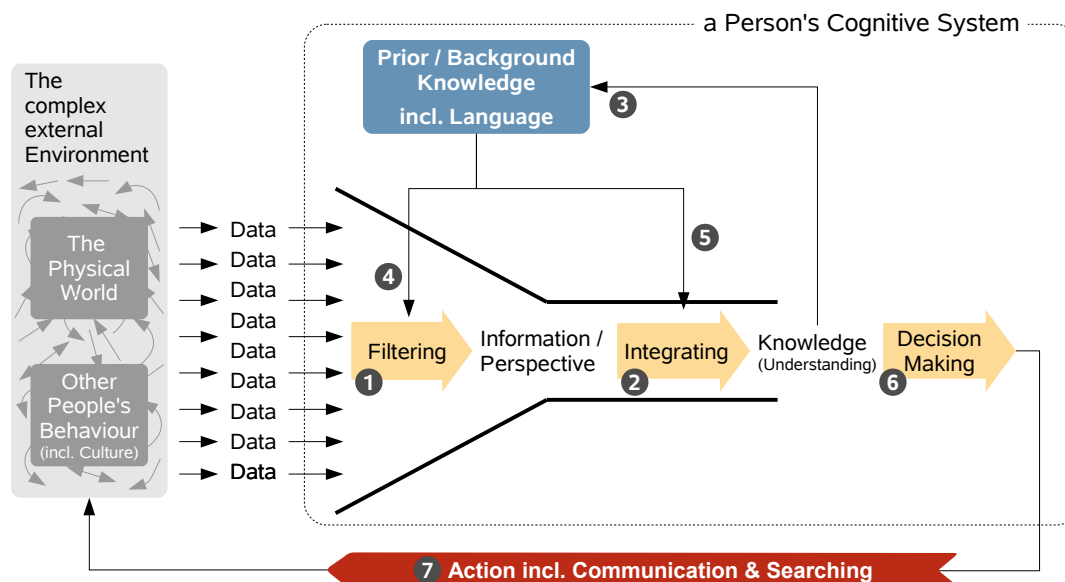


Figure 2.1.: PIA-model – Perspective Taking / Integration / Action Model (Source: Author)

explicitly defined nor are the differences of these categories explained (Argote et al., 2003; Hofer-Alfeis, 2003; Nonaka, 1991; North, 2002). This is rather surprising, since the field of knowledge management aims to be a holistic theory of the firm (Nonaka and Takeuchi, 1995; Spender, 1996) – rather than just information management with a new label. See also section 2.5.5 on page 76 on the importance of knowledge definitions.

¹⁰Many popular data formats, such as JPEG or MPEG-1 Audio Layer 3 (MP3), leverage this effect for a reduction of the memory size, by removing features from the files, which people would not notice or hear. Thus in computer science terms, the human sensory system uses lossy compression strategies for data reduction.

2.3. Literature Findings Integrated into the PIA-Model

Once relevant features and bits of information are filtered for a given task, people integrate these to get a coherent view and explanation for the situation as preparation for decision making. The further abstraction and connection of information into working models of the world is referred to as *knowledge* in this study. Graphically¹¹ figure 2.1 on the previous page illustrates this process of cognition with filtering of the data in **step 1** and integration of information to more abstract knowledge in **step 2**. In the following figure 2.1 on the preceding page will be referred to as *Perspective Taking / Integration / Action Model* or short **PIA-model**. More details will follow in the next section (2.3.3 on page 34).

Perception filtering can also be dependent on current aims, e.g. when a person enters a meeting room in order to sit down there, the person will first scan the room for a table and chairs – rather than proceeding with a detailed inspection of the carpet. Gibson (1986) modelled this application dependence in his model of ‘direct perception’ (Vicente, 2003).

Note that for the sake of simplicity, factors related to self-regulation such as the person’s current aims or motivation are left out of the PIA-model (figure 2.1 on the preceding page). Including self-regulatory processes in the PIA-model, like in the self-regulated learning model (SLR) by Butler and Winne (1995) (figure 2.4 on page 48), because these aspects would add a whole new layer of complexity to the model.

Filtering is Complexity Reduction Considering filtering as a preparation for decision making, the filtering process is a way to simplify our picture of the world around us down to the features that are relevant to a given task (Badke-Schaub et al., 2007). This measure of complexity reduction (Malik, 2008; Tsoukas, 2005b) allows us to effectively react to the complexity of the world, despite our limited cognitive resources for integrating information and decision making (Jaeggi et al., 2007; Porac and Shapira, 2001; Tsoukas, 2005b). Some authors refer to this ability as ‘human judgement’ – i.e. selecting the essential features from the total data of a complex situation for decision making (Ackoff, 1989; Davenport et al., 2001; Tsoukas, 2005b; Weick, 1993). Note that only the perceived complexity is reduced – the real complexity in the situation or the problem is not reduced¹².

Thus only complexity reduction allows us to behave and work efficiently in our complex environment – certainly at the expense of overlooking details of the actual world around us. Over time we as humans fine-tune our complexity reduction behavior in such a way, such that we overlook only irrelevant¹³ details, which do not reduce the effectiveness of

¹¹The aim of this graphic is to illustrate the common denominator of the various studies cited here – in a necessarily simplifying form.

¹²In some engineering systems, the real complexity can be reduced by restructuring many tightly inter-dependent system components into much fewer modules, which are designed to be tolerant to external disturbances. Much of the system behaviour can then described and analysed with the overall behaviour of the modules as observed at the module interfaces. This is one of the rare instances, where the true complexity can be reduced.

¹³Note that, what is relevant, may change over time or depending on the situation and thus these changes

our behaviour (Gibson, 1986; Vicente, 2003).

Prior Knowledge biases the Filtering Process The PIA-model in figure 2.1 on page 31 illustrates another important effect: Link 4 symbolizes the influence that a person’s prior or background knowledge has on filtering. Thus the way we filter the data – or in other words: the perspective we take on the actual world – may change over time and with our experiences in the actual world we create and modify our prior knowledge. Since prior knowledge creation is also dependent on the perspective we assume, a feedback loop with interesting effects is created. Section 2.3.4 on page 41 will cover prior knowledge in further detail after the next section, which covers the integration of information to knowledge.

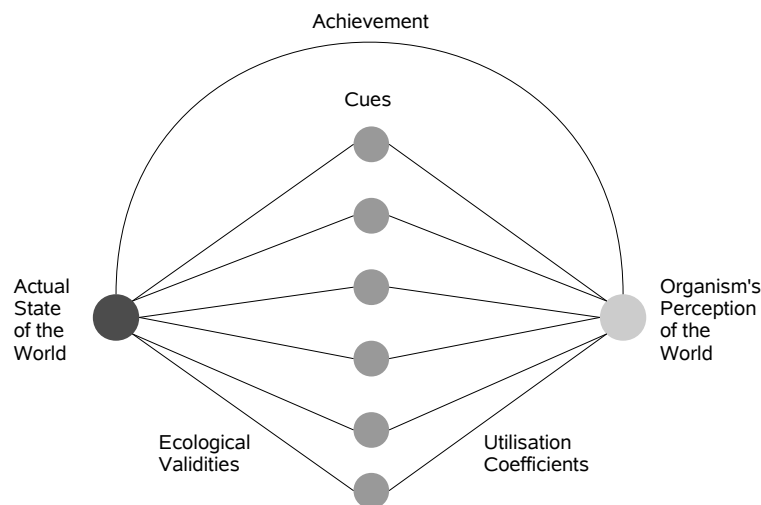


Figure 2.2.: Brunswik’s Lens Model (Source: Author following Vicente (2003))

A similar dependence of perception on prior knowledge is described in ecological psychology by Brunswik (1956): He models perception of the actual world by an organism as being mediated by a set of cues – as illustrated by his *lens model* in figure 2.2. The organism does not perceive the full complexity of the actual world but sees only the cues: a reduced set of features. To recognize and make sense of the cues, the organism employs a weighting and interpretation function that Brunswik refers to as ‘utilisation coefficients’. These utilization coefficients are optimized iteratively through readjustment of the coefficients towards the probabilistically and objectively optimal coefficients (the ‘ecological validities’) by the organism – upon positive and negative experiences with the actual world as illustrated by the ‘achievement’ feedback link. Hence the lens model includes the effect of the organisms personal history, which made it popular within a community of researchers frequently referred to as ‘neo-Brunswikians’ (Vicente, 2003). An example is the

create new challenges even for experienced people.

2.3. Literature Findings Integrated into the PIA-Model

recent refinement of the lens model by [Stewart and Lusk \(1994\)](#) to distinguish between ‘true descriptors’, ‘cues’ – actually available to the organism – and actually perceived ‘subjective cues’ – see also [Stewart \(2001\)](#) and [Vicente \(2003, p. 258\)](#).

The original lens model includes the essential elements of the PIA-model (figure 2.1 on page 31): the complexity of all available data is reduced to a set of cues (step 1 in the PIA-model), utilisation coefficients (a simple form of history dependent prior knowledge) drives the integration of the cues to a coherent judgement of the situation (step 2 in the PIA-model) and this whole process is refined by a feedback process involving experiences with the actual world (steps 1,2,3,4,6 and 7 of the PIA-model).

Nevertheless, with the PIA-model, a new model was created for this study, since it emphasizes other aspects that are important in this study, which the lens model deemphasizes: The PIA-model clearly shows prior knowledge as an important component and how it is embedded with an internal feedback loop – see also section 2.3.6 on page 47. Brunswick’s representation of the cues is problematic¹⁴ and thus the PIA-model avoids showing cues and instead illustrates that of all available cues only a few are filtered out and consciously perceived – symbolized by the filtering funnel.

Hence, in summary, people’s decisions depends strongly on how they have filtered all available data down to a limited set of information. The filtering mechanism is biased by a person’s prior knowledge.

2.3.3. Integrating information

Integrating Information causes Learning After covering the filtering and feature selection process, this section sheds further light on how the collected and filtered information is connected in order to prepare for decision making.

In his ethnographic study on copy machine technicians at Xerox, [Orr \(1996\)](#) describes various episodes of a pair of technicians on repair jobs at customer sites. Very revealing are his analyzes of the diagnosis process of copy machine failures. Orr refers to the data filtering process as *perspective taking*, which in his case is very much enhanced by social interaction of the two technicians through the telling of past experiences in narratives. The narratives include hints towards a particular perspective on the machine’s problem (illustrated by link 3 in the PIA-model). Furthermore the narratives also hint towards a possible connection of the facts and thus towards a hypothesis of the causal chain of events – consistent with the observed failure symptoms (step 2 in the PIA-model). Orr refers to this last step before decision making as *integration of the facts*. In the next step,

¹⁴As [Stewart and Lusk \(1994\)](#) propose in their extended lens model, there needs to be a distinction between the cues, which the actual world exhibits and the reduced set of cues, which a person perceives for further integration. But even the graphical representation by [Stewart and Lusk \(1994\)](#) does not directly illustrate that the number of perceived cues is intentionally much smaller than the number of available cues – which is an advantage of the funnel symbolic of the PIA-model.

the technicians frequently decide to validate their hypothesis by exchanging the part, which is suspected to be the root cause of the failure, with a spare part. This last step is illustrated in the PIA-model (figure 2.1 on page 31) with steps 6 and 7¹⁵. Blackler et al. (2000) share Orr’s emphasis on ‘perspective taking’ and attributes great influence to perspective making and taking in a corporate strategy review process within a British high technology company.

In cognitive psychology, Sternberg and Hedlund (2002) propose a model very similar to the PIA-model (figure 2.1 on page 31). Learning is modeled by the following process steps¹⁶:

- Selective encoding (feature selection in order to extract relevant information, perspective taking)
- Selective combination (“integrating information into a meaningful interpretation of the situation”),
- Selective comparison (“relating new information to existing knowledge”)

Sternberg’s model is in one aspect more refined than the PIA-model, since it distinguishes two different ways of integrating information: creating new mental models and refining existing mental models.

Similar insights can be found under the research paradigm *sense making*, which is for Weick et al. (2005) the understanding of how to connect i.e. integrate different facts – as a “springboard for action”. They further elaborate: “*To deal with ambiguity, interdependent people search for meaning, settle for plausibility, and move on.*”. Hence people search for a plausible connection and explanation of the filtered information – which is not necessarily the correct integration of the facts. Once a plausible explanation is found, it adds incrementally to the body of background knowledge¹⁷.

Visualization supports Perspective Taking and thus also Learning All models so far presented in this section, emphasize the importance of perspective taking for learning and decision making. While perspective taking and integration of the filtered information is a cognitive activity concealed in the brain of the learner, it can be supported externally: One possible support is the interaction with a teacher, who draws the learner’s attention to relevant aspects of a problem (section 2.3.1 on page 28 and 2.5.1 on page 69). A similar support effect can be achieved by visualization (Ertl et al., 2008). Learning supportive visualizations can be either manually created (e.g. a schematic drawing or a mind map)

¹⁵Compare also the argument of Weick et al. (2005) that “[...] we act our way into belated understanding.”, p. 419..

¹⁶Both citations are from Sternberg and Hedlund (2002, p. 146).

¹⁷from (Weick et al., 2005, p. 419): “[...] the concept of sensemaking suggests that plausibility rather than accuracy is the ongoing standard that guides learning”.

2.3. Literature Findings Integrated into the PIA-Model

or computer generated (e.g. a statistical graph¹⁸ such as the Shewhart process control chart¹⁹).

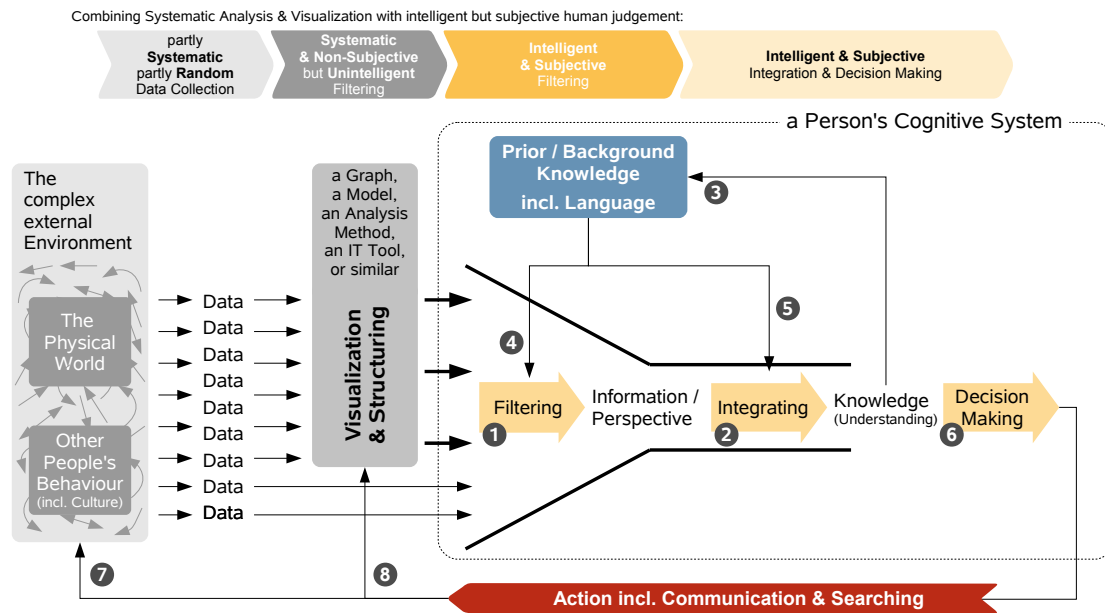


Figure 2.3.: PIA-model with Visualization (Source: Author)

The PIA-model in figure 2.3 illustrates the effect of a suitable visualization: A suitable graphical schematic or a statistical graph, can reduce the complexity of the actual problem and filter out relevant aspects of complex problems. Hence such visualizations can serve as a pre-filtering tool for complex data or a complex situation, reducing the need for filtering in the persons brain, which frees up cognitive resources that can be used for more refined filtering. As discussed before, improved perspective taking also leads to improved learning and decision making.

In addition, the use of visualization filter is an additional aid not an exclusive perspective on the problem: With the newly gained insights from the systematic perspective, a person may still have another unfiltered but more educated²⁰ look at the complex situation. Hence the strength of systematic visualization lies not in replacing human filtering but in combining systematic and thus non-subjective analysis and visualization with intelligent but subjective human judgement. As illustrated with the process arrows at the top

¹⁸For an example of a statistical graph see figure 7.2 on page 215.

¹⁹In statistical process control the Shewhart control chart show the process performance (e.g. the true dimension of a mass produced component) over the production time for monitoring the true dimension within tolerance limits (Devor et al., 1992, p. 146).

²⁰As step 3 illustrates, the systematic view of the situation may also have an education effect – i.e. refine the persons prior knowledge, which in turn enables more effective perspective taking.

of figure 2.3 on the facing page, this combines the strength and mitigates the weaknesses of both complexity reduction approaches.

In their case study on problem solving and innovation in civil construction projects, Salter and Gann (2003) found a simple visualization by ‘sketching on paper’ as one of the most effective problem solving tools – second only to conversations with colleagues. Similarly Dodgson et al. (2007) and Carlile (2002) found that models²¹ of a new product in development support the ‘design conversation’ of the participating engineers from multiple disciplines.

The wearable computing pioneer Steve Mann has pushed the concept of computer generated visualization to a new extreme with the invention of his EyeTap glasses, which allow to overlay a computer picture onto the real picture as seen through the glasses (Mann, 2001; Mann and Fung, 2002). He demonstrated the overlay of different kinds of computer generated visualizations, that were linked to the direction of view of the user, e.g. text information based on face recognition, additional graphics replacing other visual objects in reality and a semi-transparent overlay of e.g. temperature information from a heat camera (Mann, 2005; Mann and Fung, 2002). This *augmented reality* mixing or overlaying of a systematic and non-intelligent computer generated visualization onto the complex reality supports rather than replaces human decision making. This combination can be very potent, since it leverages the different strengths of man and machine. Steve Mann refers to this approach as “*Humanistic Intelligence (HI)*” and contrasts it against “*Artificial Intelligence (AI)*” approaches, which rely solely on the computer for decision making (Mann, 2001; Mann and Barfield, 2003).

As will be detailed in section 2.3.6 on page 47 the learning process itself may be iterative and may also allow for iterative fine-tuning of the visualization method (step 8 in fig. 2.3 on the facing page). This explains why software used for visualization²², can be a powerful tool for decision support: The user can iteratively and interactively refine the visualization and learn more about a problem or a situation. Since the software reduces the effort required to generate the visualizations, the user can perform more iterations and thus gain more refined insights. This effect was already realized in the 1960s by one of the pioneers of computer aided design (CAD), Ivan Sutherland, who saw his early 2D CAD system “*Sketchpad*” not just as a fancier and electronic version of a drawing table²³ but as an interactive system, which – by its visualization and interactive features – supports the engineer in iteratively finding solutions to design problems (Salter and Gann, 2003; Sutherland, 1964).

²¹These models could be computer aided design (CAD) models, simulations or similar ‘boundary objects’ (Carlile, 2002) that support visualization and in turn also communication.

²²The use of software for visualization is frequently overlooked over the other uses of software, e.g. for automation of tasks or communication.

²³Ironically many companies to this date use CAD systems purely as a more efficient tool for draftsmen to generate drawings.

2.3. Literature Findings Integrated into the PIA-Model

Conversely the lack of a suitable visual support can also lead to a reduced ability for effective decision making as the observations by [Sengupta et al. \(2008\)](#) in an experiment involving a computer simulation of a software project confirm: Many of the participating well experienced project managers failed to make suitable decisions (e.g. when to put more people on the project) and even failed to learn from their mistakes, when delayed effects of managerial decisions were not properly visualized and thus managers were not sufficiently supported in their perception of the project situation. More on the effect of a suitable perspective on projects in project management in [section 2.4.2 on page 58](#).

Thus, in summary, visualization can support learning and decision making by improving filtering – i.e. perspective taking. Moreover, interactive visualization software can support an iterative refinement of the learner’s perspective.

How Knowledge is accepted as True Weick’s notion of plausibility as standard for “truth”, raises the question how knowledge is justified, i.e. found to be true. [von Krogh and Grand \(2000\)](#) emphasize in their description of the information-to-knowledge integration step that new information and knowledge is carefully integrated with the existing *dominant logic*. Similar to Piaget’s concept, new knowledge is either integrated in the existing dominant knowledge or, when the new evidence is overwhelming, a part of the dominant logic is replaced by a new mental model. [von Krogh and Grand \(2000\)](#) in addition highlight that dominant logic may be shared, e.g. within an organization, by social interaction – again similar to Habermas ([O’Donnell et al., 2003](#)).

The source of information – in particular its trustworthiness – appears to further affect knowledge justification: [Kane et al. \(2005\)](#) demonstrated that sharing social identity with a person, who is the source of new information, makes the justification of this knowledge more likely. Yet empirical studies on knowledge justification are rare, which supports [Tsoukas \(2005b\)](#) with his argument that further research is necessary to shed more light on the processes of knowledge justification.

Yet one important aspect of knowledge justification is widely supported: knowledge justification depends on prior knowledge and thus may be subject to change.

Prior Knowledge affecting Integration Thus the integration process step, does not only integrate recently filtered information but also leverages prior knowledge to integrate the new knowledge (link 5 in the PIA-model). Therefore integrating knowledge requires retrieval of older knowledge including the retrieval of memories. This retrieval process appears to have peculiar properties that somewhat resemble the bias that prior knowledge has on perception:

[Loftus \(2003\)](#) has observed in one of her experiments, that involved showing a video of a car accident, that witness accounts of a past event (i.e. memories) can be skewed by how an investigator asks questions about the event. Similarly there appears to be hindsight

bias, i.e. knowledge about the outcome of an event (e.g. the financial crisis in 2008 and 2009), tends to make people believe that they “knew it all along”, i.e. it appears obvious in hindsight to focus on the factors leading to the event, while at the time many other factors were seen as relevant²⁴ Bernstein et al. (2007).

Further support can be found in sociology, where the well known Thomas-Theorem describes how a particular complexity reducing and subjective perspective in combination with prior knowledge drives decision making. Citing Thomas and Znaniecki (1927) from Esser (2002, p. 62):

“And the definition of the situation is a necessary preliminary to any act of the will, for in given conditions and with a given set of attitudes an indefinite plurality of actions is possible, and one definite action can appear only if these conditions are selected, interpreted, and combined in a determined way and if a certain systemization of these attitudes is reached, so that one of them becomes predominant and subordinates the others.”,

Thomas and Znaniecki (1927, p. 68)

New Knowledge is Integrated into Old / Prior Knowledge New knowledge is always integrated into a person’s body of prior knowledge (Roßnagel, 2008, p. 19). Only this integration makes it usable – i.e. accessible via associations or traces (Roßnagel, 2008, p. 44).

The aspect of integrating knowledge into existing knowledge, can also be found in the theory by Anderson (1988) on *associative networks*. He argues that similar to the biological structure of the brain consisting of networks of neurons, information items (or chunks) are stored in the brain in a network that connects chunks of information by association. He supports his theory with various studies concerning short term and long term memory recall as well as their relation to learning strategies and intensity.

Given that the integration process relies on recently received and filtered information as well as prior knowledge, the peculiarities of human memory come to bear:

Using philosophical arguments and cases, O’Donnell et al. (2003) argue with the theory of communicative action by Habermas (1989) that insights from new experience episodes add incrementally to the background knowledge, if and only if the new insights are in-line with the existing body of background knowledge. Exceptions are only situations with drastic experiences, which include undeniable evidence for an alternative truth, which is incompatible with a person’s prior knowledge. This is in-line with Piaget’s learning

²⁴A very famous example of this effect are the events that led to the Challenger space shuttle accident: The problems with gaskets of the rocket motor and low temperature were known before the launch – yet ignored. In hindsight this was a mistake but in the situation the gaskets were only one of many risk factors.

2.3. Literature Findings Integrated into the PIA-Model

model from educational sciences in which new experiences are integrated in existing mental models (*accommodation*) if possible (Clark, 2005; Piaget, 2003) – similar to Sternberg’s *selective comparison* process step. Only cases, in which the new information can not be accommodated in the existing mental models, are new mental models created (*assimilation*) – again comparable to the *selective combination* process step in Sternberg and Hedlund (2002).

Hence adjustments to the background knowledge are mostly small and incremental, an effect that is further strengthened by the perception bias caused by the background knowledge – as discussed in the last section. Yet critical experiences can still change, i.e. replace aspects of the background knowledge (Esser, 2002, p. 62) – albeit in small pieces (O’Donnell et al., 2003) – if the critical experiences contradict with a person’s constructed picture of reality. These changes in the background knowledge may then also lead to a permanent change in people’s behavior.

Also in psychology this updating effect of the background knowledge has been observed: In their study using the Iowa Gambling Task (IGT) with normal participants and amnesia patients, Gupta et al. (2009) found that the amnesia patients performed worse because of their ongoing loss of declarative knowledge (a special part of background knowledge) used for judging the success chances of a card deck. The healthy participants engaged in more incremental updating of their background knowledge regarding different decks and “*were able to draw on these long-term relational representations and stay with advantageous decks even when receiving frequent (if small) punishments.*” (Gupta et al., 2009, p. 1692). The latter finding supports Habermas’ argument that changes to the background knowledge from normal (non-critical) experiences are small and incremental.

With growing age the background knowledge becomes increasingly refined and thus critical events, which can not be accommodated, become less frequently (Piaget, 2003) – coherent with insights from lifespan psychology, which views crystalline intelligence grow at decreasing rate as age increases (Baltes and Staudinger, 1999).

In Piaget’s words:

*“Everything that we learn as children, impedes us to invent or to discover.”*²⁵.

In summary, integrating recently filtered information and prior knowledge into a coherent understanding of the situation is a preparations step towards decision making. The result is new knowledge, which may add or modify the existing body of background knowledge, which by a feedback effect illustrated in the PIA-model (figure 2.1 on page 31) again influences perception and integration.

Note that the model of perception and integration in the PIA-model is slightly simplifying in the sense that it shows perception and integration as two distinct processes,

²⁵Piaget’s quote was translated from the French original from a conversation with J.C.I. Bringuier: “*Tout ce qu’on apprend à l’enfant, on l’empêche de l’inventer ou de le découvrir.*”.

while these two processes should rather be seen as a continuum from perception including increasingly intelligent feature selection to a pure integration of highly filtered information to knowledge (Tsoukas and Vladimirov, 2001).

2.3.4. Prior / Background Knowledge and Perception

Prior Knowledge affects Perception As mentioned before, prior knowledge biases how people filter data. A person's body of prior knowledge is incrementally created over many episodes of experience. In the PIA-model (figure 2.1 on page 31) this is visualized by the feedback loop with steps 4, 1, 2 and 3 respectively.

Using a simple classification exercise²⁶ in a laboratory study with 5-year old children, Siegler and Svetina (2006) observed how children created their own mental models to solve the posed questions (the experimenters did not give any hints on how to solve the problems). Once the children learned a superior mental model, problem solving performance increased drastically. Siegler (2005) observed similar effects with children learning to perform simple arithmetic calculations.

Hence the feature selection (i.e. filtering) process depends on the prior knowledge of people – as is further confirmed in similar ways and in various studies from a broad variety of research fields:

For Butler and Winne (1995), prior knowledge and beliefs about learning and the subject play an important role in the learning process – as modeled in their self-regulated learning model (SLR) – see also figure 2.4 on page 48.

Baltes and Staudinger (1999) model prior knowledge and prior experiences as *Pragmatics* of cognition, which together with the *Mechanics* of cognition – a basic information processing performance – determines the problem-solving performance of people. Their model is based on a model by Cattell (1971), who distinguishes between *crystalline intelligence* (prior knowledge) and *fluid intelligence* (basic cognitive performance). While the *Mechanics* performance peaks with age around 25, Baltes and Staudinger (1999) see the *Pragmatics* of cognition as stable or even increasing beyond age 25. For them it is the clever leveraging of prior knowledge and experience (among a few other strategies), that allow people up until ages around 70 to compensate for their reduced basic cognitive performance.

For Piaget (2003) and Siegler (2005) children learn by creating ever more detailed mental models about problems. Thus problem solving performance is largely based on the mental models the children have learned before. Siegler and Svetina (2006) have observed in a laboratory experiment that 'encoding' of a problem, i.e. non-suitable filtering, is one of the principle causes for children to fail in creating suitable mental models – allowing

²⁶A sample questions for Siegler's classification exercise is: "Are there more dogs or animals in the picture?".

them to attain increased problem solving performance.

In organizational science a community of researchers support the idea of 'situated cognition' i.e. a mode of cognition that draws on mental schemas for interpretation of the raw data of the external (real) world (Elsbach et al., 2005). Hence they have also observed the filtering process and its dependence on prior knowledge.

So far the mental models discussed were mostly conscious: In the field of expertise research Ericsson et al. (2007) has demonstrated that perception can even be trained deliberately to improve performance – even to reach exceptional performance e.g. in chess playing or sports (Ericsson, 2005). Quickly recognizing game situations in chess can be trained to such an extend that they become a skill that is as fast and unconscious like a motor skill²⁷. Thus once conscious skills that depended on prior knowledge can be trained so far that at least conscious use of the prior knowledge is not necessary anymore.

Furthermore, prior knowledge is built up incrementally by exposure to many experience episodes (Carlile and Rebentisch, 2003; D'Eredita and Barreto, 2006b). This implies means that the body of prior knowledge and with it also the filtering mechanism of perception is very individual – with an important implication:

Already by the selection of relevant information, decision making becomes an individual and thus *subjective* activity (Tsoukas, 2005b; Weick, 1993).

Socially Constructed Prior Knowledge In the last paragraphs empirical evidence for prior knowledge driving our filtering i.e. perspective taking process was presented. Since perspective taking depends on an individual and thus unique body of prior knowledge, perspective taking driving perception is individual as well. The following paragraphs will however show that while individual, prior knowledge is not created completely without the influence of other people – e.g. family, colleagues, friends or even society at large:

A number of authors highlight that the prior knowledge that affects our perspective on a problem, is in many cases socially constructed. That means that the prior knowledge is not created mostly independently from others during our own experiences but instead that a large part of this knowledge was created due to stimuli from the interaction with others. Thus prior knowledge is affected by the people we as humans interact with, which includes family and friends but also the organizational environment and society at large (O'Donnell and Henriksen, 2002; Tsoukas, 2005b). Hence organizational as well as national culture may influence our decisions by biasing our perception of the environment and the situation (Starbuck, 2004).

In sociology there is a long standing discussion how society influences people's actions and/or how people shape society. Habermas (1989) describes a *background knowledge of the lifeworld* about society and self, which serves as reference frame for filtering percep-

²⁷An example for a motor skill is bike riding. People use the skill (i.e. ride bikes) without drawing their conscious attention to the task.

tion of the environment. The background knowledge, which is very much comparable with the prior knowledge mentioned before, is built incrementally over many episodes of experiences.

In addition, the incremental build-up and modification of the background knowledge, implies that human perception and thus also behavior depends on one's personal history. Therefore, since background knowledge accumulates over time, the feedback loop of the PIA-model in figure 2.1 on page 31 with steps 3, 4 and 5 shows a dependency of the cognitive system on the person's past. In control systems terms, this makes the entire system controlling human behavior *non-stationary* – i.e. dependent on the past. This property is where Hütther (2006) sees a principle source of human adaptability to a wide variety of circumstances and challenges.

Large parts of Habermas' background knowledge are socially constructed, i.e. the knowledge including norms and values is aligned and synchronized with socially shared norms and values by interaction with others – especially during upbringing but also while working within an organizational culture²⁸. That implies that people's perspectives on problems are very much biased by this socially constructed background knowledge²⁹. Yet it also implies that people may change the shared background knowledge by their willful actions and communication (O'Donnell et al., 2003).

Similar arguments can be found in psychology, where e.g. Anderson and West (1998) argue that climate variables, as used in psychology, measure a certain state of a shared perspective and thus bias the group's actions. Strongly shared mental models have even been shown to support problem solving performance in certain types of tasks (Badke-Schaub et al., 2007; Mathieu et al., 2000).

In summary: As the first step in decision making, humans strongly filter the available data down to a smaller set of relevant information. Concentrating on only relevant information, is essential for effective human behaviour. This filtering or feature selection process is governed by prior or background knowledge, which has been built-up in many experience episodes and is continually refined and challenged by interactions with others and the actual world. Even though the background knowledge may be biased socially by e.g. organizational culture, it remains individual and thus makes decision making subjective by virtue of the subjective feature selection process.

2.3.5. Language for Thought and Diversity in Discussions

²⁸Similar arguments regarding an organizational background knowledge, which in turn affects cognition can be found in Elsbach et al. (2005).

²⁹Dörner et al. (1999) found in his research with micro-worlds (computer simulations of e.g. running a city) that problem solving strategies are dependent on culture: German and younger Indian Managers on average chose different strategies to solve the challenge.

Language in Thought Within the context of perspective taking and integration, language can be seen from two perspectives: 1.) a special body of background knowledge, which is socially constructed and 2.) a special skill that allows us to take a particular perspective on a situation either by thought (Dörner, 2005; Tsoukas, 2005b) or by communicative interaction with others (O'Donnell et al., 2003; Orr, 1996; Rasmussen, 2001). Andriessen (2006) illustrates the effect of language by discussing how metaphors focus our attention on certain aspects while hiding others. For example, when talking about 'a team of specialists', this term hints towards a group of highly competent individuals, who are working in established working practices. In contrast, the term 'specialist work force' suggests that we are dealing with a uniform specialist resource, while the individuality of specialists is deemphasized.

Habermas even sees language as our central tool to create an intersubjective rationality by virtue of the logic inherent in language (Habermas, 1989; O'Donnell et al., 2003). Similarly Schreyögg and Geiger (2007) argue that *"knowledge is constructed in social communication processes"*, p. 83.

These claim are further supported by Hacker and Wetzstein (2004), who demonstrated in an experiment, that reflexive dialogue improves the quality of solutions in technical design – even without an expert partner for the discussion.

The empirical results of Reimann and Dörner (2004) from their laboratory experiment on the effect of self-questioning of engineers during 4 standardized design tasks, indicate that engineers who frequently pose questions to themselves create better technical designs. In particular questions that widen the perspective on the problem, e.g. by analogy to similar problems, and questions on causal effects support the thought process on the design task.

Thus language is not only a tool for communication, but also a tool to take perspective and a support for integrating information resulting in new insights and judgement.

Shared Meaning, Deep Discussions & Boundary Objects Teams work most effectively if they have developed a shared meaning of language and the problem at hand - i.e. if they have developed a shared perspective of the situation (Sandow and Allen, 2005, p. 8). Once a common language and understanding supports team communication, team discussions go deeper and lead to better results (e.g. better technical design solutions) illustrated by the following studies:

Extending the research of Dietrich Dörner on problem solving, Badke-Schaub et al. (2007) in their study on an air-craft accident and mental models of the pilot team argue that language is central to using and shaping mental models. They view mental models as a simplified perspective on the world, which support efficient decision making and also support coordination, when mental models are socially shared:

“These models allow them to integrate new information and to make predictions with little mental effort. Due to their nature, these working models are necessarily simplifications of the world.”,

“When the team members exchange their models in communication, they build up a team mental model.”,

Badke-Schaub et al. (2007, p. 7, 9)

In the same field of research, Bierhals et al. (2007) found in their empirical multi-method study that shared mental models in subgroups within teams reduced the need for explicit communication and improved performance.

Carlile (2002) has found a similar discussion enhancing effect with a 3D (CAD) model, that acts as a shared perspective on the product in development (a new car model in his case study). For Carlile the 3D model is a *boundary object* or shared artifact that facilitates deeper discussions across functional domains in a concurrent engineering project – leading to improved decision making of the group of engineers. For a similar insight see also Dodgson et al. (2007).

After his well known SECI model (see section 2.5.1 on page 69), Nonaka et al. (2000) added to his theory the concept of *Ba* (Japanese for shared context or mental space). Nonaka argues that a *Ba* supports efficient knowledge transfer.

Studying knowledge transfer via social networks (i.e. networks of personal contacts), Hansen (1999) observed that a large number of contacts (*‘weak links’*) increases access to simple knowledge. However when more complex knowledge needs to be transferred, the total number of contacts becomes much less relevant. Instead the number of *‘strong links’*, i.e. intensive contacts with a shared understanding, drives knowledge transfer. Thus Hansen’s empirical finding confirms that shared perspective supports deep and intensive discussions leading to increased knowledge transfer.

Summarizing, creating a common understanding supports deep and productive discussions and can not only be achieved by language but also by shared artifacts as a discussion basis: e.g. a model, a project plan or a schematic drawing. For Dodgson et al. (2007) the *‘design conversation’* supported by visual cues (models, sketches etc.) is an effective working mode for engineers to develop and validate their work in an interdisciplinary manner.

Validating Prior Knowledge in Discussion by Diversity in Perspective Aside from building a common understanding and from exchanging knowledge, discussions also have the effect of challenging each others perspectives on a problem. If the participants are open enough, the discussion can lead to a validation or falsification of the participants’ prior knowledge – by virtue of the strength of the presented arguments.

Colleagues who share a reasonable level of trust (Sandow and Allen, 2005) may pur-

2.3. Literature Findings Integrated into the PIA-Model

posely challenge each other in discussions or conversations to validate their perspectives – as observed in ethnographic field research by Orr (1996). In his study he describes many episodes in which teams of (Xerox) copy machine technicians use narratives (in this case: conversations enriched with stories of past technical problems) to continuously challenge each others' perspective on the technical problem – in order to integrate all relevant clues to a diagnosis and finding a fix:

“Perspective is important in diagnosis. [...] This is one of the reasons that consultations and joint troubleshooting are so popular and effective. It also provides someone to whom stories can be told and who will tell stories in return; the telling of war stories, the consideration of the present with reference to known diagnoses of the past, is an essential part of diagnosis.”,
(Orr, 1996, p. 124)

In Orr's example, language is used as a tool for thought, to build a common understanding of the current as well as past problems and as a medium to challenge and socially align each other's prior knowledge, which shapes the individual perspective on the problem. In the PIA-model (figure 2.1 on page 31), the process is illustrated by link 7 (the communication action itself), a response by the other technician, followed by at least one iteration of the internal perspective taking and integration loop (steps 1,2,3,4 and 5). More details on the learning iteration loops will follow in section 2.3.6 on the facing page.

O'Donnell and Henriksen (2002) argue in a similar manner by claiming that Habermas' lifeworld background knowledge is challenged in discourse by what O'Donnell and Henriksen call '*validity claims*'. When this process of challenging the background knowledge is occurring in a group of people, it may also have the effect of socially re-constructing the background knowledge.

Also citing Habermas' notion of validity claims, Schreyögg and Geiger (2007) advocate a '*discursive understanding of knowledge*' (p. 94), which highlights the knowledge quality improving effects of critical discourse, where this discourse relies on the critical comparison of multiple perspectives on a problem.

Similarly Walsham (2001) argues that comparing different perspectives drives learning. Therefore databases storing a single 'correct' interpretation for each problem or case are much less helpful than a controversial discussion with multiple experts highlighting the difficult points of a topic.

In his study on factors leading U.S. corporations to long-term business success, Collins (2001a) found that an important common property of successful corporations is a culture of opinion diversity and controversial discussion.

Yet diversity in teams also poses challenges: Kearney et al. (2009) argue that diversity in teams (in age, educational background, personality etc.) only supports problem solving,

when the team members share the personality trait that they enjoy engaging in cognitive effortful inspection and discussion of a problem³⁰. Thus diversity in perspective only becomes beneficial, if team members take the time and effort to negotiate a common language and understanding of the problem.

Last but not least, academia is an example of a social institution that evolved into a system, which stresses and cultivates the multiplicity of perspectives by means of publications, citations, peer review and conferences – all with the primary aim of learning (and teaching).

Summarizing, team members with a diversity of perspectives can engage in critical discussions, which by comparison of the perspectives and their implications bring implicit assumptions and subtle differences to the surface. Such deep discussions allow creation, refinement and validation of shared body of prior knowledge regarding the problem at hand.

2.3.6. Iterative Learning modelled with Feedback Loops

Feedback Loops in the PIA-Model On the analysis level of a single learning episode, learning is an iterative activity in which the learner iterates within two feedback loops³¹:

- An **internal loop**, illustrated by steps 1, 2, 3, 4 and 5 in the PIA-model (figure 2.1 on page 31), which would model more passive learning activities such as learning, while reading a text book. The internal loop leads to the incremental addition or modification of the learner's body of prior knowledge.
- An **external loop**, which includes the internal loop (steps 1 – 5) and additionally includes a decision and action (steps 6 and 7) that interacts with and thus feeds back to the actual world. The action could be in the form of a small experiment or trial³², communication (e.g. a discussion) or a searching activity (see section 2.3.7 on page 51). When a visualization technique is used as illustrated in the PIA-model with visualization (figure 2.3 on page 36), then the person may also iteratively fine-tune the visualization based on his or her growing understanding (step 8).

³⁰Team diversity is helpful if team members have a 'need for cognition' – see Cacioppo and Petty (1996). Then the team members are likely to overcome the negative effects of diversity, e.g. lack of common understanding.

³¹With these two feedback loop the PIA-model at first sight deceptively likens a control system for a machinery system from control system engineering – e.g. a cascaded double-loop air-conditioning room-temperature control system). The engineering minded reader should note however that in contrast to most technical control systems, which are stationary (i.e. have a current state, described by a few variables, which together with the governing equations completely determines the system's behavior), the human cognitive system as modelled in the PIA-model has a history and its behavior depends on this history and the prior knowledge, which was created in the course of this history.

³²An example 'experiment' or trial would be Orr's copy machine technicians exchanging a part with a spare part in order to test whether this part caused the machine's malfunction (Orr, 1996).

2.3. Literature Findings Integrated into the PIA-Model

Thus external loop activities involve actively probing the actual world in addition to the refinement of the body of prior knowledge from the first loop.

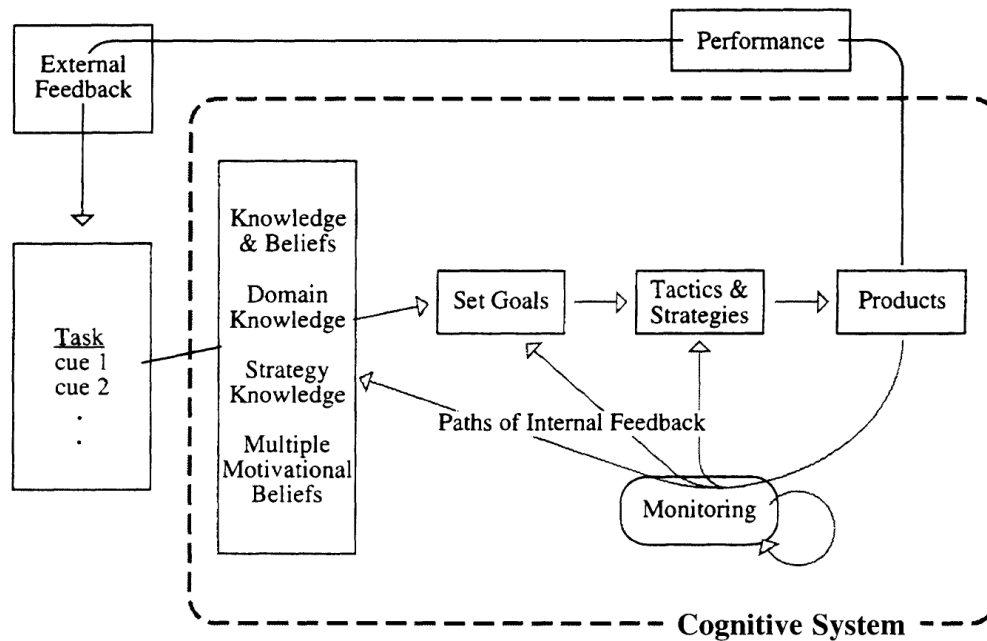


Figure 2.4.: Self-Regulated Learning (SRL) Model (Source: [Butler and Winne \(1995, p. 248\)](#))

Similar Double-Loop Feedback Effects in Literature [Butler and Winne \(1995\)](#) model academic learning in a similar iterative manner with their *Self-Regulated Learning (SRL) model* as shown in figure 2.4. Like the PIA-model, their model contains a body of prior knowledge (in their case split up into four categories) that has an initial state and is modified iteratively as the learning episode progresses. As in the PIA-model, there is an internal feedback loop that directly feeds the learner's self-assessment of the learning progress back into the prior knowledge affecting the learner's goals and strategies. In addition there is an external feedback loop outside of the learner's cognitive apparatus, which is an external learning performance assessment – e.g. an exam ([Butler and Winne, 1995, p. 248](#)). Similarly but more simplified [Goldberg and Cole \(2002\)](#) model the learning rate in schools as a feedback system that involves student motivation, academic performance and the teacher's expectations of the student.

[D'Eredita and Barreto \(2006b\)](#) describe learning as episodic activity: i.e. learning takes place iteratively in many episodes (learning experiences) and creates knowledge that is connected to that episode. The connection to past episodes is recalled upon attention to the subject. [Racsmany and Conway \(2006\)](#) found a similar episodic structure of knowledge

in a series of six psychological experiments.

When only learning activities such as reading, searching or discussion are considered, the SRL-model is sufficiently similar to the PIA-model allowing the transfer of some of insights applicable to the SRL-model: Drawing on a number of other studies and their model, Butler and Winne argue that the continuation of the iterative learning process by the learner is dependent on the learning progress. If the learning progress is at (or above) the learner's expectations the learning process continue with further iterations. When the progress begins to fall short of the expectations, the learner is likely to adjust his or her learning strategies. But if that does not improve the learning progress over several iterations, the motivation of the learner will become depressed, which may cause the learner even to completely disengage from the learning task.

Thus correspondingly in organizational work settings, knowledge workers are unlikely to continue with a learning effort for many iterations, if there is no substantial learning progress – in-line with their expectations – visible to them. This feedback mechanism has an important implication:

Those who learn successfully in a particular situation, are likely to be motivated to learn even more, while those who do not see progress at any point within a learning episode, are bound to completely abandon the learning effort. In general this feedback effect is situational and depends on the person but also on many situational factors – e.g. the availability of relevant information in a particular case. However if one takes into account that some employees by personality, skill or prior experiences are more open to on-the-job learning than others, even a small difference in this openness will be amplified by the described feedback effect and will lead to either substantially supporting or attenuating learning³³.

Other authors echo this insight: Roßnagel (2008) argues that older employees in particular may have lost practice for the skill of learning in formal learning settings, since they have gone through their formal education many years ago and also because corporations send older employees less frequently to seminars. Hence older employees on average have a reduced self-efficacy regarding formal learning before attending a seminar. He suggests to build-up self-efficacy by targeted training in order to avoid a self-fulfilling prophecy regarding the learning performance.

Further support can be found in the field of problem solving psychology, where Dörner et al. (1999) describes problem solving as a feedback loop, which is similar to the external feedback loop of the PIA-model and consists of the following phases:

³³From the engineering theory of control system, a self-reinforcing effect will either reinforce itself without limits leading to something like a resonance catastrophe or cause a system (such as an operational amplifier) to saturate. Many engineering systems are however also limited by other factors (e.g. increasing resistive forces) or the process is limited by a new kind of bottle neck that has become effective. Similarly learning will not spin out of control either but rather other factors will limit learning, e.g. more difficulties to find good information.

2.3. Literature Findings Integrated into the PIA-Model

1. Defining Aims
2. Collecting Information and creating (mental) Models [equivalent to the internal feedback loop from the PIA-model]
3. Prognosis (using the information and prior knowledge)
4. Planning and Decision Making
5. Feedback
6. Self-Reflection (adjusting the mental model) [this is the external loop back to step 2]

Thus Dörner's model also contains two cascaded feedback loops that affect cognition mediated by the adjustment of mental models with internal and external stimuli – similar to the internal and external feedback loop in the PIA-model (figure 2.1 on page 31).

Iterative Learning with Multiple Actors The iterative learning process, can also take place jointly with multiple actors: Orr (1996) describes 'diagnosis' (i.e. learning to understand the cause of a technical problem) as an iterative process that involves repeated adjustment of the copy machine technicians' perspective on the problem eventually leading to a solution. Similar to the PIA-model he describes two processes that are equivalent to the internal and external feedback loops of the PIA-model: In the diagnosis process a pair of technicians jointly create hypotheses about the cause of the copy machine's malfunction. To test these hypotheses, the technicians frequently exchange parts, that are the suspected cause, against a spare part. This is an example for the external loop from the PIA-model. But Orr also emphasizes the importance of another activity in diagnosis: the telling of narratives about earlier problems. By talking about an earlier problem, one technician offers a particular perspective on the available data of the problem³⁴ to the other technician. The teller of the narrative is acting with communication and thus employs the external loop of the PIA-model, while the listener passively compares the offered perspective against his own perspective and thus employs the internal feedback loop from the PIA-model.

Summarizing, the PIA-model contains an external and an internal feedback loop that allows describing learning as an iterative process within these loops. This description of iterative learning is consistent and similar to various alternative descriptions in literature. By the nature of feedback loops, the iterative learning process may amplify or attenuate

³⁴Here the 'data of the problem' is any data of the problem that the technicians have already obtained and reviewed and thus is directly available to them in their thought process. Examples are observations of the technicians about the state of the machine as well as reports of the malfunction by the users of the machine.

itself depending on only small differences in the initial and boundary conditions – i.e. the predisposition of a person for on-the-job learning or an actual learning condition.

The amplification and attenuation effects make the dependency of learning on personal disposition and situational factors non-linear.

2.3.7. Searching for Information

Searching and Feedback In the last section learning was modelled with the PIA-model (figure 2.1 on page 31) as an iterative process with an internal and an external feedback loop. The external loop symbolizes actions such as experiments (or trials), communication and activities to deliberately gain more relevant information – i.e. searching. Search activities may include, searching of databases but also leveraging personal social networks³⁵ to get information from others.

As with the learning activity as a whole, the level of searching activity is subject to the feedback effect as well: if the learner during a particular learning episode experiences searching as helpful, he or she is likely to engage in further searching activity, while upon negative experiences with searching, knowledge workers are likely to abandon the searching efforts and to seek alternative strategies. If even the revised learning strategies lead to an alternative support of learning, the learner is likely to disengage with the learning task.

Chiou and Wan (2007) operationalized the learners expectations about the usefulness of searching for him or her with a common psychological construct: *self-efficacy*. In their empirical study on internet searching, they found that searching self-efficacy decreases dynamically upon encounter of positive and negative search performance results during the course of a searching episode. Self-efficacy is an important factor, affecting whether a learner decides to do a search (Debowski et al., 2001), choose alternative strategies or to stop learning³⁶. Thus the findings of Chiou and Wan (2007) confirm the application of the cascaded feedback PIA-model to predict the dynamics of searching – incl. the amplification and attenuation effect of positive and negative search experiences.

Whether search is perceived as helpful, i.e. whether people maintain and build self-efficacy for searching during the learning episode, depends also on the searching strategies: i.e. the questions that a learner asks other people or the search terms a learner uses. If the learner is progressively learning during the learning episode, he or she will also continuously adjust his or her perspective on the problem and through the new perspective

³⁵In the following the term ‘*social network*’ labels any network of personal acquaintances and is not limited to only recently popular web-based communities.

³⁶Bandura (1997) describes self-efficacy as a dynamic phenomenon, which can be ‘trained’ by approaches such as ‘guided mastery’. Within the PIA-model self-efficacy would be a special part of a persons’ prior knowledge, reflecting the persons prior experiences and expectations regarding a particular task. Like any other part of the prior knowledge, it can be modified gradually over time, which is consistent with Bandura’s observations and suggestions. Thus self-efficacy could be seen as a operationalization of a special part of the prior knowledge as a psychometric construct and therefore represents a form of complexity reduction.

the learner will see new gaps of information and thus continuously be able to ask new questions³⁷. Thus searching and learning may support each other vice versa: with out a progressing learning effect, searching becomes difficult and without successful searching, learning becomes difficult.

More on searching and knowledge management, which is not essential to the line of argument of this theory chapter, is covered in appendix section [A.2 on page 288](#).

2.3.8. Tacit Knowledge and Implicit Learning

The process of integrating filtered data in order to construct new mental models (knowledge) can be conscious as well as unconscious (Scott and Dienes, 2008; Siegler, 2000). This effect is frequently referred to as explicit (conscious) vs. tacit (unconscious, non-verbal) knowledge or explicit vs. implicit learning.

Especially the research stream knowledge management (KM) focuses on the challenges that emerge from attempting to ‘transfer’ tacit knowledge from one person to another – see section [2.5.1 on page 69](#). For many KM researchers tacit knowledge is one of the principal barriers for knowledge transfer (Rolf, 2004; Schreyögg and Geiger, 2005). Of those KM scholars a fraction further aims to overcome this challenge primarily by externalization (i.e. conversion of tacit to explicit knowledge) – see e.g. Abou-Zeid (2002); Coffey and Hoffman (2003); Hofer-Alfeis (2000); Nonaka and Takeuchi (1995); North (2002).

In this section the current state of research regarding tacit knowledge from different fields will be outlined, which will show that the phenomena associated with tacit knowledge or implicit learning are not well understood in detail yet. There is however evidence regarding two relevant issues: a) implicit learning rarely happens in complete isolation from explicit learning and b) tacit knowledge or an implicitly learned skill can be ‘transferred’ or taught to other people by interactive methods (e.g. working together or joint practicing). Given these two insights and tacit knowledge and the previously outlined challenges surrounding learning, I therefore claim:

- a) The challenge of tacitness is in many cases overshadowed by and connected to the general challenge of learning. Thus research on organizational improvement measures should focus on learning in general first and address the implicit learning challenge with interactive and social knowledge sharing approaches – as already researched and suggested in many earlier studies.
- b) Since implicit and explicit learning mostly happens in a mixture form, detecting and measuring explicit learning can be used as a proxy for the entire learning effect – with both explicit and implicit parts.

³⁷The copy machine technicians in Orr (1996) always think of new places in the machine to inspect after reshaping and adjusting their perspectives on the problem (in their case predominantly by telling narratives).

This claim is further detailed in the following elaborations:

Implicit Learning under Scrutiny from Different Fields From a learning psychological perspective, conscious and unconscious learning is referred to as *explicit* and *implicit* learning. Kuhn and Dienes (2005) demonstrated using an experiment involving the recognition of sequences of musical tones, that implicit learning is more than just memorizing chunks of information in an unconscious manner. After a training session, the participants were able to classify tone sequences by using an implicitly learned rule.

Along these lines, Dienes and Perner (1999) and Pothos (2007), citing results from experiments with artificial grammar (AGL), argue that the information integration process leading to an understanding, e.g. of a mechanism or a situation, is not always conscious but can also be associative: While explicit learning involves a conscious and deliberate learning effort, implicit learning – leading to tacit knowledge – can occur unconsciously during episodes of experience. Dienes and Perner (1999) further argue that implicit learning creates associative “first-order connectionist networks” – similar to those described earlier by Anderson (1988). Supporting the theory of associative neural networks, (Sun et al., 2005) with a neural network computer simulation predict explicit and implicit learning behavior in more detail.

Other types of implicit learning, e.g. the implicit learning of hidden covariation detection (HCD)³⁸, have found to be weaker than first expected at discovery in the field of psychology (Roßnagel, 2001).

Given the unconscious nature of implicit learning, it does not come as a surprise that a property of this form of learning is that people are usually not able to verbalize what they have learned or to recall the learning episodes (Manier et al., 2004; Nonaka and Takeuchi, 1995; Tsoukas, 2005b). Frequently this effect appears in the form that, when an expert tries to explain something, his explanation is hard to understand because the speaker bases his arguments on many tacit assumptions and insights – which are not equally obvious to the listeners³⁹.

In the field of knowledge management, much of the discussion focused on a similar distinction between *explicit* and *tacit knowledge* by Polanyi (1966), which was later popularized in the management literature by Nonaka (1991) – see also section 2.5.1 on

³⁸In an experiment Roßnagel (2001) has shown a series of pictures of long haired and short haired women to participants and described them as kind and capable respectively. The participants were however not given any direct hint about the covariation of long hair with kind and short hair with capable. In earlier experiments other authors showed that participants implicitly learned the covariation. Yet in more recent experiments, Roßnagel and others found the implicit learning effect to be rather weak (Roßnagel, 2001).

³⁹The reader may think of some professors, who have a hard time teaching their students, since they have lost the understanding for what could be difficult to a newcomer. With many years of experience and practice, for them some of the basics have just become intuitively obvious – i.e. tacit – and thus not worth to mention or actively think about. It is actually part of the professors’ skill to not have a need to think about the basics – which frees cognitive capacity for other focal points.

page 69. An insightful and recent discussion of Polanyi's conception can be found in Tsoukas (2005b).

Polanyi illustrates the nature of tacit knowledge with his famous 'blind man's stick' example: When a blind man uses a stick to probe his environment, he focuses his attention entirely on the tip of the stick in order to learn more about the objects in his environment, which the tip touches. If the man wants, he can also focus on how the stick feels in his hand but that would only distract him from his original task: exploration of the environment. The blind man has developed a high skill level in using his stick as a tool for exploration – *"making it feel as if it [the tool] is an extension of [his] own body (Polanyi and Prosch, 1975)"* (Tsoukas, 2005b, p. 127).

This principle applies not only to physical tools but also to intellectual tools: e.g. when we as humans speak, we don't think of the rules of grammar but focus on what we want to express. *"As we learn to use a tool, any tool [also an intellectual tool], we gradually become unaware of how we use it to achieve results."* (Tsoukas, 2005b, p. 127).

These examples illustrate Polanyi's conception the to-from structure of tacit knowledge: "tacit knowing requires three elements: subsidiary particulars, a focal target, and a person who links the two." (Tsoukas, 2005b, p. 103). In the language example, the rules of grammar are the subsidiary particulars (i.e. the tacit knowledge), which the person leverages in order to attend to a focal target – here the contents of the expression.

Tsoukas goes even further by arguing:

"We must [...] learn to rely [...] subsidiarily on particulars to attend to something else, hence our knowledge of them [the subsidiaries] remains tacit." (Tsoukas, 2005b, p. 147)

and therefore:

"We achieve competence, by becoming unaware of how we do so." (Tsoukas, 2005b, p. 150)

Thus for Polanyi and Tsoukas, most of our actions involve the intense use of tacit knowledge. Hence a large body of tacit knowledge, like that used by an experienced practitioner, enables particularly effective behavior: *"a practitioner's ability to follow rules is grounded on an unarticulated background."* (Tsoukas, 2005b, p. 103). This argument is also supported by Schön (1992) for whom tacit knowledge is acquired and practiced in action (p. 176).

This further implies that we have many skills in both tacit and explicit forms. For example, most of us will know how to switch gears in a manual transmission car consciously by the following steps: 1.) press the clutch, 2.) operate the gearshift lever ("the stick"), 3.) release the clutch. Yet most of us will rely much more efficiently on their tacit gear shifting skill and just think consciously: "gear up!" or even simpler just: "go there!". Most

people will also experience reduced gear-shifting performance, when trying to consciously control the gear shifting process. The practice in gear-shifting (as a motor task) is not in the conscious knowledge about the process but in the implicit skill.

The insights by Polanyi and Tsoukas are based on examples, cases and deductive reasoning. Yet the described properties of tacit knowledge are also being discussed in psychology:

For Sternberg, tacit knowledge is one of the most important factors in practical intelligence – i.e. an intelligence measure that predicts on-the-job performance (Sternberg, 1997; Sternberg and Hedlund, 2002). Therefore he and some other scholars measure tacit knowledge using Sternberg’s *tacit knowledge inventory for managers (TKIM)* survey scale (Colonia-Willner, 1999) by how well managers choose from a variety of possible strategies for a number of realistic mini-cases⁴⁰.

Regarding the high efficiency of tacit knowledge for decision making, some studies observed higher recognition speed, when the recognition was based on a tacit skill rather than an explicit recognition process (Boldini et al., 2004; Hintzman and Caulton, 1997). Yet these findings could so far not be confirmed without contradictions (Dewhurst et al., 2006).

In the neural sciences, the study by Jaeggi et al. (2007) shows results that high performers in a dual perception and decision task, keep the neural activation levels low, while the low performers show increased neural activation levels during periods of task overload (Activation levels are measured by MRI). This could be a hint that the high performers leverage tacit knowledge more efficiently (with a lower level of activation) and thus keep more cognitive free for concentrating on the focal task.

However, overall, the link between the use of tacit knowledge and decision making performance compared to degree of explicit knowledge usage, appears to be a topic of ongoing research (Dewhurst et al., 2006).

In addition tacit knowledge is special when it comes to knowledge justification – i.e. conscious and deliberate validation of knowledge. Tacit knowledge by its unconscious nature escapes conscious and deliberate testing and validation (Schreyögg and Geiger, 2005) – e.g. using deductive logic (Tsoukas, 2005b). It remains unclear how and if tacit knowledge is validated.

⁴⁰In Sternberg’s *tacit knowledge inventory for managers (TKIM)* survey scale, 9 mini-scenarios, which represent common situations in the business world, are presented to managers by a description that is about 5 sentences long. For each of these mini-scenarios managers rate the quality and effectiveness of about 10 strategies to handle the situation (Colonia-Willner, 1999). The instruction is to ‘scan’ the strategies and rate them. Thus given the short rating time and the brevity of the description, it is likely that managers give an intuitive answer – i.e. an answer expressing their tacit skills for business situations. Yet it remains unclear how Sternberg and other scholars like Willner ensure that answers are truly intuitive rather than based on conscious reasoning. Alternatively they need to follow Polanyi with the assumption that decisions are mostly based on tacit knowledge and only on a small fraction based on explicit knowledge or cognitive performance.

Interaction of Implicit and Explicit Learning Another important facet of implicit and explicit learning is that there is evidence that neither occurs in complete isolation from the other but always in a mixture with varying degrees of implicitness and explicitness (Schreyögg and Geiger, 2005; Tsoukas, 2005b; Wong and Radcliffe, 2000).

Based on an experiment involving children learning basic arithmetic operations, Siegler (2000) demonstrated that insights on how to solve a problem in some cases first surfaced in an unconscious (i.e. tacit) form. The children only later became able to explain the insight – i.e. verbalize the explicit skill. Another form of implicit to explicit knowledge conversion was detected by Fischer et al. (2006), who observed increased formation of explicit knowledge for a group of participants that received a special treatment: a night-long sleeping break between tests.

Based on theory from psychology and simulated using a neural network computer model with an explicit and implicit learning part, Sun et al. (2005) argues that implicit and explicit learning happens always in conjunction.

The fact that rates of forgetting are similar for explicit and tacit knowledge (McBride and Doshier, 1997) in addition to the previously mentioned findings regarding conversion of tacit to explicit knowledge suggests that tacit and explicit may be stored in the same neurons in the brain and just differ by the form of connection to the neurons and thus their accessibility – similar to the concept by (Anderson, 1988) in which working memory and long term memory differ only by their activation state. Yet this hypothesis requires further investigation.

Implications of Tacit Knowledge for this Study Summarizing the literature on tacit knowledge and implicit learning: the PIA-model in figure 2.1 on page 31 would need to be extended by two modes of integrating information and two modes of prior knowledge: implicit or explicit learning leading to tacit or explicit knowledge. In normal practice knowledge workers unconsciously leverage a large body of tacit knowledge in addition to a usually much smaller portion of explicit knowledge, which is used consciously for decision making. The mechanisms how explicit and tacit knowledge is converted into the other form in the human brain are so far not well understood. Yet there are many indications that learning is rarely purely explicit or implicit but usually occurs as a mixture of both.

The unit of analysis of this study is on the level of an individual employee and not on the level of episodic micro-processes in a person's brain. Therefore understanding every detail of the mechanisms surrounding tacit knowledge is not necessary to learn about the effect of the organizational environment on learning in general. Thus I do not see the necessity to model the unclear interactions of implicit and explicit learning for the purpose of this study in the PIA-model and see this decision as an acceptable measure of complexity reduction for the purpose of this study.

As many authors stress, transferring tacit knowledge poses a challenge – yet **the**

challenge of tacitness is in many cases overshadowed by and connected with the general challenge of learning. When a specialist has a tacit skill, which he needs to pass on, the skill's tacitness limits the specialist in his ability to support the learner's learning process. Yet by demonstration of the skill (joint practice), the specialist may still draw the learners attention to the relevant information, which the learner can then integrate and acquire the skill – usually in a mixed form of tacit and explicit knowledge. It depends on the situation, whether the reduced teaching ability of the specialist or the general challenge of learning the skill is the most limiting factor for knowledge transfer.

For on-the-job learning in problem solving situations, tacitness of a specialist's knowledge is even less important, since only a fraction (albeit a large fraction) of all learning activity is with another person acting as an informal teacher or learning companion.

Tacit components certainly becomes a problem in the special case, when somebody tries to capture knowledge of specialists in a document or a database. However the challenge regarding tacitness is very much reduced and frequently overcome, when personal interaction is used in the transfer of the skill (from tacit to tacit). The latter recommendation is far from new and echoed by a large group of scholars (D'Eredita and Barreto, 2006b; Fahey and Prusak, 1998; Hansen et al., 1999; Nonaka et al., 2000; O'Donnell et al., 2003; Orr, 1996; Reagans et al., 2005; Sandow and Allen, 2005; Schreyögg and Geiger, 2005; Tsoukas and Vladimirov, 2001).

Nevertheless in all cases a big part of the challenge is the learning process. The challenge of tacit knowledge becomes effective only indirectly and acts on the learning process by reducing the learning support that the 'teacher' may give to the learner – in person or via documents.

Therefore to go a step beyond the current state of research and add value, the focus of this study is on the other (and frequently more important) barrier of knowledge transfer: the learning process (Jacobson and Prusak, 2006).

2.4. Using the PIA-Model to Explain Industrial Practice Models

A number of practices used in various industries, utilize the effects described in the the PIA-model (figure 2.1 on page 31). Practice based management models such as the European Framework for Quality Management (EFQM), modern project management and the Toyota production system are discussed in this section as examples of leveraging the modeled effects.

Due to their practical effectiveness, all three industrial practice models received wide spread acceptance – in industry as well as the engineering sciences, where all three practices have found their way from industry onto the research agenda (Goldratt, 1997; Serrano

2.4. Using the PIA-Model to Explain Industrial Practice Models

et al., 2008; Walgenbach and Beck, 2000). The primary difference of these practice models to other management theories is that these models were initially developed from practical insight rather than theory.

2.4.1. Commonalities of Industrial Practice Models

The three industrial practice models have the following principle commonalities:

1. **Create a shared understanding first.** Before searching for solutions, all three practices first mandate spending time on analyzing and creating a shared understanding of the current situation (status) with respect to the organization's aims. This combined analysis and teaching effect is achieved by **visualization** of the situation in a framework model.
2. **Engage in deep discussions regarding the solution.** The shared, systematic and simplified visualization of the situation (based on the model) allows all participating organizational agents to engage in a deep and constructive discussion with the aim to find and agree on a way or to get from the current state to the common goal.
3. **Repeat and iteratively improve using feedback loops.** Steps 1 and 2 are iteratively repeated to improve both the suitability of the shared perspective and the solution.

How these principle commonalities link to the perception and integration model, will be discussed later in section [2.4.5 on page 66](#) after describing the properties of the three mentioned industrial practice models in more detail. The reader may recall the effect of visualization illustrated with the PIA-model in section and figure [2.3 on page 36](#).

2.4.2. Project Management

In project management, a complex and interlinked activity with many actors is broken down into a finite number of activities – referred to as the *work break-down structure (WBS)* (Wysocki, 2006, Chpt. 4) or as the *activity definition* (Project Management Institute, 1996, Chpt. 6). Each activity in the project plan is an abstract and complexity reducing representation of a group of smaller tasks that are related, e.g. a single person within a time frame is responsible (Wysocki, 2006, Chpt. 1).

Next, the activities are put in sequence and the web of the most important dependencies are visualized e.g. by a PERT chart (for an example see figure [2.5 on the next page](#)). Since not all but only the most important dependencies are modelled, also this step is leading to complexity reducing representation. It follows the estimation of the duration of the individual activities and a back-planning from a due or delivery date taking into

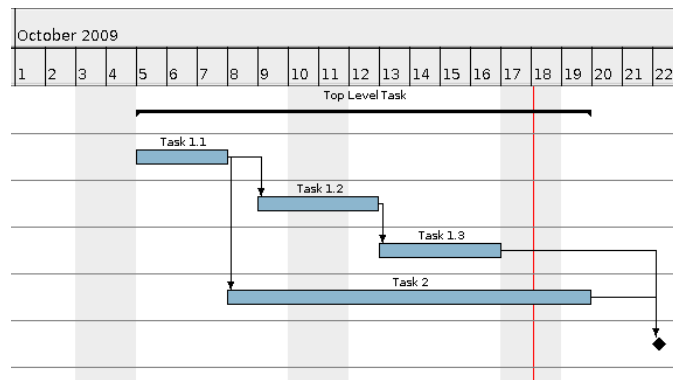


Figure 2.5.: Gantt Chart Example (Source: Author)

account resource limitations and leading to due dates for all individual activities (Project Management Institute, 1996, Chpt. 6).

The Gantt chart is however only the tangible outcome of the three initial planning steps: Very important and valuable is also the shared understanding about the dependencies and the expected duration of the activities plus an agreement on when to work on which activity in a coordinated manner. The resulting plan is effectively an agreement of all participants about the overall and intermediate goals of the project. Important for the project manager is furthermore a visualization of the *critical path*, which is the longest sequence of activities, which determine the overall project duration (in the example in figure 2.5 the critical path is the sequence: “Task 1.1”, “Task 2”). In conventional project management, the project manager will focus his actions on this critical path (Project Management Institute, 1996; Rand, 2000).

Given the importance of a suitable perspective on the project including a suitable focus of the project manager on the critical activities, Goldratt (1997) advocates for an alternative to the critical path: the *critical chain* – which takes into account resource-constraints and psychological aspects of deadlines⁴¹. The value of his approach derives largely from the novel perspective that he offers. Even though Goldratt does not present any empirical evidence, the novel perspective has been intriguing enough to start a new line of discussion including case studies of industry applications (Best, 2006; Bevilacqua et al., 2009; Yang, 2007) and theoretical or empirical analyses (Herroelen et al., 2002; Raz, 2003; Steyn, 2001).

After the project has started, modern project management approaches include a status monitoring process, with which the actual progress is compared with the planned progress in addition to a continually updated forecast for the required remaining time in the active activities. The status can be a simple information such as “X% complete”, or the current

⁴¹To fight the “student syndrome” leading to procrastination, when there is much safety hidden in activities, Goldratt advocates the use of buffers at strategic locations.

2.4. Using the PIA-Model to Explain Industrial Practice Models

state plus a prediction⁴² about the activity completion date, e.g. “*The status gets a yellow light, since we are currently lagging behind in the activity and need more resources in order to meet the deadline. Yet with more resources we expect, that meeting the activity’s deadline is still feasible.*”⁴³. Both types of status reports can be aggregated, e.g. across departments, via statistical or qualitative methods – leading to a picture of the project state with systematically reduced complexity. As an alternative to this conventional approach Goldratt (1997) instead suggests to monitor the actual completion times and the resulting state of the various project buffers.

The purpose of the project status is to effectively support a discussion with the aim to decide on corrective actions⁴⁴ early in order to bring late activities back on track or to reschedule affected activities leading to make-up plans. Again the systematically generated status information is only a starting point for the discussion to which individual and possibly richer perspectives on the project’s state may be added and used in conjunction for robust decision making.

Since the status and predictions about the activity completion dates are continually updated, project management approaches with this updating feature, effectively include an internal feedback loop that continually refines the predictions and corrective actions.

The structuration and visualization of the project activities and additionally of the actual project progression, facilitates learning across projects, allowing to continually refine the systematics to control the progress of projects⁴⁵.

In summary, project management methodologies offer different ways of visualizing the scope of the project work and the current project status, supporting effective decisions on corrective actions to keep the project on course. As the comparison to the alternative critical chain approach by Goldratt (1997) shows, different perspectives on the project are conceivable and given their importance are the topic of a intense debate.

2.4.3. The Toyota Production System

The Toyota Production System (TPS) is mentioned here as an example, since it has gained importance beyond its original application at Toyota. Many other companies apply principles of the system under the headings *lean production / manufacturing*, *Just-in-time Production* or *Continuous Improvement (Kaizen)*.

⁴²See also activity 10.3 ‘Performance Reporting’ from the Rev 3 PMI framework (Project Management Institute, 1996, Chpt. 10).

⁴³Wysocki (2006) describes 5 types of status reports in Chpt. 10 on p. 321ff.

⁴⁴One of the simplest, yet not always effective, corrective actions is putting more man-power on an activity.

⁴⁵While the organizational actors using a project management methodology will have learning experiences, which they transfer across projects, most project management approaches do not systematically support across-project learning. Hence there is no explicit long term focus in project management – in contrast to EFQM of section 2.4.4 on page 62 and the TPS of section 2.4.3.

The system was developed in the 1950s by Taiichi Ohno, Shigeo Shingo and Eiji Toyoda for mass producing cars with finite customizations based on the work of W. Edwards Deming, whose ideas (Deming, 1985) eventually lead to the TQM (Total Quality Management) approach, and Walter A. Shewhart (the prime father of *Statistical Process Control* (Devor et al., 1992; Shewhart, 1931)) (Liker, 2004).

The system consist of a few principles on how to design flow production systems as well as management principles that need to be applied with concrete methods and processes for the individual organization, plant or department. Thus one strength of these principles is their abstract nature that allow for application and utility in many different settings.

The production principles frequently have two sides: a perspective on the production process and linked hints on how to improve production. For example, processes should be designed according to the following principles (labelled with the original Japanese technical terms):

- **Muri** – Decompose a complex process into its simplest parts [process perspective] and **standardize** these process parts in order to be able to monitor and continuously improve the standardized process steps [improvement action].
- **Mura** – Make the entire process lean – i.e. look for unnecessary inventories [flow perspective] and reduce them as much as possible – e.g. by stabilizing the previous process step using statistical process control with *key performance indicators (KPIs)*⁴⁶ [improvement action].
- **Muda** – Elimination of any wasteful activity and focus on the value creating activities. To support finding wasteful activities, the TPS has 7 categories of waste: 1.) over-production, 2.) unnecessary motion, 3.) waiting, 4.) excessive transport of parts, 5.) over-processing, 6.) inventory 7.) rework and scrap [value stream perspective] (Liker, 2004, p. 28). How the different kinds of waste can be eliminated depends on the actual process [improvement action]. The progress of improvement can be tracked with KPIs such as actual work time over cycle time.

These three principles all have in common, that the create a new level of transparency on the processes - either by an abstract model of the work in process steps or by complexity reducing KPIs that reflect process performance in a few numbers. Thus new transparency is created by a shared and simplified perspective on a complex process. Transparency and visualization are frequently used techniques of the TPS: An example is the guidance for part design, which should aim at making any misalignment or missing components directly and visually obvious. Another example are the *Andon boards* at central locations

⁴⁶Key performance indicators or KPIs are numerical measures that indicate the performance of a process in a numerical and therefore simplified manner. Examples are the average throughput time of a certain type of parts, defect rates, yield rate, work hours per part or profit per part.

2.4. Using the PIA-Model to Explain Industrial Practice Models

on the shop floor of a flow production line, which visualize the current state of the process to all workers.

It should be noted, that if the shared perspective is labelled as simplified here, it is simplified with respect to the full complexity of the process. Commonly this new systematic shared perspective is more detailed than what can be exchanged easily in discussions without such a common perspective.

The newly agreed upon shared perspective is not only used to describe the status in a detailed way but also to discuss and agree on targets. Especially KPIs lend themselves for target setting but also for monitoring improvement, which is important either to facilitate feedback – in case of failure – and equally important – in case of success – to applaud the participants.

In the scientific community, especially the value stream perspective focusing on wasteful activity has received attention – under the name *value stream mapping*⁴⁷ (Bevilacqua et al., 2008; Lian and Van Landeghem, 2007; Serrano et al., 2008).

Hence, in summary, much of the strength of the TPS derives from using a simplifying but suitably focusing model to create a shared and simplified perspective on a complex process. The TPS contains only very few prescriptive and specific hints for process improvement, since many of the suggestions in TPS are very general and can also be found in other manufacturing optimization approaches. The shared perspective and the general nature of the hints support the development of domain specific and therefore highly specialized and effective solutions. The effectiveness of the process improvements are continually monitored and refined along with the shared perspective in a feedback approach called *Kaizen*.

2.4.4. EFQM

Until the 1960s, quality engineers focused on classical statistical process control (Shewhart, 1931) with applications purely in production and the ISO 9001 series of standards (Wilkinson and Dale, 1999). Yet in the 1960s, quality engineers began to realize that many drivers of quality are located outside of the manufacturing shop floor (e.g. the relationships to suppliers, human resource management, strategic management, etc.) (Walgenbach and Beck, 2000). *Total Quality Management (TQM)* systems began to evolve.

Inspired by ideas that form the basis of the Toyota Production System (TPS) and the *Baldrige Award* (Evans and Jack, 2003; Sims et al., 1992) in the U.S.⁴⁸, the EFQM organization was founded as a non-profit organization in 1988 by 14 larger European corporations such as British Telecom, Volkswagen, Dassault Aviation and Philips (European

⁴⁷in German: “Wertstromanalyse”

⁴⁸Following the successes of Japanese firms with statistical process control based on the work of Shewhart (1931) and Deming (1985). An initiative was started to promote total quality management to the U.S. industry with the Baldrige Award for outstanding U.S. implementations of total quality management.

Foundation for Quality Management, 2003) as a European TQM organization. Since then EFQM has grown to over 800 members and focuses on training auditors (EFQM assessors) and the *European Quality Award (EQA)* (European Foundation for Quality Management, 2003).

Hence the EFQM model was initially created from industrial practice and, due to its popularity⁴⁹ and the support of the European commission (Walgenbach and Beck, 2000), only later (in the 1980s) began to attract scientific interest and investigation – e.g. Bou-Llusar et al. (2009); Burkhard (2006); Ehrlich (2006); Evans and Jack (2003); Kujala and Lillrank (2004); Rusjan (2005).

EFQM goes beyond pure process modelling and aims to provide a framework to assess the entire organization – which is shown in figure 2.6. For each block in the figure there are a number of sub-categories, e.g. Category 3 *People* has 5 sub-categories, including 3a: “*People resources are planned, managed and improved*” and e.g. 3c “*People are involved and empowered*” – see also (Bou-Llusar et al., 2009) for details.

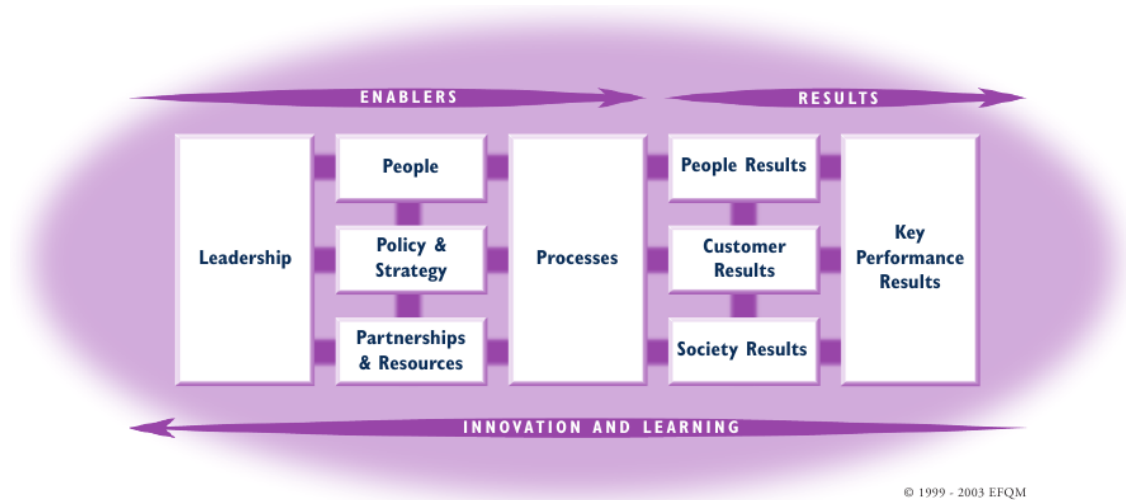


Figure 2.6.: The EFQM Excellence Model (Source: European Foundation for Quality Management (2003))

EFQM inherited⁵⁰ the central idea from statistical process control: continuous improvement (category 5b) based on systematic measurements of quality, cost, throughput times and adherence to schedules – all to the expectations of the customer (category 5c-e). This is effectively an internal daily-business short-term feedback loop for iterative

⁴⁹A recent example for a full EFQM implementation is the reinsurance Hannover Rückversicherungs AG (Heinrich and Kohlenberg, 2008).

⁵⁰Given their common heritage, the element continuous improvement based on measurements is shared amongst EFQM and the Toyota Production System TPS.

improvement.

Yet the scope of EFQM goes beyond these smaller short-term improvement iterations, extending the concept of monitoring and incrementally improving an organization's processes with *key performance indicators (KPIs)* for all aspects of the organization – incl. soft aspects such as leadership and human resources⁵¹. Improving the entire organization certainly is a medium to long term effort.

In contrast to other holistic KPI-based approaches such as the *balanced score card* (Kaplan and Norton, 2007), the EFQM framework is a predefined, systematic, holistic⁵² and complexity reducing perspective on all aspects of the organization. This predefined nature of EFQM has the advantage that its users need to engage with this – for them foreign – view of their situation, leading to new insights. Using the EFQM model does however not imply that it is the only and most suitable perspective on the organization. In the contrary, a number of EFQM practitioners advocate the combination of EFQM with other models, e.g. the Balanced Score Card (BSC), which allows for a better focusing on the specific and current issues of an organization and thus can be an ideal complement to EFQMs overview perspective (Lamotte and Carter, 2000; Schmidt, 2008).

For the assessment of the enablers and the result indicators the EFQM society has developed the *RADAR* method, which describes the assessment process as a multi-stage and mostly qualitative yet strictly structured process with teams of trained usually external assessors⁵³. In contrast to most other corporate benchmarking approaches, the RADAR process includes numerical scores (for complexity reduction). But instead of relying on pseudo objective and overly simplistic linear aggregation schemes using weighted sums of subscores, the RADAR process allows the assessors at every level⁵⁴ to use their sound and intelligent judgement to set the numerical score covering and simplifying the entire complexity within an enabler or results category. Subjectivity in this assessment is reduced by working in teams of assessor. In addition, the RADAR process not only evaluates the current performance of all parts of the organization but assesses whether the local performance is the result of a well defined process complete with integrated review and continual improvement mechanisms (integrated feedback cycles) – see also European Foundation for Quality Management (2001).

As in project management and TPS, the assessment of the situation is not only used for understanding the status quo but also to set goals and monitor improvement.

Particular to the EFQM model is furthermore the split of assessment categories into

⁵¹In the EFQM model the term 'people resources' is used instead human resources to emphasize the individuality of people.

⁵²EFQM encompasses all parts of an organizations.

⁵³The RADAR methodology emphasises many of the quality criteria also known in the social sciences, e.g. systematic analysis based on categories or inter-subjective qualitative judgement (Wengraf, 2001).

⁵⁴The only exception is the very top level for the *EFQM excellence award*, for which the scores of the different enabler and result dimensions are aggregated by a weighted sum. Yet for practical applications this sum is not required.

enablers and results, which are also called leading and lagging indicators respectively. The lagging result categories, cover direct business success categories, such as financial results, customer satisfaction or corporate image. The leading indicators track results that only with a time lag will lead to the direct business results, e.g. good leadership positively affects productivity and thus also positively affects the financial results (Schmidt, 2008).

Noteworthy is that the EFQM model does not include a strict description⁵⁵ of how the leading indicators, possibly in a combination, directly act on the lagging indicators. Only recently have researchers begun with attempts to empirically confirm the causal links between leading and lagging indicators (Bou-Llusar et al., 2009; Evans and Jack, 2003). Yet for the effectiveness of the EFQM model a detailed and mathematically strict description of the causal linkages is not necessary: Strict and detailed descriptions of the linkages would become very complex – especially when relevant yet industry specific factors are considered. Thus avoiding a strict linkage definition is effectively a complexity reduction⁵⁶. The fuzziness of the model leads practitioners to focus first on monitoring and improving all categories of the model in iterative steps, which – at least according to the common professional insight of the EFQM member organizations – will support and improve the organization’s effectiveness. Again for this improvement cycle modelling the causal linkages of the model strictly in mathematical form is not necessary.

The EFQM model furthermore includes very abstract guidance on how to support and improve the enabler and result categories with the *Fundamental Concepts of Excellence*. These include for example insights from W. Edwards Deming with “*Excellence is visionary and inspirational leadership, coupled with constancy of purpose [including communicating this vision within the organization].*” (European Foundation for Quality Management, 2003). Yet given the abstract nature of this guidance, they are applicable to many different kinds of organizations, but also require significant skill and effort in applying them in a concrete context.

Summarizing, the EFQM approach, like project management, focuses and spends significant effort on building a common and deep understanding of the state of the organization – shared by all relevant organizational actors – as a sound platform for an deep and intelligent discussion in order to find context specific solutions. EFQM covers all aspects

⁵⁵An example for a strict description would be a causal graph (Pearl, 2003) – as frequently used in structural equation modelling (SEM) (Pearl, 1998).

⁵⁶Describing and empirically proving the causal links implicit of the EFQM model would require first to operationalize all leading and lagging indicators as empirically useful numerical constructs, which are general and equally valid across multiple firms in different industries. In addition a range of environmental factors (e.g. the specific market environment) would need to be monitored and modelled during the time of the study. Next, given the sizable variable number, data over a large number of firms would need to be collected. Only then with suitable assumptions on causality could the causal linkages be demonstrated empirically. The two earlier cited EFQM / TQM validation studies by Bou-Llusar et al. (2009) and Evans and Jack (2003), which claim to validate most causal linkages, use questionable data and methods given the complexity of the problem and the diversity of the surveyed organizations.

of the organization (e.g. human resources) and balances short term with long term goals.

2.4.5. Link to the PIA-Model

As mentioned in section 2.4.1 on page 58, there are 3 principal commonalities amongst the three industrial practice models. The effect and effectiveness of these commonalities are discussed in the following with reference to the PIA-model.

1. **Visualize to create a shared understanding first.** All 3 approaches spend significant time and resources on creating a shared understanding of the current situation (status) and of the organization's aims.

Reaching the shared understanding is usually facilitated for example by a mental model⁵⁷, a visualization or a key performance indicator (KPI), which show the situation in a simplified (i.e. complexity reducing) and systematic manner. Hence applying the industrial practice models effectively creates a pre-filter to aid and reshape the team members filtering of the problem's complexity as illustrated by step 1 in the PIA-model with visualization (figure 2.3 on page 36).

The improved filtering also improves the understanding (step 2) and leads to better decision making (step 6). Thus this is how the industrial practice models facilitate improved local and contextual decision making by those directly involved – rather than using prescriptive instructions that may not fit the context.

The improved filtering is achieved by opening up a novel way to analyze the problem, which is achieved by offering a visualization tool, an analysis method or by qualification of the team members (i.e. by adding to the team member's prior knowledge: link 3 in the PIA-model).

Since the shared perspective is created by a team of employees and regularly reviewed, its subjectiveness is decreased and its robustness and usability is improved.

Furthermore the industrial practice models do not only require the shaping of *any* shared perspective, but give specific hints and prescriptions on how to systematically analyze and visualize the situation leading to a single and contextually relevant perspective within the limits of a prescribed framework. This perspective might be equally suitable as other perspectives but – by its systematic nature – exhibits the great advantage that it has a stable bias – i.e. the perspective does not depend on the current situation or the hopes, prior experiences and wishes of the participating organizational agents⁵⁸. In contrast, subjective perspectives of individuals or even groups of people are never completely without situational and subjective bias.

⁵⁷e.g. in the TPS: looking for all wasteful actions in a work process

⁵⁸Within the PIA-model (figure 2.1 on page 31) 'hopes, prior experiences and wishes' are modelled in a simplifying manner with the component 'prior / background knowledge'.

However despite its bias, subjective perspectives may more suitably represent the situation – especially since non-systematic perspectives frequently allow to cover more complexity. Therefore a value greater than the sum is derived from a fusion of an unbiased but possibly simplifying systematic perspective and a subjective but complex perspective. This fusion is achieved by discussion and subsequent learning of all participants indicated by step 7 of the PIA-model.

Hence the result of this perspective shaping process is an agreement of a detailed – yet complexity reducing – shared view of the situation with little bias and therefore high robustness (Argote (1999, Chpt. 4), Prusak (2005)). The artifacts created in the process of perspective shaping, e.g. a project plan or a KPI system, are only a support tool, used to create socially constructed and thus shared knowledge in the heads of the organizational participants, which drives the perception of the situation. See also section 2.3.4 on page 41.

2. **Engage in deep discussions regarding the solution.** All 3 industrial practice models contain little guidance on how to solve a concrete problem. At best vague and context-independent general solution hints are provided, which require a significant and intelligent effort to apply them in an actual context. Therefore the newly created deep⁵⁹ shared understanding of the situation allows the participating organizational agents to engage in a deep and constructive discussion in order to agree on goals, track improvements from an earlier status assessment and to develop or adapt solutions for the specific context of the firm.

In the PIA-model these discussions integrate the information from the shared and individual perspectives into new knowledge (step 2), which is the basis for decision making regarding the solution (step 6) and action (step 7) finally.

3. **Repeat and iteratively improve using feedback loops.** The three practices all include regular reviews to reassess the situation and the effectiveness of earlier decisions – allowing for readjustment or correction of the organization’s actions as well as the shared perspective. The latter may include also an iterative fine-tuning of the visualization method (step 8 in the PIAmodel fig. 2.3 on page 36). These reviews effectively created an outer feedback loop to the implementation of actions (steps 7 and 8 in the PIA-model).

Hence the planned or ongoing actions as well as the shared perspective are frequently refined in conjunction.

⁵⁹In the previous paragraph the shared perspective was labelled ‘detailed – yet complexity reducing’. A compromise between detail and complexity level is necessary to facilitate effective discussion of a situation. Also the links between the facts may be complex and therefore – given cognitive constraints of the human brain – a limitation of complexity in the facts and the inter-linkages is necessary for humans to cope with the problem (Boisot and Canals, 2004).

2.5. Alternative Perspectives Covered in Literature

At first sight, it may be surprising that all three industrial practice models focus much effort on creating a shared perspective⁶⁰ and leave the discussion and solution finding step to those who apply the general practices in actual local contexts – rather than giving more prescriptions regarding the solutions directly.

The PIA-model, derived from an entirely different tradition of scientific studies, can be used to explain the effectiveness of this approach: Following Weick et al. (2005), Orr (1996) and many others – as discussed in section 2.3.3 on page 34 – decisions strongly depend on how we perceive or make sense of a situation.

In addition, a stronger focus on more prescriptive solution hints (e.g. ‘best practices’) is not a good option either, since such prescriptions would need to be very complex to account for all contextual peculiarities that different organizations may have. Alternatively sharing of best practices without or little modification is limited to a small range of contexts (Matson and Prusak, 2003).

Hence, summarizing, the strength and generality of the three different industrial practice models, beyond simple ‘best practice’ sharing⁶¹, derives from indirectly supporting decision making – in a contextually relevant manner – by creating and agreeing on a robust shared perspective first. Since many organizational actors participate in the reasoning and decision making process, the first step for change management (Beer et al., 1990; Gioia and Chittipeddi, 1991; Orlikowski and Hofman, 1997; Tsoukas and Chia, 2002) is made: A shared understanding for the decided actions and possibly also compromises is widely spread within the organization. Thus these industry practices support organizational learning including individual learning and organizational change in a holistic manner.

2.5. Alternative Perspectives Covered in Literature

Earlier in this chapter (in section 2.3 on page 28) the PIA-model (figure 2.1 on page 31) model was presented and will be used for the interpretation of the results from the survey data in the later chapters. This section gives an overview of a number of issues and lines of argument that are subject to intense discussion in the literature – but which for different reasons have not been included in the model – either because covering these aspects adds complexity to the model, which is unnecessary for this application or because the argument is incompatible with the literature, which forms the basis of the PIA-model.

⁶⁰In his discussion of an organizational matrix structure, Bartlett and Ghoshal (1990) argues that the strongest effect from organizational structure derives from the fact that all organizational actors must assume a particular perspective. Thus he implies that if managers can flexibly assume different perspectives on the organization, the formal organizational structure becomes less important.

⁶¹Certainly ‘best practice’ sharing can also be done with sensible modifications to a local context. Yet such modifications usually require first an abstraction of the features of the best practice followed by an application of these features – after a structured analysis step – in a different and new way to the new context. I would argue that such an approach is more similar to the general approach of the three practices presented here rather than to ordinary best practice sharing.

Therefore the aim of this section is to compare the PIA-model to other similar or even opposing perspectives in order to sharpen the understanding of the model further and its relation to other widely discussed points of view. This also implies that this section is useful but not essential for understanding the other parts of the study.

The next subsections cover the following alternative perspectives on knowledge, learning and knowledge management:

- Knowledge as Object rather than a Personal Skill
- Identifying Application Relevant Knowledge (Knowledge Maps)
- True Knowledge vs. Diversity of Perspectives
- Valuation of Knowledge
- Importance of Knowledge Definitions

2.5.1. Knowledge as Object rather than a Personal Skill

In the field of management science, knowledge management (KM) is seen as the field of research that provides relevant insights regarding knowledge creation, retention (storage) and transfer within organizations (Argote et al., 2003). Other management scientists further include the use of knowledge within the scope of KM (De Long and Fahey, 2000).

Yet given the choice of the term ‘Knowledge’ as an important part of the KM research paradigm, KM analyses the problem through a very particular perspective leading to claims that in some aspects stand in contrast to the general insights described before in this chapter. In addition in large parts of the KM literature a definition of knowledge is either not made explicit at all or only a vague definition is provided (Schreyögg and Geiger, 2007) and there is only a limited agreement on the definition of knowledge – which is part of an ongoing discussion in the field of KM (Fahey and Prusak, 1998).

This section highlights the differences between some of the most important claims from KM and the general insights presented before (in section 2.3 on page 28), which are used in form of the PIA-model as basis for this study. Many of these differences have also caused discussions within the field of KM – as will be discussed later.

Most influential for KM is the *SECI* model presented in (Nonaka, 1991; Nonaka and Takeuchi, 1995). Central to Nonaka’s model is the distinction between two forms of knowledge: explicit and tacit knowledge. Referring to Polanyi⁶², Nonaka sees tacit knowledge as a genuinely personal type of knowledge that can not be verbalized, while explicit knowledge is verbalized knowledge that surfaces directly in discussions and can be captured in documents.

⁶²Tsoukas claims that Nonaka has misinterpreted Polanyi (Tsoukas, 2005b, Chpt. 6). Hence it can be argued that Nonaka misinterpreted or at least overly simplified Polanyi’s notion of tacit knowledge.

2.5. Alternative Perspectives Covered in Literature

Based on this dichotomous notion of knowledge, the *SECI* model illustrates how knowledge is converted between these two types of knowledge in a cycle with 4 modes of knowledge conversion (Nonaka et al., 2000, p. 9):

1. **socialization** (from tacit knowledge to tacit knowledge – e.g. by working together)
2. **externalization** (from tacit knowledge to explicit knowledge)
3. **combination** (from explicit knowledge to explicit knowledge)
4. **internalization** (from explicit knowledge to tacit knowledge)

Later Nonaka et al. (2000) added the concept of *Ba*, the Japanese term describing a shared context in which social knowledge transfers take place.

Particularly illustrative is Nonaka's example of a Matsushita engineer trying to learn from the Osaka Hotel Head Baker the perfect dough kneading technique in order to implement it in the next generation Matsushita bread baking machine. Nonaka describes this episode as *externalization* process converting the tacit knowledge of the baker to explicit knowledge that is then transferred to the engineer – facilitated by the context of working together.

Many authors in KM have used the SECI paradigm to describe knowledge transfer and conversion – exemplary papers are: Abou-Zeid (2002); Hussi (2004). Very frequently the application of the model goes along with the implicit assumption that knowledge has properties similar to that of money or a stock pile of a commodity:

- it can easily be identified,
- it can be counted (and its value measured),
- transferring involves mostly sending it, while receiving it does not involve much effort on behalf of the receiving party,
- it can be disconnected from people (externalized),
- it can therefore be managed directly

For examples see: North (2002) and Hofer-Alfeis (2003). As the arguments supporting the PIA-model showed and the following arguments will further detail, none of the above listed implicit assumptions describes knowledge and learning well.

The analogy with money yet has one deceiving advantage: the model is easy to understand for people with a traditional management training and existing management techniques can be applied with minor modifications⁶³. Consequently the aim of many

⁶³Following Weick et al. (2005) with his argument that plausibly rather than lengthy validation is the standard for adopting new knowledge, it is not surprising that this model was widely adopted in the management and computer science literature.

authors is to disconnect knowledge and capture it in databases. For example, Coffey and Hoffman (2003) try to capture the knowledge of NASA design experts in a database, given the long time scales of NASA projects. All these approaches emphasize knowledge-as-objects or knowledge stock to the detriment of knowledge flow – as Fahey and Prusak (1998) remark. Hopefully NASA can reap their efforts when the time comes and the database allows their young engineers to learn. Yet only if the young engineers put effort into learning from the database and if the lessons are truly useful for so far unforeseeable application, will their initiative payoff⁶⁴.

Yet it is unclear if Nonaka was fully understood, if one considers for example the following quote from Nonaka and Takeuchi (1995):

“Information is a flow of messages, while knowledge is created by that very flow of information, anchored in the beliefs and commitment of its beholder. Thus understanding emphasizes that knowledge is essentially related to human action.” (Nonaka and Takeuchi, 1995, p. 58)

Furthermore when Nonaka describes his SECI model as a cycle, one could also interpret the phases 1.) *externalization* (tacit knowledge of person A to explicit), 2.) *combination* (transfer of explicit knowledge) and 3. *internalization* (explicit to tacit knowledge in person B) as a teaching / learning relationship. However this is not how it is widely interpreted by many knowledge-as-object authors.

A special subgroup of these knowledge-as-object authors further claims that knowledge objects can be attached to process steps – for examples see: Binner (2008) and Kwan and Balasubramanian (2003). Their aim is to provide a managerial tool for verifying whether the person responsible for the process step has the required knowledge. Most of these arguments focus on explicit knowledge and while most of these authors acknowledge the existence of tacit knowledge, their focus on explicit knowledge leads them to ignore the tacit part (Fahey and Prusak, 1998).

Nonaka’s publication and widespread adoption of his model yet also called for further testing of the model and many critics:

In line with the findings presented in earlier sections, D’Eredita and Barreto (2006b) and Tsoukas (2005b) re-interpret the Matsushita bread baking example in a different way that contrasts with Nonaka’s model: In their view the joint dough kneading experience of the head baker and the engineer, made the engineer focus on the right bits of data (filtering) and thus allowed the engineer to learn the kneading skill even though the

⁶⁴Haas and Hansen (2005) demonstrated that in some cases, even searching takes too much effort in comparison to the returned value of the findings. In his study, different teams from a consulting company were analyzed how much they searched for information in the firms knowledge database, which was incentivized, and how successful the teams’ projects were. Quite unexpectedly those teams, which simply relied on their experience and little searching, found ways to use their time more effectively and produce better business results.

head baker could not describe his tacit kneading skill with words. Hence the non-verbal interaction of working together was pivotal in this case. Since the engineer was consciously reflecting on his learning process, he was able to acquire the skill in tacit as well as in explicit form. D'Eredita and Barreto (2006b) and Tsoukas (2005b) therefore claim that the baker's tacit knowledge was not converted and transferred at all. Instead the baker only helped the engineer to learn the bread kneading technique by himself. Since the baker was not able to verbalize his technique, his help was limited to the demonstration of the technique, which however turned out to be sufficient. Therefore the authors (with the support of Schreyögg and Geiger (2005)), claim that tacit knowledge can not be converted directly and without loss or change into explicit knowledge or vice-versa. The process of knowledge transfer should be modeled as a teaching / learning process instead.

Looking at this bread baking episode from very far away, one could still use Nonaka's model to describe the process as a conversion of the Baker's tacit knowledge into explicit knowledge, which is owned by the engineer. Like all models Nonaka's model is a simplified representation of the true mechanism and has limited predictive power (and usability). His perspective is particularly useful when analyzing knowledge transfer between large groups, e.g. organizational units or corporate divisions. For these high level types of analysis, Nonaka's model is abstract enough for large groups of actors but still highlights that forwarding of documents and instructions should be supported by face-to-face meetings and if possible personnel transfer (Argote, 1999, Chpt. 5).

If Nonaka's model despite all weaknesses is useful, the question arises, why an alternative perspective and model was used for this study: The aim of this study is to gain a more detailed understanding of how knowledge transfer happens between individual knowledge workers and how it can be supported by an organization. Since this research question requires an investigation at the individual level, I claim that the perspective illustrated by the PIA-model (fig. 2.1 on page 31) is more effective, since it highlights the following aspects of knowledge, which are particularly relevant at the individual level:

Knowledge is personal – i.e. in the heads of people (Okhuysen and Eisenhardt, 2002; Schreyögg and Geiger, 2005; Schön, 1992; Tsoukas, 2005b) and thus requires personal interaction to make use of it (Salter and Gann, 2003; Sadow and Allen, 2005). Other definitions are certainly possible but less useful, since with these other definitions clearly differentiating information from knowledge becomes difficult. More importantly, when knowledge is stored in text books in the book shelf, then the difficult step of getting that knowledge into somebody's head, becomes overly deemphasized, which diverts attention away from the learning step in knowledge transfer.

The **transfer of knowledge requires an active engagement** with the subject mostly **on behalf of the learner** – as argued for in section 2.3.1 on page 28. Despite this simple and old insight, the receiving part of the knowledge transfer, i.e. the more difficult part, is frequently neglected by the knowledge-as-object KM fraction – e.g. see

(Hofer-Alfeis, 2003; North, 2002). The required effort is even increased when the learning process also requires searching for information first (Haas and Hansen, 2005).

Thus knowledge in the heads of people cannot be managed directly like a simple other corporate resource (such as money or material). What can be managed is the organizational environment that supports individual and personal learning driven by the initiative of the individual employee – and this is what this study focused on (see section 2.7 on page 80).

2.5.2. Identifying Application Relevant Knowledge in Maps

Knowledge is hard to identify in detail as a set of small knowledge objects. Since knowledge is the integration of information, its essence is the connection of new information with existing knowledge – as discussed in section 2.3.3 on page 34. Knowledge therefore has little meaning when taken out of context (Tsoukas, 2005b).

Nevertheless it is possible to identify roughly larger bodies of knowledge – e.g. mechanical engineering knowledge about steam systems. One could even devise an exam to test whether a predefined scope of skills is covered. This method is frequently used when aiming to test for a minimum and standard skill set.

Yet knowledge management frequently aims to manage in particular those skills, which are non-standard capabilities of a corporation and give the corporation a competitive edge. Hence it is not surprising that a number of authors – e.g. Hofer-Alfeis (2003); Ward (1998) – aim at mapping explicit non-standard organizational knowledge.

Attempting however to determine the scope of knowledge accurately, rather than testing of a predefined standard scope, is more difficult – especially since many skills are mixture of explicit and tacit knowledge. Therefore even the experts owning the knowledge can only identify the explicit portions. The management of INCAT, an Australian fast ferry ship yard, does not fear losing their know-how to competitors, when allowing e.g. a Hong-Kong ship yard, to build their catamaran design on license. As a principle reason they cite *casual ambiguity*, the inability of even their expert engineers to fully explain which knowledge it is that allows them to perform. This is in-line with Tsoukas argument that we need to first become unaware of a skill to fully master it – as discussed on p. 54 (Tsoukas, 2005b).

Hence the attempts to map the particularly valuable non-standard explicit knowledge can be coarse at best⁶⁵. Furthermore, as discussed before in this section, knowledge in the heads of people can not be managed directly like any other commodity. Therefore

⁶⁵ Aside of mapping attempts in Hofer-Alfeis (2003); Ward (1998) cited earlier, which focus on mapping the explicit knowledge, mostly disconnected from people, there are also other efforts to map the knowledge in groups of people – e.g. see the social network approach by Cross et al. (2001). Social networking has a very different aim: facilitating personal communication incl. possibly teaching/learning situations. Due to their human centered approach, social networking techniques are much more promising.

it is not only difficult to create knowledge maps, their usefulness in application is also very questionable – in particular in comparison to the effort and cost to create coarse knowledge maps.

Yet mapping standardizable skills might be an effective measure to ensure a minimum standard qualification level throughout an organization or at various steps in a process and is practiced in many organizations since many years in the form of formal employee qualifications programs (with seminars, workshops etc.).

2.5.3. True Knowledge vs. Diversity of Perspectives

As already discussed in detail in section 2.3.5 on page 45, **diversity of perspectives** supports learning (Collins, 2001a; Orr, 1996; Walsham, 2001). Yet this diversity should have some limits – i.e. some **commonalities** in the perspectives and with it a shared language to describe the problem is important for effective problem solving in groups as well (Badke-Schaub et al., 2007; Carlile, 2004; Nonaka et al., 2000; O'Donnell et al., 2003).

So far this effect was discussed with reference to the PIA-model and how humans integrate filtered information to knowledge that they hold for the (current) truth (section 2.3.3 on page 34).

It holds however also an important insight with respect to the knowledge-as-object perspective: Organizations should not aim to enforce and disseminate a single 'true' stock of knowledge but rather should cultivate a (possibly limited) diversity of perspectives within the organization. Certainly these different perspectives will also have different levels of quality. However a diversity of perspectives offers a multitude of alternative true views highlighting different aspects of a problem. Thus in a practical application of the knowledge a critical comparison of different perspectives can lead to an overall improved and contextually relevant perspective on a problem.

Hence knowledge management efforts need to take into account that knowledge has different levels of quality (Schreyögg and Geiger, 2007). Thus like nature-approximating models in Physics, even partially flawed knowledge may have practical usefulness – albeit better quality knowledge could be more useful. Conversely not all what is treated as knowledge in organizations is flawless knowledge. Hence Schreyögg and Geiger (2007) have observed knowledge quality control processes in some firms.

Thus summarizing, a deep understanding of a problem frequently derives from a true understanding and a critical examination of multiple different perspectives on a problem. Thus aiming for a single true stock of knowledge should not be the aim of knowledge management efforts. The aim should be the cultivation of a diversity of high quality perspectives on an issue.

2.5.4. Valuation of Knowledge

In principle the valuation of knowledge is important, since in corporate environments time to ‘manage knowledge’ is limited. Thus knowledge management competes with other value creating processes – most frequently the core short term value creation process of the business (the daily business). Hence to be able to prioritize knowledge management activities with other business activities it would be necessary to estimate the value of the knowledge that is managed. Some authors from the knowledge-as-objects school of thought, therefore argue that knowledge management should aim at explicating and storing or sharing particularly valuable knowledge (Bornemann and Alwert, 2007; Bornemann, 1999; Hofer-Alfeis, 2003).

Yet aside from the difficulties with ‘managing knowledge’ (section 2.5.1 on page 69) there are some properties of knowledge that create significant challenges to assign a monetary value to knowledge:

The **value of knowledge stems from its potential for future applications** (Fahey and Prusak, 1998). Hence if one can predict future challenges and resulting knowledge applications for an organization and if one understands which knowledge is needed or useful for these future challenges, then one can estimate the value of knowledge. Meeting these requirements is at best difficult under practical conditions:

Most firms operate in increasingly turbulent markets and environments (Leibold et al., 2004; Spender, 1996). Hence predicting future challenges is frequently inaccurate. Since knowledge is hard to identify – as discussed in the last paragraph, it is difficult to link objects or fields of knowledge accurately with future challenges.

Thus knowledge hoarding approaches – aiming to fill knowledge databases – entail significant risks of over-investment in authoring and storing information (Lam and Chua, 2005). Aside from the waste of resources in such over-investments, too much irrelevant or out-dated information further decreases acceptance of knowledge databases, since searching becomes harder. Searching can even become so difficult and require more effort than value drawn from the findings (Haas and Hansen, 2005). In fact many larger corporations have put intense efforts in building up vast databases (North, 2002; Voelpel et al., 2005). Most of these efforts were not worth the effort – at least from the perspective of the users and therefore did not become a natural part of the organization without continuing attention and subsidies (North, 2002, p. 316).

Hence hoarding knowledge disentangled from its uses bears the risk of over-investment in documenting and storing knowledge as information (Fahey and Prusak, 1998). From this challenge, the following questions arise for managers: Who decides what knowledge is valuable? A manager or a larger group of practitioners who are using and will use the knowledge? Who decides how much effort to invest in documentation and what quality

documents need to have⁶⁶?

2.5.5. Importance of Knowledge Definitions

When designing organizational measures to support knowledge management, it is important to **share a proper definition of knowledge** among the members of the organization. In particular it is important to understand the subtle difference between information and knowledge in order to avoid becoming trapped in just launching a new information management effort under the new name of knowledge management. (Schreyögg and Geiger, 2007).

Understanding the nature of knowledge implies understanding the most important features of knowledge and challenges surrounding knowledge management – and this understanding is created in a discussion of a suitable knowledge definition with the organizational actors. Schreyögg and Geiger (2007) for example propose a ‘*discursive understanding of knowledge*’, which defines knowledge as an active and individual construction in critical discourse processes (as mentioned section 2.3.5 on page 45). Hence their definition highlights that there are different quality levels of knowledge, which e.g. lead to a knowledge review process at Shell (Schreyögg and Geiger, 2007, p. 94). Thus if a common understanding of knowledge is shared within an organization, the chances for success of measures supporting KM and learning are greatly increased (Fahey and Prusak, 1998; Hussi, 2004).

Since this integration step is a learning step, it requires the active involvement and also the necessary prior knowledge on the part of the learner. 2.3.1 on page 28. This personal concept of knowledge is in-line with the concept of *knowing* (Cook and Brown, 1999; Okhuysen and Eisenhardt, 2002). Similarly Schreyögg and Geiger (2007) ‘discursive understanding of knowledge’ as well focuses on the knowledge construction (in discourse) as a process or activity. Thus action emphasizing definitions of knowledge such as ‘knowing’ or ‘learning’ are most suitable to describe the challenges surrounding knowledge.

2.5.6. Summary – Alternative Perspectives

In the preceding subsections the argument was made that the knowledge-as-object view, which is popular in the knowledge management research stream, is not very useful as a basis to design knowledge management systems. The comparison of knowledge to other organizational resources such as capital or assets, hides too many aspects of knowledge Fahey and Prusak (1998):

⁶⁶Documentation can be created with different levels of effort and different resulting quality. The lowest level may be storing an individual experts notes without further editing – which then are only useful to remind the same expert at a later time about his insights. On the high quality end, the documentation may have undergone several steps of editing, making it understandable for a wide audience without contextual knowledge and checking for legal aspects e.g. in operating manuals.

- Knowing and learning requires the active engagement of people – especially when learning new knowledge (section 2.3.1 on page 28).
- Knowledge is hard to identify (section 2.5.2 on page 73).
- Knowledge cannot be disconnected from people and their context – without loss.
- The value of knowledge can hardly be estimated in sufficiently fine granularity to support the daily work.
- Knowledge cannot be directly managed.

Thus defining that knowledge can only be found in the heads of people and everything else is data and information – most suitably captures the properties of knowledge. For supporting knowledge flows in organizations, the focus should be on the activities surrounding knowledge (e.g. knowing or learning). Further details on the knowledge definition used for this study in the following section.

2.6. Assumed Perspective & Definitions

In this section the definitions chosen for this research are presented:

2.6.1. Knowledge Definition

As argued for in the preceeding sections, learning is a more suitable paradigm for the purpose of this study than the notion of knowledge. Just for reference, knowledge, information and data are defined for the purpose of this study as follows⁶⁷:

- **Data** is any sensory input that we are receiving e.g. through our eyes or any input that we could receive, e.g. if we looked in the right direction. In the end data is sensory input but it can be brought to our senses by means of information technology.
- **Information** is a filtered extract of all available data, either by deliberate choice or very strongly also by unconscious filtering processes driven by prior knowledge – see section 2.3.2 on page 30 and following. Information can be in the heads of people but can also be conserved in written form, e.g. in a university text book or in newspaper commentary.
- **Knowledge** is the integration of information and prior knowledge into understanding – see section 2.3.3 on page 34. The process of integrating knowledge is called learning.

⁶⁷The definitions used here are in principle following Ackoff (1989) and Tsoukas (2005b).

2.6. Assumed Perspective & Definitions

Knowledge can only be in the heads of people. A text book in the book shelf is not knowledge, it is only a carefully arranged stream of information, which is intended to aid the learner.

Certainly data, information and knowledge can also be defined in other ways. e.g. some scholars might argue that a newspaper commentary contains knowledge, while a newspaper report (aiming to only portrait a situation in words) contains only information.

The definition presented for this study has however one important advantage – making it useful in organizational application⁶⁸: Information and knowledge are clearly distinct categories, since one may be stored electronically and on paper, while the other is only in the heads of people and involves a challenging process called learning. Thus when an organizational measure is aimed at information management, it is clear that the goal can be full automation of a data filtering and information transmission process. In contrast to that, the people bound definition of knowledge already implies that a measure aimed at the organizational knowledge, can not lead to full automation but requires the involvement and support of people. Hence it becomes clear to all organizational participants, that measures aimed at information management are in principle different from knowledge management efforts.

2.6.2. Organizational Learning Definition

To emphasize the active character of knowledge transfer (section 2.3.1 on page 28), a learning perspective was assumed for this study. Another reason for learning (rather than e.g. teaching or knowing), is that learning is frequently the most limiting factor in the knowledge transfer process and the application of knowledge (as was discussed in section 2.3.8 on page 56).

There fore and as discussed in the motivation section 1.2 on page 17, gaining insights on supporting organizational learning with organizational measures is the aim of this thesis.

Unfortunately “[...] definitions of organizational learning show as little convergence as definitions of leadership.” (Berson et al., 2006, p. 579). Hence the term ‘organizational learning’ is frequently used rather liberally and often even without explicit definition: Argyris (2002b) writes about organizational learning but really focuses on managerial learning, without clarifying a distinction between organizational and managerial learning and therefore implies that managerial learning is strongly linked to organizational learning. Argote (1999) studies the organizational learning effect of American ship yards during the second world war building the ‘Liberty Ships’ (Lane, 2001) by using the productivity increases as a proxy for organizational learning.

⁶⁸See Fahey and Prusak (1998) who argue for the importance of a practical working definition of knowledge in application.

More explicit definitions can be found in the following publications: Following [Brown and Duguid \(1991\)](#) organizational learning integrates practice, individual learning as well as change – and leads to innovation. [Crossan et al. \(1999\)](#) model organizational learning in their *4I framework* by “*social and psychological processes: intuiting, interpreting, integrating, and institutionalizing*” on p. 523. Hence they link multiple levels: the individual, the group and the organizational level. Hence organizational learning begins with individual learning, yet also conversely individual learning is influenced by group and organizational learning ([Vera and Crossan, 2004](#), p. 225).

Since individual learning is a key element in organizational learning, this study focuses on the individual learning part of organizational learning.

For simplicity but yet not in contrast with the afore mentioned authors *organizational learning* is defined for the purpose of this study as consisting of two major components:

- Individual learning
- Organizational Change – including group learning and change management⁶⁹ (i.e. institutionalizing)

2.6.3. Individual Learning Definition

Individual learning was already discussed in detail in the previous sections (especially in section [2.3.6 on page 47](#)), where it was described with the support of various publications. Most noteworthy is:

- **On-the-job learning is particular** (i.e. special) **to the problems** at the workplace of an individual employee. Thus this study is not concerned with formal learning of standardized subject matter – as is common in school, academic or seminar settings – as studied by [Roknagel \(2008\)](#) and [Maurer et al. \(2003\)](#).
- **Learning** as defined for this study is **directly linked to perception and to decision making** (incl. problem solving activity) – i.e. the use of knowledge. Thus it is similar to the definitions used by [Berings et al. \(2005\)](#); [Brehmer and Dörner \(1993\)](#); [Dörner et al. \(1999\)](#); [Sengupta et al. \(2008\)](#), which focus on problem solving.
- Learning affects not only the understanding of a particular phenomenon but also includes the **refinement of our perspective** on this or other phenomena (see the PIA-model in figure [2.1 on page 31](#)).

Compare also the operationalization of this learning definition with the *learning index* in section [5.4 on page 145](#).

⁶⁹For references on change management see ([Beer et al., 1990](#); [Orlikowski and Hofman, 1997](#); [Tsoukas, 2005a](#)).

2.7. Research Gap and Research Question

2.7.1. Research Gap

As outlined in the previous sections many streams of research already offer findings relevant to knowledge intensive work. In particular the mechanics of cognition and learning have been described as well as a number of peculiar properties of knowledge such as tacit knowledge.

A large variety of factors affecting learning, knowledge creation and transfer has been presented. Frequently studies cover a particular cause-effect relation (or a limited set of factors): e.g. effect of factor A on learning, while holding factors B,C and D constant. Yet is unclear, which factors most strongly affect learning, when considering a larger more encompassing set of factors.

When spontaneous (non-formal) learning processes are considered in detail, organizational factors are not considered (Berings et al., 2005). The effect of organizational factors on learning are only studied as factors causing the participation in formal on-the-job training activities (e.g. Maurer et al. (2003)).

Furthermore most research focuses either on the individual use of knowledge outside of organizational contexts or on the non-individual use of knowledge in organizations. Hence organizational factors are either not covered in depth or, if they are, then the learning processes are not investigated in detail on an individual level.

2.7.2. Research Question

Given that individual learning is key component of organizational learning, yet little is known on how to support it by support it with organizational means, the following research question was chosen:

Which are the most important organizational features that support or hinder on-the-job learning?

With ‘on-the-job (individual) learning’ as defined before in section 2.6 on page 77 as spontaneous learning, while solving problems at work.

Since organizations are under constant pressure to adapt to the external environment (Malik, 2008), it is essential that the organizational members can focus their efforts on a few important factors, rather than aiming to optimize all learning supportive factors of the organization.

This main question leads to a number of sub-questions:

- How can we measure ‘organizational learning’?

- What are the most important organizational barriers and promoters (factors) of on-the-job learning?
- How can insights about individual learning in organizational contexts, be used to design organizational measures to support learning?

The next chapters will outline the research methodology (chapter [3 on page 83](#)), the actually used research methods – i.e. the survey (chapter [5 on page 137](#)), the development of the statistical analysis algorithm BOGER (chapter [6 on page 171](#)) and result interpretation (chapter [7 on page 205](#)) of this study followed by the implications (chapter [8 on page 263](#)).

3. Theory of Science & Methodology

Chapter Contents

3.1. Scientific Paradigma	84
3.1.1. Methodology as Quality Standard for Methods	84
3.1.2. Science – a High-Quality Form of Investigation	84
3.1.3. Basic Assumptions about Reality and the Value of Research . . .	86
3.1.4. General Value Proposition of this Research	91
3.1.5. Perspective Validation and Empirical Results	92
3.1.6. Iteration Improves Research Quality	93
3.1.7. Methodological Approach	95
3.1.8. Optimizing Methods – Cost vs. Benefit and Quality	96
3.2. Choice of Methods for this Study	97

This chapter covers the following topics of scientific theory, which form the theoretical basis for the discussions regarding the applied methods in the remaining chapters:

- 1.) **Basic assumptions** about science (and in particular about **quality criteria** for scientific results) underlying this study are presented. Section 3.1 on the following page
- 2.) The selection and **sequence of the mixed methods** used for this study are summarized. Section 3.2 on page 97

This chapter is particularly relevant to those readers who wish to inspect the methods of this study and thus need to understand the basis on which the combination of methods used in this study has been chosen. Further details on the individual methods are described in the chapters on the qualitative stage (5.2 on page 139), the quantitative survey (5 on page 137) and the statistical analysis (4 on page 101 and 6 on page 171).

3.1. Scientific Paradigma

3.1.1. Methodology as Quality Standard for Methods

To judge the methods used in later sections, one first needs to describe quality standards for methods. Arbner and Bjerke (1997) refer to such a quality standard as *methodology*. A methodology would define criteria such as reliability, validity, context relevance, generalizability, value to practitioners (e.g., predictive power) and research cost. When considered individually, most researchers would agree to aim for these criteria – especially when an ideal research design can be found that perfectly meets all quality criteria (including cost and duration). Differences quickly surface, however, when the researcher needs to compromise (e.g., due to a limitation in resources, time and access to organizations).

Different basic and *personal* assumptions will lead to different optimal trade-offs between value to practitioners, context relevance, robustness of the results, accuracy and research cost (Arbner and Bjerke, 1997, Chapter 1).

Thus this section, firstly, makes explicit the basic assumptions underlying this study and, secondly, describes the quality criteria (the methodology to select appropriate methods) for this study. In section 3.2 on page 97, this methodology is applied and the actual choice and sequence of methods for this study is summarized.

3.1.2. Science – a High-Quality Form of Investigation

Useful and valuable knowledge about organizations can be created by researchers but also by organizational actors, consultants and other investigators. Yet scientific investigation sets itself apart from other forms of investigation by the following widely accepted basic quality standards for the investigation process (Arbner and Bjerke, 1997, p. 23) :

- The investigations should be **systematic** – i.e., follow more or less strictly defined rules. Such explicitly described and followed **methods reduce subjective bias** and make the process of obtaining the results verifiable – i.e., testable by others.
- The line of arguments should also be systematic. All **statements** must be either **supported** by citations, deductive reasoning or empirical results. Conclusions should be confined to those claims that can be supported with new evidence or literature. In particular, claims regarding **causality** should be **carefully supported** by more than just statistics showing associative relations¹.
- The methods, the sourcing from other literature, the interpretation of the results and the reasoning behind the conclusions should be exposed to peer review by

¹Statistical evidence that two variables are correlated, i.e., have high and low values at the same time (associative relation), includes no support for causality. Only in combination with further evidence (e.g. from literature) on the direction of causality, can association be interpreted as partial evidence for a causal claim.

publications within reasonable time intervals. By this “**Principle of Publicity**” (Arbnor and Bjerke, 1997, p. 24), the research may serve as an inspiration for others and allow others to test the results.

- A representative sample of the **existing literature** relevant to the topic needs to be taken into account – in order to find a suitable perspective on the problem and to avoid overlooking relevant aspects.
- If possible, **researchers** should work **in teams** or at least discuss one another’s research approaches and methods – in order to reduce subjective bias by inter-subjectivity.
- Since the **aim** of research is to better understand the subject in question, the result will be an **abstract description** of the subject – i.e., a **model**, or what Arbnor and Bjerke (1997) refer to as *ideal-typified language* (as described earlier on p. 89). Such abstract descriptions are often created to allow some form of prediction in practical application. In that case, the **predictive power** of the model or abstract description becomes a principal quality criterion².
- In addition, the **methods** chosen should **reliably** yield **robust** and sufficiently **accurate** results. The quality requirements for the result of the analysis also imply quality requirements for the data sample: The **sample** should be **sufficiently large** for robustness and accuracy. In addition, the **sample** of cases or people observed should be **representative** of the sample the researcher aims to make claims about. (More details on this are discussed in chapter 4 on page 101).
- The **limitations** arising from any weaknesses from the employed methods or data sample should be made **explicit**. This implies that the **strength of any new claims** arising from the research should be discussed.
- There are many empirical methods to reliably detect associative relationships – i.e., two states or events always occurring in conjunction or in sequence. However, in most cases **association** alone is not sufficient: Building models of a phenomenon based on association, requires in addition insights about the causality between factors, including the **direction of causality**. If a research project aims to detect and support causality with appropriate evidence, an analysis should include either

²Starbuck emphasizes that when a deeper understanding of a social phenomenon is reached, the principle quality test of this new theory is successful prediction of future events: “*When researchers attempt to improve social systems, they must acknowledge the values guiding their proposals, use their theories to predict outcomes, and revise their theories when the predicted outcomes do not occur.*”, (Starbuck, 2004, p. 1249)

3.1. Scientific Paradigma

a direct³ or indirect⁴ method to get from association to causation. (More details on this are discussed in section 4.1.5 on page 109.)

These quality requirements are independent of the basic assumptions and aims of the research, and they are widely accepted.

Aside from the use of explicitly defined and robust research methods, the requirement of publicity enables the scientific community to further improve insight into a topic by: 1.) testing research results by peer review (and thus identifying robust insights – which will survive) as well as 2.) iteratively refining and further developing insights by sharing new results with other scholars. Thus scientific results shows a superior quality and robustness not only in the rigorous application of robust methods but also when a claim stands the test of time – i.e., when the results are confirmed by other researchers with new perspectives, new data and other methods.

3.1.3. Basic Assumptions about Reality and the Value of Research

Knowledge about Organizations is Socially Constructed In the theory chapter 2 on page 21, the PIA-model (figure 2.1 on page 31) was presented as a model describing the process of learning in episodes. The principal insights are not limited to learning in work settings but apply equally to learning of researchers during the course of their investigations.

Like organizational actors, researchers must make sense of the situation in the organization under investigation. Given that organizations are complex social systems, researchers need to reduce this complexity (Starbuck, 2004, p. 1238) by describing their observations in more abstract categories, which are the necessary basis for building models and creating abstract insights that allow application and transfer to future cases. In section 2.3.2 on page 30, this process of complexity reduction was described as the highly subjective process of *perspective taking*, which can be supported (but not replaced) by systematic analyses or visualizations (section 2.3.3 on page 35). As illustrated by the PIA-model with visualization (figure 2.3 on page 36), the analysis methods can also be iteratively refined in the process, which may improve but also subjectively bias the results.

³A direct way to support a causal statement with empirical results is by active experimentation (Hitchcock, 2007). The complex system or process under study is actively manipulated, and the effect of the intervention is observed. Interactions of variables can be detected by design of experiments (DOI) – see also Box (1994).

⁴In a number of research fields, interventions or manipulations of people, economies and organizations are not a feasible or ethical option. In these cases, researchers frequently combine statistical analysis of association with other methods to find the direction of causality – e.g., using qualitative methods (Starbuck, 2004). Other researchers make the strong assumption (Hitchcock, 2007) that sequence indicates causation, such as in longitudinal studies (e.g. Maurer et al. (2003)) or data mining in biology (e.g. Opgen-Rhein and Strimmer (2007)). In modern statistics, a number of methods are under development that allow one to draw conclusions regarding causality with weaker assumptions, which thus require less support by other methods (Pearl, 1998). Yet even these modern methods do require assumptions - albeit weak assumptions.

Scientific knowledge about organizations is also constructed by the subjective creation of categories, which are used as a specialized and subjective language to describe organizations (Starbuck, 2004). Given that researchers mostly work embedded in research communities, these categories and perspectives are affected by national culture (Dalmedico, 2004) as well as the culture of the research community. Therefore a number of management scientists view knowledge from organizational research as *socially constructed* (Arbner and Bjerke, 1997; Goldberg and Cole, 2002; Malik, 2008; Starbuck, 2004; Tsoukas, 2005b). Social construction was already illustrated with the PIA-model in section 2.3.4 on page 42. In the following, researchers who see knowledge as socially constructed are referred to as constructionists.

The challenge of subjective perspective taking in principle applies to all sciences. Yet given the nature of their problems, some sciences are less strongly affected by it.

For example, the natural sciences are frequently concerned with problems in which the complexity lies not so much in the operationalization and selection of relevant variables but more in understanding the interaction of a few variables. The simple mathematical model⁵ $F = m \cdot a$ from Newton's laws of motion is a good example: One of Newton's major contributions was that an object's mass and the forces acting on the object are the only two variables that can predict the object's acceleration with a very high degree of accuracy⁶. All other variables, such as the size and shape of the object, are not relevant unless they lead to forces acting on the object (e.g., the force of air resistance depends on the object's shape, cross-sectional area and current velocity). In contrast to organizational science, the number of candidate variables for a model is finite (and small), and all variables are well defined and quantifiable. In addition, Newton was able to perform experiments in which he controlled (as in manipulated) all variables and the environment. With the experiments and systematic empirical methods, he was able to show that only the acting forces and mass determine acceleration in the relationship described by the mathematical model. Newton's model stood the test of time and still serves as a good approximation⁷ used in many engineering sciences.

In organizational science, subjective perspective taking is more challenging because:

- Manipulative experimentation as a way to validate hypotheses (especially about causality) is frequently not possible or ethically acceptable, and thus researchers have to resort to 'natural experiments' (Starbuck, 2004). Further details follow in

⁵Force is equal to mass times acceleration.

⁶Newton's model is precise enough for most common engineering purposes, yet some applications, such as GPS, require taking into account relativistic effects – as described by Albert Einstein's theory of general relativity.

⁷Albert Einstein refined Newton's model with the special theory of relativity. Yet the refinements only become significant when dealing with bodies traveling at speeds that are a substantial fraction of the speed of light. Hence it applies to only a few engineering applications. An exception are GPS receivers, which have to compensate for relativistic effects connected with earth's rotational speed.

3.1. Scientific Paradigma

section [3.1.3 on the next page](#). But natural experiments or other analyses usually rely on strong assumptions and cannot be easily replicated under the same conditions by other researchers.

- Variables are hard to identify and hard to operationalize. For example, a person's background knowledge is hard to operationalize, yet it is very relevant (section [2.3.4 on page 41](#)).
- There are many more potentially relevant variables (Starbuck, 2004). The more potentially relevant variables exist to describe a scientific problem, the more subjective complexity reduction is required, and the more subjective the insights from the investigation will be.

These challenges commonly require organizational researchers to use subjective assumptions and subjective judgement in order to get to their results, which is why organizational science studies rarely offer proof without doubt. Given that organizational science, like other sciences, uses the principle of peer review, theories are inspected and tested by a community of researchers. Thus results that prevail over extended periods of time at least have an inter-subjective support. Still, the community of organizational researchers is embedded in various cultures (national culture, scientific culture etc.), and the views of the researchers testing a new theory will not be completely independent, and therefore organizational science results are at least in part subject to social construction.

In addition, and possibly more severe, the organizational mechanisms themselves may be socially constructed – e.g., they may be influenced by corporate or national culture (Malik, 2008; O'Donnell et al., 2003; Tsoukas, 2005b; Voelpel and Meyer, 2006). As described in section [2.3 on page 28](#), this effect can also be illustrated with the PIA-model (fig. [2.1 on page 31](#)). The organizational actor's complexity-reducing perspective on a problem or a situation is largely determined by his or her socially constructed background knowledge, and thus the actor's behavior is also affected by this socially constructed perspective (section [2.3.4 on page 41](#)).

In summary, organizational science results are at least in part affected by social construction, due to severe challenges with applying scientific methods to complex organizational problems and since the observed organizational mechanisms themselves may be socially constructed.

The aim of the method mix used for this study is to reduce subjectivity and the effect of social construction – yet from the preceding arguments, it must be clear that the results of this study can never be completely unaffected by social construction and subjectivity.

Simple Prescriptive Strategies do not match Organizational Complexity The socially constructed and thus subjective nature of knowledge creation about organizations is only

one challenge stemming from the complexity of organizations. Complex systems require equally complex models to fully describe their behavior (Malik, 2008). Even if modeling is desired to reduce complexity, too much complexity reduction may lead to overly simplistic recommendations that work in some situations and fail in others – as the following example illustrates.

Simple strategy prescriptions (e.g., in the form of *best practices*⁸) are popular among many management scientists and practitioners (Matson and Prusak, 2003). Prominent examples are the diversification wave in the '80s, followed by a counter-movement back to 'concentration on the corporate core competencies' related to the out-sourcing hype in the '90s. Not surprisingly, neither out-sourcing nor in-sourcing is a superior strategy in general. They may, however, be strategies leading to superior results in certain conditions. The difficult part is describing the conditions in which a particular strategy is effective. Given an unlimited number of organizational structures existing in a large number of different commercial environments, such a model accounting for a high complexity in boundary conditions is bound to become very complex itself. In addition, this complexity – including a large number of variables – makes scientific validation in most cases infeasible.

These challenges with complexity explains why simple models and simple strategies are much more commonplace than complex and contextual models. Yet those simple models have important drawbacks:

“[...] formal strategy models cannot offer contextually sensitive and time-sensitive advice, nor can they formally suggest novel ways of acting.”, Tsoukas (2005b, p.369)

Starbuck further illustrates another drawback of complex models, such as computer simulations with many variables and delayed effects: the behavior of these models is not simple and thus hard to understand. Due to the lack of transparency, policy makers frequently reject complex models in favor of simpler models – even though the complex models describe reality much more closely than simpler ones (Starbuck, 2004, p. 1238).

How Organizational Research can Contribute Value by Perspective Setting If simple models do not directly apply in many contexts, and complex models are commonly infeasible to create and validate, the question then arises: What can organizational research contribute at all to practitioners and other researchers?

Management scientist William H. Starbuck⁹ argues for a more active role of the researcher:

⁸Best practices can still have value – if they are adapted to local circumstances rather than blindly applied (Matson and Prusak, 2003). Yet going successfully beyond blind application of best practices requires substantial understanding of how a particular best practice is effective.

⁹Starbuck formerly served as the editor of the management science A-Journal Administrative Science Quarterly and still serves on various boards of journals related to organizational science and organizational behavior.

3.1. Scientific Paradigma

“Rather than realities, the social systems I was studying proved to be arbitrary categories created by observers or social conventions. I became an advocate for research that actively attempts to change situations rather than merely to observe what happens spontaneously.”, Starbuck (2004, p. 1233)

In particularly, he recommends searching for and analyzing natural experiments, and he advocates for the researcher engaging with practitioners in organizational change efforts to test hypotheses:

“Because systems are almost always close to their equilibria, they do not have to display the capabilities that they would have when displaced from their equilibria.”

“Thus, I began urging myself and my colleagues to search for natural experiments and to become engaged in efforts to improve social systems. Natural experiments occur when exogenous events displace social systems from their normal equilibria. In these situations, one can see some of the systems’ adaptive and reactive capabilities, which opens the possibility of discovering why equilibria exist.”, Starbuck (2004, p. 1249)

Similarly, Arbner and Bjerke (1997) argue for an active role of the researcher. However, in their *Actors approach* to research, they focus on creating and refining an understanding of the organization together with and for the practitioners. They argue that causal models, modeling an ever-changing status quo, and simplistic direct attempts to change the organization, e.g., by ‘motivating employees’, is much less effective than creating a more refined shared understanding and a meaning for the change among all relevant organizational actors (including managers and employees) (Arbner and Bjerke, 1997, p. 279). Hence the focus of Arbner and Bjerke (1997) is on **perspective setting** – similar to the effect of the industrial practices described in section 2.4 on page 57.

In more abstract terms, Arbner and Bjerke (1997, p. 60) describe the results and value added by scientific research following the actors approach as consisting of:

1. *‘descriptive languages’* – a systematic description and summary of the empirical observations,
2. *‘ideal-typified languages’* – a description of the gained/refined understanding of the phenomenon and
3. *‘emancipatory interactive action’* – the presentation of the insights to practitioners or other researchers stimulating them to reflect on their own perspectives and understanding for further improvement.

This definition fuses the analysis of empirical observations leading the researcher or researchers to a refined understanding of the phenomenon that can be shared with other

researchers and organizational actors. Sharing this novel perspective on the organization allows all involved to mutually refine one another's perspective on the phenomenon. The practical value of the research derives primarily from the refined perspective of the organizational actors, who will be able to make better decisions based on more effectively filtered information.

Perspective setting is also the primary objective of management science for Malik (2008). He describes his research as merely a description of his personal and subjective understanding about mechanisms within organizations, which he has arrived at during the course of his experiences in various organizations. He argues that the value in his research lies in the particularly insightful perspective he offers with his *viable systems model*¹⁰.

In summary, despite the complexity of organizations and the contextual differences between them, organizational science can contribute by helping the organizational actors to reflect on and refine their perspective, which creates a 'springboard' for more effective decisions and action (see Weick et al. (2005) in section 2.3.3 on page 34).

3.1.4. General Value Proposition of this Research

As already mentioned in section 3.1.1 on page 84, research designs are optimized to meet multiple objectives (validity, reliability, robustness, cost, duration etc.). In most circumstances, designing a research approach also involves making sensible compromises that maximize the value of the research (given the budget and a limited time frame). Therefore the intended value of the research in this study needs to be defined first – which is the purpose of this section.

The aim of this study is to contribute to science and support practitioners in the following manner:

1. Search for relevant literature to get an exposure to multiple and systematic perspectives on the phenomenon.
2. Systematically observe and describe the phenomenon (in this case, learning in organizational contexts).
3. Fuse the summarized results from the observations with insights from literature to gain a deeper understanding of the phenomenon for typical cases.
4. Present this understanding and the methods to other researchers for inspection and in order to enable others to test and extend the insights.

¹⁰The *viable systems model* models the control mechanisms in organizations by analogy to the control mechanisms of complex biological organisms – following the work of Stafford Beer. Rather than giving prescriptive advice about strategy, Malik provides practitioners with a particular perspective for understanding their organization (in Malik's way), allowing them to draw their own conclusions, leading to contextually adapted and relevant decisions.

5. Present this understanding to practitioners¹¹ in a suitable form that is understandable by the audience – thereby offering a new perspective on the phenomenon and promoting a deeper understanding. Practitioners may then combine their own perspectives with the new perspective and reflect on their actions and behavior regarding the phenomenon in order to improve decision making (for this study, the practitioners are managers who shape the working environment to support learning).

This study is applied research in the sense that it aims to bring benefits to practitioners. Basic research or methodological research similarly aim for the above value goals by emphasizing item 3.) rather than 4.).

In summary, this study aims to systematically observe the phenomenon of interest (here, on-the-job learning) and integrate existing insights on the topic. This study contributes by offering a novel, high-quality and robust perspective to practitioners and researchers.

3.1.5. Perspective Validation and Empirical Results

Earlier in this chapter, on page 89 and in section 3.1.4 on the previous page, perspective setting among organizational actors was proposed as the primary method of transferring scientific insights to practitioners. In the theory chapter (section 2.3.5 on page 43), exposure to the opinions of others by reading or discussions was cited as the most effective way to develop and refine one's own perspective on a problem. This raises the question of whether and why empirical investigations are necessary and valuable at all for creating, refining and validating perspectives.

As described in section 2.3.6 on page 50, Orr's copy machine technicians test their diagnosis by exchanging mechanical parts. Without this validation step, the technicians' storytelling could easily drive them into a perfectly plausible but completely false direction.

Similarly, Starbuck (2004) cites an example on medical doctors and recommends an analogous approach to organizational research: The medical doctor, as cited by Starbuck, claims that good medical doctors do not rely on diagnosis alone, since there are many more combinations of symptoms than diagnoses, and symptoms may be more or less pronounced. Therefore instead they validate their initial hypothesized diagnosis with careful treatment (e.g. low doses of medication) – in order to proceed with treatment in the same direction, only if the patient responds in the expected manner.

In both examples, empirical information can be very helpful in narrowing down the set of all possible and internally consistent perspectives on a problem. Compared to discussions with other researchers, for example, empirical methods have one distinctive feature that makes them effective: they are systematic. As already discussed in section 2.3.3 on

¹¹The target audience for the results of this study are practitioners from the focus organization as well as from other organizations and other researchers.

[page 35](#), when a systematic and less subjective but unintelligent presentation (visualization) of a problem is combined with a person's subjective and intelligent perspective, it can be a strong tool to understand and solve difficult problems, which can also be useful in scientific research.

Empirical methods are socially constructed tools that highlight only certain aspects of a problem while hiding others. Yet, despite some built-in biases (by design) and possibly even systematic flaws of the method, the systematism inherent in the method largely immunizes it from biases caused by the researcher's hopes or expectations.

Thus, similar to the analysis methods of the industrial practices described in [section 2.4 on page 57](#), the value of empirical results lies in the different nature of biases – i.e., a systematic bias rather than a subjective one. The comparison of the empirical results to the researcher's current perspective, including current expectations, hypotheses and hopes, allows the researcher to reflect. This may lead the researcher to adjust his or her current perspective or even decide to proceed with further empirical investigations with a new focus.

To put it simply: an answer certainly depends on the question asked¹², but a systematically acquired result may still be surprising and may lead open minds to change their perspective and/or follow up with further questions. A constructionist researcher uses empirical results not as source of ultimate truth but to ground his or her process of perspective refinement on systematically acquired (thus differently biased) data – and, if possible, refines the perspective over several iterations, as discussed in the next section.

3.1.6. Iteration Improves Research Quality

As described in the last sub-section, empirical methods are powerful tools to validate the researcher's perspective. Yet empirical methods will be systematically biased by the design of the method – e.g., the design of a questionnaire. The design of the empirical method is driven by the researcher's perspective.

In [section 2.3.6 on page 47](#), based on various studies, learning processes were described as iterative processes, iterating between perspective taking, integrating the filtered information to new insights, leading to new decisions to look for further information (step 7) and possibly to adjust the methods (step 8) (see the external feedback method in the PIA-model [figure 2.3 on page 36](#)).

As for other iterative learning processes, iterating between refining the researcher's perspective through interaction with other opinions and conducting empirical investigations is an effective approach to increasing the quality of the research results. Given that reducing various biases is one of the most formidable challenges in research, iteration between perspective setting and multiple different empirical methods primarily improves

¹²... and the questions asked depend on the researcher's current and subjective perspective.

3.1. Scientific Paradigma

the robustness of the research results.

In addition, iterative research approaches are also recommended specifically for scientific learning about organizations in literature.

As mentioned earlier, Starbuck advocates focusing on natural experiments in organizational research. In order to reduce biases due to subjective values underlying the researcher's perspective, Starbuck further recommends an **iterative research approach**:

“When researchers attempt to improve social systems, they must acknowledge the values guiding their proposals, use their theories to predict outcomes, and revise their theories when the predicted outcomes do not occur.”, Starbuck (2004, p. 1249)

Like Starbuck, Arbner and Bjerke (1997, p. 306) stress the importance of an iterative approach for data collection, analysis and discussion:

“The actors approach’s presumption of socially constructed reality places the creator of knowledge in a situation that is distinct from the two other approaches. Actors creators of knowledge consciously work under the assumption that they not only change the actors, but are at the same time changed by the actors. This situation constitutes the basis of how growth of knowledge in the actors approach ought to take place (a process of mutual development that creates meaning). In this process, creators of knowledge intend to develop insights that make it possible for them to look at the situation from a new perspective, which in turn changes the initial prerequisites of the study.”, (Arbner and Bjerke, 1997, p. 306)

Scientific research in many fields today strongly focuses on frequent publication of research results. This increases the overall quality of research by increasing the amount of external review of the results and methods, either by peer review before publication or by other researchers commenting on a publication. It furthermore allows others to build their own research on the results of prior research.

Thus in all research areas that rely on publications, the contributing scholars increase knowledge in an iterative manner – even if the methodology of individual studies does not include an iterative approach.

Using a sequential (i.e., non-iterative) research approach, with a fixed sequence (e.g., literature research, method design, empirical data collection, analysis¹³) may be a suitable approach to keep research efforts short and small, allowing for more (public) iterations by publication. Using an iterative research approach *within* a research project has the

¹³The categories *exploratory* and *confirmatory* research make sense only for strictly sequential research efforts. Exploratory studies will be the preparation of confirmatory studies, and the results of confirmatory studies will trigger further exploratory research.

disadvantage that the project will have multiple stages and thus grow larger and take longer. The advantage of iterative research projects is the increase in quality *before* publication, which is particularly useful if a research stream is flooded with too many low-quality publications – too many for other researchers to read and comment on. Then the longer publication cycle time pays off in a deeper and more substantial scientific discussion.

Even in statistics there is a branch of scholars who recommend iterative model building: the advocates of Bayesian statistics (Lee, 1997) suggest that statistical results should be iteratively refined as new results and insights emerge by iteratively refining prior likelihoods to posterior likelihoods¹⁴.

In summary, the quality of scientific research stems only in part from its systematic methods. The other important driver of quality is the iterative nature of research efforts over time – either by regular publication (and public discussion with other scholars) and/or by iteration between different methods within a study.

Whether and how many iterations within a study will pay off depends on the researcher's level of knowledge, the quality and progress of existing literature¹⁵ and the quality of the researcher's current results¹⁶.

3.1.7. Methodological Approach

Building on the widely accepted minimal set of principles regarding good science from section 3.1.2 on page 84, this section merges the generic principles with those that depend on the basic assumptions from section 3.1.3 on page 86.

Researchers with a positivist worldview¹⁷ focus mostly on the general quality standards described in section 3.1.2 on page 84, frequently with a special emphasis on methods. Positivists aim for objective models as complexity-reducing representations of an objectively perceivable reality, which implies that, if the methods are sufficiently robust and accurate, the created knowledge will be an absolute and everlasting truth.

Researchers following a constructionist approach (Tsoukas, 2005b)¹⁸ also aim to meet the general quality standards – yet with a different aim for the results: an insightful, i.e. useful, yet (inter-)subjective perspective on the issue in focus, possibly including a model

¹⁴In Bayesian Statistics, the investigator starts with an educated guess (the prior likelihood), which is successively refined ('updated') by new empirical data (to a posterior likelihood). It can be shown that with sufficient updating with data, the estimated result robustly converges to the true result. Thus with sufficient data the process is even robust when starting with a false hypothesis – in which case only the convergence may be slower.

¹⁵Especially in fields in which the general quality of studies is low, an increase may improve scientific progress by deeper discussions.

¹⁶Low-quality results or only a slight increase in insight may lead the researcher to the decision to further iterate within a study.

¹⁷In the terms of Arbnor and Bjerke (1997), an *analytical* worldview, labeled as 'behavioral science' by Starbuck (2004).

¹⁸In the terms of Arbnor and Bjerke (1997), the *actors* approach.

to illustrate results. In contrast to the absolute knowledge of positivist scholars, the aim is robust and useful knowledge, but given the challenges of subjective perspectives, the knowledge created by constructionist scholars is at most inter-subjective¹⁹ rather than objective and everlasting.

Since organizations are complex systems composed of many individual and different actors, this complexity needs to be reduced by the researcher. As discussed in section 2.3.2 on page 32, complexity reduction is always subjective. Therefore, to mitigate the effect of subjectiveness, the following techniques can be applied in addition to aiming for the general quality criteria from section 3.1.2 on page 84:

- **inter-subjective** methods – merging the insights and judgment of multiple researchers (e.g., conducting research in a team during the design as well as analysis phases); and
- **iterative research** approaches – iterations within a study between analysis, discussions, further reading and further empirical data collection²⁰, leading to an iterative refinement of the researcher’s perspective, grounded in empirical observation (as described in section 3.1.6 on page 93)

3.1.8. Optimizing Methods – Cost vs. Benefit and Quality

While high-quality research results need to be an aim, the efforts in conducting the research should remain in an appropriate proportion to the quality and value of the insights aimed for. Thus when selecting a particular research method, the expected quality and value of the results²¹ should be weighed against the expected costs of using the method. Cost here includes any kind of effort associated with the method, especially the time it takes to apply it. Even though neither research costs nor success is accurately predictable in most cases, the *expected* cost effectiveness needs to be taken into account when selecting a method. Hence the value of a research approach should also be judged by its *expected return on research investment*.

Thus when selecting a method, the researcher needs to optimize for the expected value of the insights, including their quality and cost. The combined results of two different medium-quality methods may be more valuable and robust than the application of a single very high-quality and high-cost method.

¹⁹*Inter-subjective* research results are insights that are not entirely subjective, since they represent a shared consensus of a group of researchers, but since the results are only the consensus of a limited number of subjective opinions, inter-subjective results are not perfectly objective either.

²⁰In such iterations, natural experiments as suggested by Starbuck (2004) may become useful.

²¹The true value of the results could vary, e.g., by the level of detail of the insights: the results could be a series of associations or they could be a causal graph (i.e. associations with the causal direction as well).

Given that in the constructionist view of science, every finding is only a (possibly temporary) stepping stone for deeper and iterative understanding of a subject, high (i.e. cost-effective) but not perfect quality must be the aim. In addition, in order to mitigate the effects of subjectiveness, there are more quality objectives to meet, and thus a constructionist view of science will lead to other compromises when optimizing methods.

In summary, research should aim for cost effectiveness and optimize for value (including quality) against cost in order to optimize not a single study but the process of research by iteration. In addition, different basic assumptions (see section 3.1.3 on page 86) lead to different optimal compromises in the method design.

3.2. Choice of Methods for this Study

As described in section 2.2 on page 23, there is a lot of literature that is relevant to knowledge-intensive work. Yet many different perspectives (e.g., definitions of terms) are used, and there is substantial disagreement among the dominant theories from different areas of research. Furthermore, the large body of literature varies widely in quality.

Thus there is no lack of literature but rather an abundance of incoherent literature, which is why finding a suitable perspective on a problem can be particularly challenging.

A side effect of the large amount of literature is the large set of factors that have been reported to drive or inhibit learning. Therefore this study aims²² for a ranking of the most important factors affecting on-the-job learning. A ranking in turn requires a quantitative analysis – which led to the decision to conduct a single²³ survey with a broad set of potential factors. To reduce the candidate factors down to a reasonable level beforehand, insights from literature and unstructured interviews were used.

After a first literature review, a few qualitative interviews were conducted – not to get solid and presentable evidence but to support further literature research and to design the structured survey (see section 5.2 on page 139). After collecting the data from the structured survey and after overcoming some difficulties with the data analysis (see chapter 6 on page 171), the results were interpreted and compared with further literature research, which led to further refinement of the perspective (and its illustration in the form of the PIA-model).

In summary, the study was conducted with a mixed set of methods and involved discussions with other researchers during all of the following phases:

- Literature Research

²²See also the research question in section 2.7.2 on page 80.

²³In the spirit of iterative research, it would have been desirable to field multiple surveys and/or qualitative stages. But given the large number of variables, the survey results are only usable when the variables included cover a certain critical breadth, which in turn greatly increases the required sample size (critical mass). Thus a usable survey requires a substantial effort, and therefore I chose to field a single broad survey rather than multiple narrow and short ones.

3.2. Choice of Methods for this Study

- Qualitative Semi-structured Pilot Interviews and Interactive Pilots of the fully Structured Survey (including interviews to understand the participants' interpretation of the questions)
- A Single Structured Survey with a large set of potentially relevant variables²⁴ and a large sample size
- Statistical Analysis with a model for statistical inference²⁵
- Interpretation of the Results
- Further Literature Research
- Model building (as perspective setting)
- Write-up & Presentation

The sequence of methods gives this study the following properties with respect to the quality criteria described earlier (in sections 3.1.2 on page 84 and 3.1.7 on page 95):

- Mixing literature research with qualitative interviews and a quantitative fully structured survey (including a pilot study²⁶) made it possible to refine the **research** findings **iteratively** and to choose from many perspectives in literature on the use of knowledge in organizations – in order to achieve a high level of quality for the published insights.
- Multiple methods, sources and discussions with other researchers increased the robustness of results and **reduced subjective bias**.
- The decision against a detailed analysis of the qualitative interviews (involving interview transcripts) and the decision to focus on a single but wide and large sample size made it possible to **keep the research costs** (especially in terms of time for the entire study) **at a reasonable level**.
- The main empirical source of evidence (the survey) was analyzed with a **robust** statistical **method**²⁷ that semi-automatically reduces the number of variables in

²⁴The broad literature review in chapter 2 on page 21 raised many potentially relevant factors rather than yielding a single more reduced set of factors that is widely accepted.

²⁵Despite obtaining a large sample size, the challenges with noise in the data combined with the relatively large number of variables (section 5.12 on page 163) made it necessary to design a novel and more suitable statistical model-building algorithm – see chapter 6 on page 171. A new statistical method, however, was not part of the original planned sequence of methods.

²⁶The initial literature review, the qualitative interviews at the beginning of the study and the pilot study for the survey provided a necessary basis that allowed the creation and refinement of the fully structured survey. Thus the unstructured early work in the study was a first step to iteratively narrow down and refine the perspective on the problem.

²⁷see chapter 6 on page 171

the final statistical model and achieves high internally validated predictive power. Thus it yielded a **ranking of the surveyed influence factors** – an added value so far not found in the literature.

4. Statistical Theory on Algorithmic Modelling

Chapter Contents

4.1. The Theory regarding Model Selection and Fitting	103
4.1.1. The Process of Statistical Inference	103
4.1.2. Theory-based vs. Algorithmic Modeling Approaches	105
4.1.3. Overall Aim of Statistical Inference	105
4.1.4. Modeling Stochastic Processes	108
4.1.5. Automation Limits of Statistical Analysis & Causality	109
4.1.6. No Principle Difference between Model Selection and Fitting . .	111
4.1.7. Model Selection Criteria: Model Fit vs. Model Error	114
4.1.8. Variable Selection vs. Model Selection	116
4.2. Practical Challenges with Algorithmic Model Selection . . .	119
4.2.1. Measures for Model Fit with Reality (R^2)	119
4.2.2. Biased Model Fit Estimation and Overfitting	123
4.2.3. Challenges in Model Selection	128
4.2.4. Estimates for Predictive Error	129
4.2.5. Examples of Robust Algorithms and their Properties	134

As mentioned in the last chapter, the research question (section 2.7.2 on page 80) aims to rank the most relevant factors affecting learning, which implies that many potentially relevant variables need to be considered for inclusion in a statistical model. Therefore a systematic procedure to select the most important variables for the statistical model is required, which could even be automated.

Since automatic as well as manual model selection pose a number of challenges, this chapter illustrates the relevant background from statistical theory that underlie the choice and design of the statistical analysis algorithm. Those readers interested primarily in the

algorithm used for this study may skip to [section 6.2.1 on page 179](#), which has references back to this chapter.

In recent decades, many statistical regression algorithms have been developed that feature not only parameter estimation but also model selection. Frequently the central part of such an algorithm is a mechanism that automatically composes the statistical model by choosing variables included in the model from a larger subset of potentially relevant variables.

Many of these algorithms have rightfully been criticized for fragile results, which is why a number of researchers argue that data-driven model selection (sometimes referred to as *data mining*) is unsound ([Chatfield, 1995](#); [Zhang, 1992a](#)) and that many regression methods should exclusively be used for confirmation of a priori specified models ([Backhaus et al., 2006](#), p. 8). In the following, this line of argument will be referred to as the theory school of thought.

In this chapter, I will argue that parameter estimation, as performed during model fitting in ordinary linear regression, is in principle not different from model selection: both parameter estimation and model selection specify the shape of the model. Therefore I argue that in principle data-driven model selection is possible and necessary but that in practice many automatic model selection algorithms are not robust, since biased statistical estimators are used for optimization of the model, including its fit. Conversely, with low-bias statistical estimators, stable algorithms are feasible, and consequently stable regression algorithms can be used for exploratory research. This claim is supported by a detailed discussion of the challenges and causes of model-fit estimator bias, complete with a presentation of modern low-bias estimators and examples of robust model selection algorithms.

While the following discussion focuses on regression with continuous outcome variables, similar implications apply to classification with categorical outcome variables.

The mathematical arguments in this and subsequent chapters will use the following notation:

y	a scalar
$y_{i,j}$	a scalar element of a matrix, addressed by indices i, j
\mathbf{y}	a vector \mathbf{y} (in sans-serif font)
\mathbf{X}	a matrix \mathbf{X} (in capitals with sans-serif font)
$\mathfrak{D}_P(y, \mathbf{x})$	a distribution \mathfrak{D} of a stochastic process P . The distribution is described mathematically here as the probability density function (PDF), which is multi-variate over the variable vector y, \mathbf{x} .
$E_x^P[g(\mathbf{x})]$	the expected value of function $g(\mathbf{x})$ for the stochastic process P over variable x

4.1. The Theory regarding Model Selection and Fitting

To put my argument into a general context, I will first give an overview of the process of statistical inference and how the approach of the theory school of thought differs from an algorithmic approach. Next I will highlight the limits of automation in the statistical inference process – particularly with respect to causality. Next, I will present the central argument as an illustration of why there is, in principle, no difference between parameter estimation and model selection. To round the discussion off, common model-fit criteria for both parameter estimation and model selection are presented.

4.1.1. The Process of Statistical Inference

Before making a claim about algorithmic modeling, this section outlines how algorithmic modeling fits into the entire process of statistical inference and in particular why the whole process is still far from automatic but requires the researcher’s sound (and subjective) judgment at a number of steps.

Step 1: Problem Formulation Statistical analysis starts with a pragmatic formulation of a problem, which requires a wealth of knowledge about the context of the measurement (Chatfield, 2002). Therefore pre-investigations, such as exploratory qualitative research, frequently become necessary.

If the aim of analysis goes beyond generating descriptive statistics and towards learning about the mechanisms generating the data (statistical inference¹), the following questions need to be answered: Does an underlying mechanism, i.e., a stochastic process generating the data, exist? If so, how stable is this process over time? Is it stationary, i.e., independent of its history (past states of the observed system)? For social systems: Does the mechanism of interest depend on the history and perception of the involved actors – which may change? If the researcher assumes that there is no underlying and stable mechanism, e.g., when a radically constructionist position is assumed, then statistical inference has no use. In all other cases, even if the mechanism has validity only for a limited time and context and is subject to a constructed (subjective) perception of the problem by the involved actors, then statistical inference is useful in finding and/or validating a model that approximates these mechanisms (Voelpel and Meyer, 2006). For simplicity, in the following arguments, these underlying mechanisms will be referred to as “reality” or the “true” statistical process, where the quotes indicate that we might not deal with an absolute and true reality as in the natural sciences but instead with a limited subjective perspective on the true but not directly and fully observable reality. As discussed in section 3.1.3 on page 86, subjectivity is mainly

¹see section 4.1.3 on page 105 for a detailed discussion of *statistical inference*.

4.1. The Theory regarding Model Selection and Fitting

introduced by the individual choice of information selected for the investigation – e.g., through a subjective choice of issues surveyed by a structured questionnaire.

Step 2: Data Collection Next the researcher needs to decide which data is relevant and how it should be measured and collected. In survey design, this involves choosing which questions to include and how to formulate them. If all cases cannot be collected or if the entire population cannot be surveyed, the researcher in addition needs to consider how to collect a representative sample of the larger population of all cases.

A prerequisite for making these decisions is some prior knowledge or at least suspicions about the investigated mechanisms².

Step 3: Statistical Analysis The analysis may start with cleaning data. This may include removal of inconsistent data items – e.g., removal of incomplete surveys and surveys with inconsistent answers to logically connected test questions. In addition, some researchers decide to remove a small fraction of outliers, such as surveys whose results are very far away from the researcher’s hypothesized model. Next the researcher needs to choose an appropriate method of statistical analysis, given the problem and the available data. For parametric methods only, the researcher further needs to decide on and thus restrict the model’s shape - e.g. by assuming only linear relationships as a sufficiently accurate approximation. Next the model needs to be fitted to the data sample, which for parametric models is parameter estimation. Finally, the fitted model needs to be evaluated for quality (e.g., goodness of fit) and robustness, i.e., the risk of deriving false (random) results with the fitted model.

Step 4: Interpretation and Presentation of the Results The statistical results derived from the fitted model need to be interpreted. The numerical results need to be translated into a meaning within the context of the study.

The procedure for the interpretation should be as systematic as possible and should be made transparent for inspection by others³. If available, the results are compared to existing research and/or other sources of insight about the problem, such as literature or a qualitative analysis. Finally, the results need to be presented in a form that is understandable to those who use the results, e.g., for decision making.

²see also D.R. Box’s reply to and in (Breiman, 2001b) on p. 216

³For an example, see the guidelines of Brambor et al. (2006) on the interpretation of multiplicative interaction models.

4.1.2. Theory-based vs. Algorithmic Modeling Approaches

Section 4.1.1 presented a general description of the analysis process. This section covers in further detail the two principal approaches to statistical model selection and evaluation:

- purely theory-driven model selection based on an a priori specified model for confirmatory research, and
- algorithmic model selection (automatic model selection) for exploratory research.

How these two approaches differ in the analytical process is outlined in further detail in table 4.1 on the next page.

Both approaches commonly lead to a publication discussing the theory (from literature), the methods and interpretation, commonly including a critical coverage of potential weaknesses in the methods and new questions for future research.

Hence the main difference between the two approaches is model selection: The statistical model used in the theory approach is completely determined by theory. While there is an evaluation of the model (model fit and parameter significance), there is no assurance – by statistical means – that the chosen model is the best possible model. Hence the theory approach relies entirely on the correctness and completeness of the initial literature research, which needs to lead the researcher to the single “correct model” even though there might be other plausible alternative models. Thus knowledge from studies using the theory approach is only refined iteratively by publication and subsequent inspection by others and the test of time (as already discussed under the heading of iterative research in section 3.1.6 on page 93).

With the exploratory (algorithmic) approach, the researcher does not use or need strictly formulated hypotheses about cause-effect links before model fitting. He or she just needs to decide which variables to collect data for. Additional knowledge about causal links and their direction is only used at the end during the interpretation stage. Thus the algorithmic approach includes an iteration internal to the study for further refinement before exposing the results to the inspection of others and the test of time (section 3.1.6 on page 93).

Note that some mixtures of the two approaches are also possible, e.g., starting the theory approach with many alternative and plausible hypotheses.

4.1.3. Overall Aim of Statistical Inference

In *statistical inference*, the aim is primarily to learn about the mechanisms of the underlying process (the process that drives “reality”) in order to find ways to predict⁴ or

⁴Note that prediction does not require finding causal links, it just requires the detection of associative patterns and thus is frequently simpler. In the following, the focus will be on manipulation rather than prediction, since the aim of this research is to provide advice on how to design (i.e., manipulate) the organizational environment to support learning.

Analysis Process Step	Theory Approach	Algorithmic Approach
<i>Problem Formulation & Data Collection</i>	Insights from a detailed literature review , and an optional qualitative pre-study, are used to formulate strictly defined hypotheses , defining the data collection requirements and causal relationships. Thus most of the statistical model is determined at this step.	A literature review plus possibly a qualitative pre-study are used to define the data collection requirements. In contrast to the theory approach, the aim is to gain only a preliminary and broad understanding of the problem in order to define a set of potentially relevant variables from which only a subset will later be included in the statistical model.
<i>Data Collection</i>	Collected variables are strictly limited to those listed in the hypotheses.	Data collection of the set of variables with potential relevance .
<i>Statistical Analysis</i>	The mathematical structure of the model is created, based on the before formulated hypotheses and corresponding set of variables. In the case of parametric models (e.g., the multi-variate linear regression model), this also determines the principal shape of the model. Next, the model is fit to the data – i.e., the parameters are estimated. Finally, the quality and robustness of the model and/or of individual parameters are evaluated – e.g., by a model fit estimate and Student's t-test for significance of the parameters.	Depending on the available data, a parametric model might be used, which requires an assumption on model shape. For non-parametric models, this shape limitation step is omitted. In both cases the next step is automatic variable selection and parameter estimation , which are performed by an algorithmic procedure. The final step is the evaluation of model quality and robustness .
<i>Interpreting and Presenting the Results</i>	Based on the statistical results on associative patterns in the data, the hypotheses are either confirmed or rejected. If confirmed, the detected associative relationships are translated into causal claims based on the theory inherent in the hypotheses from the problem formulation stage.	Based on the model built by the algorithm rather than on hypotheses, further literature research with a new search focus follows as an iterative refinement of the initial research (section 3.1.6 on page 93). Possibly after further manipulation of the statistical model, the detected associative relationships are translated into causal relationships based on the refined theory.

Table 4.1.: Principle Approaches to Model Selection

manipulate the real process in a targeted manner. The tool for gaining a deeper understanding of the underlying process is commonly a statistical procedure or algorithm applied to collected data sample from a larger population. Modeling “reality” implies that the aim is to make claims – based on a collected sample – regarding the entire population or the future⁵ and not just the sample (from a larger finite population or past events).

Researchers therefore aim to fit statistical models to collected data from the process of interest in order to learn about the mechanisms of “reality” from the structure and shape of the model, based on the assumption that the model sufficiently reflects “reality” even though it is only built on a sample, which is smaller than the entire population. Thus the aim can be stated as follows:

Statistical inference aims at using a representative sample of a larger population to create a model, which sufficiently fits “reality” in order to allow predictions and/or inferences about the mechanisms driving “reality”.

As will be discussed in section [4.1.5 on page 109](#), statistical analyses can only find associative relationships (concurrent occurrence of events), which may or may not be based on a causal link. Thus assumptions about the causal links between independent and dependent variables of the model must be made a priori. Therefore the researcher should choose only those independent variables for the model which are assumed to have a causal effect on the dependent variable.

Noteworthy as well are stochastic processes, processes that, given a particular variable configuration, still show (limited) random behavior. An example is the stochastic process of getting lung cancer from smoking: Smoking does not (deterministically) lead to lung cancer in all cases. One can even get lung cancer without smoking. Smoking “only” strongly increases the risk of contracting lung cancer. One reason for this randomness may be that some relevant variables (latent variables) are not observed and therefore missing from the statistical model ([Hitchcock, 2007](#)). In the case of lung cancer, this might be, e.g., a yet-unknown genetic predisposition that has so far been overlooked in studies about the disease. In addition to these seemingly random but actually hidden deterministic effects, there are also true sources of randomness – such as radioactive decay⁶.

In statistical inference, there are a number of methods that deal with stochastic processes, which are described in the next section.

⁵Whenever a population is mentioned here, it should be understood in the following way: A population could be a population in an ordinary sense – i.e., a group consisting of a finite number of people. Other statistical processes have no finite maximum sample size – e.g., behavioral processes in society across generations. In these cases, the aim of statistical inference is to make predictions about the future. In other words, we build a model from past events with the aim of understanding the data-generating process, which we assume to be constant over time, in order to make predictions of the future.

⁶While the rate of radioactive decay is highly predictable, the point in time in which a particular atom decays is a truly random process not depending on any other external variables (according to the current state of scientific knowledge).

4.1.4. Modeling Stochastic Processes

In statistical inference, there are two common approaches to deal with stochastic processes:

- **Classifier Models.** The model directly models the probability of discrete events, such as the probability of getting lung cancer or not. A frequently used method is the *logistic regression* (Efron, 1986), which models the probability of discrete events, i.e., ordinal (categorical) data, with a linear model – equivalent to ordinary linear regression models dealing with metric (continuous) data. In machine learning, algorithms creating models for the probability of discrete (categorical) events are referred to as *classifier algorithms* – examples are Arcing (Breiman, 1998), Random Forest (Breiman, 2001a) and PART (Frank and Witten, 1998). Classifiers are popular in bio-informatics as well as in the social sciences⁷.
- **Expected Value Models.** In many applications, a model of the expected (i.e., average) behavior of a stochastic process is sufficient and of principal interest. In organizational research, a stochastic process may be driven by factors of the working environment and various situational factors, such as an employee's current motivation or the personal relationships of the participants in a meeting. Despite the relevance of these situational variables on the process, the researcher might not be interested in them since they cannot be used as levers for organizational improvement⁸. In such a case, the researcher might sensibly decide not to observe these variables and thus to accept some situational randomness and instead focus the statistical model-building effort on the organizational actors' mean behavior – and how it is affected by the working environment.

The most common example of an expected value model is the ordinary multi-variate linear regression (Backhaus et al., 2006). The fit of a linear regression model minimizes the error between the model and the data. It is therefore tolerant to noisy data (i.e., data from a stochastic process).

Neither of these two approaches to statistical inference fully model a stochastic process of a continuous (i.e., metric) random variable. A complete model of a continuous random variable would be a mathematical representation of a multi-variate distribution $\mathfrak{D}_P(\mathbf{y}, \mathbf{x})$. This distribution would be a function of all dependent and independent variables (vector \mathbf{y}, \mathbf{x}) related to the stochastic process. Modeling a distribution instead of the expected value (a simple scalar) would require a parametric or non-parametric model of the multi-variate probability density function describing the distribution and a large sample size to

⁷For example, in political science, a classifier algorithm may be used to predict the likelihood that a person with a particular configuration of socio-demographic variables (profession, age, preferred newspaper etc.) will vote for a particular political party.

⁸Nevertheless, it might be worthwhile to include these variables in a statistical model to reduce unexplained variance.

estimate this function. Yet so far no practical algorithms for full stochastic modeling are available. Pearl’s proposal to perform structural equations modeling (SEM) with distributions rather than correlations is a notable but nascent start (Pearl, 2003) of a discussion in this direction in the field of statistics. Alternatively, some classifier algorithms (such as Random Forest (Breiman, 2001a)) use very many categories to emulate the modeling of a continuous random variable – which, however, also increases the required sample size⁹.

Due to this study’s particular application of statistical inference, the following discussion focuses on expected value models of stochastic processes.

4.1.5. Automation Limits of Statistical Analysis & Causality

The intention of this section is to clarify that algorithmic model selection procedures can not automatically generate causal models when observational data (i.e., “natural experiments”) is used. To automatically create causal models data from truly manipulative experiments would be needed instead¹⁰. Thus with observational data algorithmic model selection can only detect association (concurrence of events or correlation) and therefore always relies on assumptions about causality.

Exceptions are more recent techniques from the field of *causal inference*, which detect causality using statistical methods – yet these methods also rely on assumptions on causality even though these assumptions are weaker and not as specific as the direction of causality for a specific link. For examples on causal inference, see (Heckman, 2005; Pearl, 2003; Rubin, 2004; Winship and Morgan, 1999).

For the sake of the central argument about model selection, the more advanced techniques of causal inference will not be covered in detail here. Instead the basic challenges with causality are illustrated with an example about studying the effect of smoking on the risk for lung cancer (following the arguments in Hitchcock (2007)):

Given observational data on smokers and non-smokers and their actual history of lung cancer (or lack thereof), the following main challenges arise when a researcher tries to infer causation:

- **Probabilistic Causation** Smoking does not inevitably lead to lung cancer; at most, it increases the *risk* of lung cancer. Thus instead of using a simple ‘*A always causes B*’ ($A \rightarrow B$) relationship model, a probabilistic model is necessary: ‘*A increases the probability of B*’. Or, if B is not a simple dichotomous outcome but instead a continuous variable, e.g., wage B is dependent on education A, then a probabilistic model of the following form is necessary: ‘*A changes the distribution of B*’.

⁹This approach has the downside that the categories, representing small bands of the continuous random variable (e.g., 0.1 to 0.17), do not have a sequence anymore; thus some information in the dataset is discarded, which in turn leads to an increase in the required sample size.

¹⁰For the research question of this study, manipulative experimental data, would require a manipulation of the organization and is thus not feasible.

- **Spurious Correlations with Latent Variables** Many challenges arise from latent (not observed) variables. For example, a correlation between yellow fingers and lung cancer may have been observed. Thus if smoking was not observed, then the researcher might be tempted to conclude that yellow fingers *cause* lung cancer. However, in that case, smoking is a latent variable that actually drives two independent statistical processes: coloring of the fingers and lung cancer. From the data alone, one may get an incorrect understanding of the causal directions. While this is an obvious example, others may be less so, and the existence of an underlying latent variable may be overlooked entirely.

The purely associative correlation between yellow fingers and lung cancer is referred to as *spurious correlation*. While there are some methods to detect spurious correlations, there are always cases in which they fail or in which weaker assumptions about causality need to be made nevertheless. The only failsafe method to detect causality is in a true experiment, where the experimenter can enforce a treatment in a truly random manner.

- **Temporal Sequence is no Guarantee for Causal Direction** Occasionally longitudinal studies are touted as the ultimate method to detect causality and its direction. However, temporal sequence is no guarantee for causality – again due to possibly overlooked latent variables, as the following example illustrates: Before a storm, the mercury column of a barometer will fall. But if we try to constrain the mercury column, the storm will occur nevertheless. In this simple example, it is obvious that the weather and atmospheric pressure are the latent variables driving both events in sequence, and the temporal sequence does not guarantee causation. Thus temporal sequence might be a good indication but not proof for causation. Determining causation from temporal sequences requires the assumption that an effect with a lead time on the event of interest is not caused by any hidden latent variables or other similar effects.

Hence causation is a serious challenge, since most scientific models provide value only by either understanding the causal links for later practical use of the resulting mental model or directly when models are used for forecasting. If the direction of causality has been misunderstood, then interventions based on the model will not have the desired effect.

In summary, the causal structure of the problem needs to be confirmed by means other than observational data. During the interpretation of statistical results from model selection algorithms, the new insights about association need to be fused with other knowledge about the causalities of the problem – e.g., from literature or a qualitative pre-study – ideally in an iterative research approach.

Along these lines, [Lukacs et al. \(2007\)](#) cite [Soule \(1987\)](#):

“Models are tools for thinkers, not crutches for the thoughtless.”

in order to emphasize that hypothesizing, and thus theoretical considerations, is “*at the heart of science*”.

4.1.6. No Principle Difference between Model Selection and Fitting

While there is lot of controversy surrounding algorithmic model selection ([Chatfield, 1995](#)), this section will show that model selection and the fitting of a parametric model (which is widely accepted) are very similar – in principle. Therefore sound algorithmic modelling is feasible in principle and it is a number of practical challenges¹¹, which cause many algorithmic model selection approaches to behave in a non-robust manner.

Given the overall aim of statistical inference from [section 4.1.3 on page 105](#), the aim of statistical model building is to fit a model to “reality”, which allows predictions of a dependent variable using a set of independent variables that has been limited to those variables in line with the a priori causal assumptions.

In [section 4.1.2 on page 105](#), two principle approaches to statistical inference were compared: the theory approach and the algorithmic approach. The following discussion focuses on the ‘statistical analysis’ step from [table 4.1 on page 106](#).

For the theory approach, [Chatfield \(1995, p. 420\)](#) describes the process of model building as consisting of the following steps:

1. model formulation/specification (including data pre-processing)
2. model (parameter) estimation
3. model validation (checking model fit)
4. combination of data from multiple sources (meta-analysis)

Under the theory approach, the set of variables included in the model is determined before the statistical analysis by the a priori developed hypotheses. Step 1 consists of a translation of the verbal hypotheses into a mathematical model and thus includes a decision on the shape of the model (e.g., a non-parametric model or a particular parametric model). In step 2, using the sample data, the shape of the model is finally determined based on the prior decision regarding shape flexibility. For parametric models, this model fitting step centers on the estimation of the model parameters, e.g., the regression coefficients in ordinary regression.

How well the resultant model fits “reality” is determined not only by step 2 (parameter estimation) but also by step 1 (model specification). Recently a number of researchers began to advocate for removing this artificial separation between parameter estimation

¹¹The details of the practical challenges are described later in [section 4.2 on page 119](#).

and model selection – as a consequent next step in improving model-based statistical analysis: Burnham and Anderson (2004), for example, calls for “*inference based on the full set of models*” [p. 262], while Chatfield (1995) and Faraway (1992) demand consideration of uncertainty for the entire modeling process, including model selection, not limited to parameter estimation only.

Therefore, for judging the success of the entire model-building process, a single standard or fitness criterion is needed, not separate standards for judging each step separately.

The parameter estimation in step 2 is commonly validated in step 3 using a criterion for the model’s fit with “reality” in the form of an estimator, such as R^2 (usually based on the collected sample). If a model fit estimator is good enough for validating the model estimation step, then it should also be good enough to validate the entire process, including the model specification step – as long as the employed statistical estimator for model fit is robust enough to assess both steps jointly. Because both steps essentially are concerned with defining the shape of the model, there is no principal reason to treat them differently.

For confirmatory research, supporters of the theory approach may argue that assessing only the success of the model estimation step is sufficient, since the specified model can be assumed to be true based on prior research – i.e., the model formulation already reflects the mechanisms of “reality” and the researcher is only aiming to quantify the variable effects.

In practical applications, statistical models will not perfectly fit “reality”: there will be some unexplained variance. While some of this variance will be due to genuinely random variation, most of the unexplained variance will stem from (non-observed) latent variables that have not been included in the model (Chatfield, 1995, p. 426) – as was discussed in section 4.1.3 on page 105. Thus the model does not perfectly reflect “reality”. Such models can be at best good approximations of “reality” – reflecting our limited knowledge of the real world.

Citing the popular saying coined by Box and Draper (1987):

“Essentially, all models are wrong, but some are useful”, p. 424

the claim that a model is true can hardly be supported in practical settings (see also (Burnham and Anderson, 2004, p. 262).)

Supporters of the theory approach may add to this that their theory-based model is, like any other model, only an approximation, but the *best* possible approximation of reality, given the available data.

However, Breiman (2001b) and Lukacs et al. (2007) (independently from each other) observe that there are multiple good models, and therefore in most cases there is not a single best model that stands out from all other good models. They argue as follows: Most statistical inference procedures aim to maximize a model fit measure. Yet these

model fit estimates have estimation errors and biases. Thus if one sees these errors as a kind of measurement error within a limited tolerance range, it becomes very likely that statistical inference yields multiple models that have model fit scores that fall within this tolerance range. Consequently, all models with a fit within the tolerance range need to be considered as equally good. Mathematically there will still be a single most optimal model, however it must be considered *equally* good compared to any other models within the tolerance range. See Breiman (2001b, p. 206) for an example.

Yet if an underlying stochastic process is truly understood theoretically, the theory approach with a theory-based model may be used for a quantification of the effect strengths of the different terms. However, researchers who use statistical models, which are determined and fixed a priori, to confirm their hypotheses run the risk¹² of overlooking other theoretically plausible models that fit the data better or equally good within the tolerance of the model fit estimate. Hence to avoid this risk, assessing (i.e., validating) the entire process of model building, including model specification, becomes a necessity. For this assessment, predictive power with respect to “reality” (i.e., model fit) suggests itself as a quality criterion for both steps: model selection and model fitting.

A number of researchers have correctly pointed out that model selection using conventional R^2 estimates or similar measures is very unstable in a number of applications (mostly depending on properties of the data) – see Anderson and Burnham (2002); Breiman and Spector (1992); Chatfield (1995); Grünwald (2007); Kapetanios (2007); McCann and Welsch (2007); Yuan and Yang (2005); Zhang (1992a). These examples demonstrate that a particular algorithmic model selection method is not sufficiently robust, but that does not imply that automatic model selection in general is impossible. These studies merely highlight the importance of making the model selection algorithm and the involved statistical estimators robust enough for different kinds of data.

The supporters of the theory approach often refer to these examples as illustrations that exploratory research using model-based statistical inference is *in general* not possible, and that most methods lend themselves *only* to confirmatory analysis (Backhaus et al., 2006, p. 8). This is true for a wide range of conventional methods (such as ordinary linear regression) and early algorithmic approaches (e.g., step-wise regression based on t-test to enter/delete (Breiman, 1992)). Especially for noisy and high-dimensional data, the robustness requirement is often difficult to meet – as section 4.2.2 on page 123 will show. Thus starting with a fixed model based on theory may be a very good and practical choice if only weaker algorithmic methods are available. However, showing that many early algorithmic approaches fail the requirements for robustness does not imply that model-based exploratory research is infeasible *in general*.

¹²Running this risk might be an acceptable and good research design choice when all other design alternatives exhibit other and more severe weaknesses and downsides – as discussed in section 3.1.8 on page 96.

Thus if an algorithm meets the robustness requirements, it can be a useful tool to automate model selection and fitting – all within a limited space of plausible models. The limit is necessary since the detection of causality can not be automated, as discussed in section 4.1.5 on page 109.

In an algorithmic modeling approach, instead of fully specifying the model form and variables a priori based on assumptions, the researcher would make assumptions on which variables may be relevant and thus need to be collected. Furthermore, he or she would make assumptions on how the causal linkages are directed but would not assume how and which of these variables enter the final model.

In summary, if automatic parameter estimation is possible, then explorative automatic model selection must also be possible – as long as the involved algorithms and model fit criteria are sufficiently robust. It follows that the model-building process as a whole needs to be assessed by a single model fitness criterion, which is common for both the model specification and estimation step. However, designing sufficiently robust model selection algorithms is difficult, which frequently makes theory-based model building a viable alternative. Yet there are examples of robust model selection algorithms, which will be presented in section 4.2.5 on page 134.

4.1.7. Model Selection Criteria: Model Fit vs. Model Error

In the preceding section I argued for a single model fit with “reality” criterion for both model selection and fitting. The most common measure of model fit is the coefficient of determinance: R^2 . It is popular because the meaning of R^2 is easy to understand and because there exists a simple estimation method.

Yet since there are challenges with robustness of the simple R^2 estimation method (for certain types of data), recent statistical literature has featured a discussion on generalizing model fit criteria and improving robustness – on which this section contains an overview.

In general, for expected value model, model fit criteria can be divided into two classes:

1. **Predictive Error** measures – gauge the average *mismatch* between the statistical model and “reality”, including random noise. Analogous and in the same criteria class, **model fit** measures gauge the *fit* of the statistical model with “reality”.
2. **Model Error** measures – gauge the mismatch between the current statistical working model and an ideal expected value model¹³.

Predictive Error *Predictive error* and *model fit* measures (such as R^2) estimate the mismatch or fit of the statistical model with “reality”. Since the statistical model is an expected value model here, the measure will include model inaccuracies – e.g., due to

¹³ *Expected Value Models* – as discussed in section 4.1.4 on page 108.

linearization¹⁴ and the average mismatch due to random “noise” from stochastic processes (see sections 4.1.3 and 4.1.4 on p. 105).

Thus, for most practical applications, predictive error model fit criteria are preferable, since they estimate the predictive power (and accuracy) of the statistical model for real-life applications, which may include random, i.e., stochastic, effects.

Model Error *Model error* measures estimate the difference between the model and an ideal model that perfectly reflects the expected value of the stochastic process – i.e., “reality” without any random noise. Hence a measure of model error assesses the accuracy of the stochastic modeling process compared to the maximum achievable accuracy with an expected value model.

Thus model error measures are more suitable for analyzing the performance of statistical fit algorithms with different datasets, since the measure neutralizes the effect of noise that may be specific to a dataset.

An example of a model error measure, used for ranking different models, is introduced in Breiman (1996b):

$$PL = PE(\text{Working Model}) - PE(\text{Ideal Model}) \quad (4.1)$$

PE is the *predictive error* – here, the sum square errors estimated with a reduced bias using a test set. The *ideal model* is the true underlying (expected value) model to which truly random noise is added. PE(Ideal Model) is the model fit of the *ideal model* with the data including random noise¹⁵. A similar measure – derived from information theory – is the Kullback-Leibler information quantity, which measures the difference between a true model and a particular working model (Burnham and Anderson, 2004, p. 267).

Suitability for Model Fitting As will be shown in the following, both model fit criteria classes lead to similar results when used as objective function in model fitting.

As noted in Burnham and Anderson (2004), the true stochastic process is unknown in most practical applications. For selecting the best statistical model for one and the same true process, it is not necessary to know the ideal model, since the ideal model term in the Kullback-Leibler information quantity (Burnham and Anderson, 2004, p. 267) and in Breiman’s predictive loss measure remains constant: from equation 4.1 and the fact that the ideal model is fixed for a particular application, i.e., $PE(\text{Ideal Model}) = \text{const.}$, it follows that minimizing the predictive loss and minimizing the predictive error of the

¹⁴Model inaccuracies are discussed in more general detail in section 4.2.2 on page 123 under the header *model flexibility*.

¹⁵Since the real process is noisy, there is no perfect expected value model that perfectly predicts the actual data. Thus the PE(Ideal Model) fit will be high but not necessarily perfect (i.e., 100%).

working model both lead to the same parameter estimates and model selection result:

$$\arg \min \text{PL} = \arg \min \text{PE}(\text{Working Model}) \quad (4.2)$$

Therefore Breiman suggests minimizing the predictive error, which is maximizing the model fit R^2 based on a robust estimator with little bias. A number of authors agree that a robust predictor of the model fit with “reality” is a practical and suitable joint model selection and parameter fitting criterion (Efron, 1986; Faraway, 1992; Li et al., 2006; Zhang, 1993).

Summary As discussed in the previous section, algorithmic model selection is feasible when based on a robust estimator for the model fit – i.e., both model selection and model fitting performance need to be optimized for and validated by a single and common robust quality criterion. Recent statistical literature contains proposals for a number of new model fit indicators, which fall into two abstract categories: *predictive error* and *model error*. Despite the differences of the two categories, using two equally robust and accurate measures from each category will yield the same model selection and fitting results – as long as frequently applicable assumptions hold. Practical examples for robust predictive error measures are presented later in this chapter in section 4.2.4 on page 129.

4.1.8. Variable Selection vs. Model Selection

Variable Selection / Parameter Selection Many model selection and fitting algorithms first select a model, then fit the model, and finally compare multiple models and select the best (or a few good ones)¹⁶. Some algorithms further reduce the problem to automatically selecting variables independently from the rest of the current model. A classical example is step-wise regression, which uses Student’s t-test for deciding whether to delete or add a variable and the associated regression parameter to the current model (Breiman, 1992; Faraway, 2002; Miller, 1984). More recent examples are forward step-wise regression with AIC as model fit criterion (Atkinson and Riani, 2007) or with a bootstrapped model fit criterion (Breiman, 1992), least angle regression (LARS) (Efron et al., 2004), and Lasso (Lutz and Buhlmann, 2006).

While these variable selection approaches, may work fine for certain types of stochastic processes (i.e. kinds of data) and model types, there are two principle challenges with variable selection – that may lead to erroneous results with other types of data and models:

¹⁶There are a few notable exceptions. Breiman’s random forest grows a number of tree-like models, involving variable selection and fitting at every step (Breiman, 2001a). Similarly, the decision-tree-based algorithms *PART* (Frank and Witten, 1998) and *cForest* (Strobl et al., 2007) combine model selection and fitting. Other examples are support vector machines (SVM) (Breiman, 2001b; Rasmussen and Williams, 2006) and neural networks (Guyon, 2007; Witten and Frank, 2005).

collinearity of variables and non-independence of the model terms. Unfortunately, the data of this study belong to the latter group.

Challenge 1: Collinearity Collinearity occurs if information relevant to the stochastic process is shared by two or more variables. The simplest case is when two variables correlate. In that case, collinearity is easily detectable [Backhaus et al. \(2006, p. 90\)](#). Yet there are also cases in which process-relevant information is shared by three or more variables. This is commonly referred to as *multi-collinearity*, which can be detected by removing and adding variables to a model and monitoring the estimates for the parameters related to the other variables. If multi-collinearity is present for the added/removed variable(s), the parameter estimates will change for all other variables affected by this collinearity (see [Backhaus et al. \(2006, p. 91\)](#) and [Brambor et al. \(2006, p. 70\)](#)).

Hence strong multi-collinearity causes the affected parameter estimates in ordinary regression to become fragile (non-robust) – even though the affected terms themselves are valid effects with a true impact on the stochastic process ([Brambor et al., 2006, p. 70](#)). Thus strong multi-collinearity may bias and weaken stability¹⁷(significance) measures, such as Student’s t-test, that measure the stability of an individual term included in the current model. A weakening of a term’s stability due to collinearity may be acceptable if the data otherwise contains little noise. It is foremostly the combination of collinearity, noise and small sample sizes that causes the problem to become critical.

Therefore, if a model contains sufficiently strong multi-collinear variables, the parameter estimates and the assessments of the parameter stability (significance) of the collinear variables become dependent on the composition of the rest of the model – i.e., a parameter may be stable in conjunction with a particular group of independent variables but unstable when combined with another group of variables. In step-wise regression, for example, finding that a collinear parameter is insignificant does not imply that it is insignificant in all models. Thus the significance test cannot be used to remove a variable permanently for the rest of the search. Instead of seeking the best variables/parameters in the model, the search for the best model needs to be enlarged to assess all possible combinations of variables in a model.

Guyon et al., in their work on feature selection in the field of machine learning, claim more generally that uni-variate feature selection, i.e., variable selection, does not provide useful and reliable information for robust model selection with collinear data; instead, only full models need to be compared. Yet in practical applications, the number of possible unique models can be very large: e.g., rather than testing 30 parameters individually, performing a full model search in order to choose 10 out of 30 potentially relevant variables, more than 30 million models need to be tested. Therefore Guyon et al. suggest a filtering

¹⁷For the BOGER algorithm developed for this study, a stability measure is defined in text box [6.2.1 on page 190](#) as an alternative to conventional statistical significance tests.

approach in which the potentially relevant variables are screened first, and the number of candidate variables for the model are reduced before beginning with a full model search (Guyon, 2007; Guyon et al., 2006).

Challenge 2: Models with Non-Independent Terms The second challenge with by-variable model selection is the type of model: The mathematical model of ordinary linear regression is composed of completely independent terms:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + \dots \quad (4.3)$$

Here b_i are the regression parameters and x_i the independent variables.

Hence removing a term will not change the behavior of the model due to changes in any of the other terms¹⁸. The interaction term x_1x_2 is no problem either, since it can be treated as a new and separate variable. However, when either x_1 or x_2 is removed, the interaction should be removed as well (Brambor et al., 2006).

In contrast to such a model with independent terms, already small changes to the model cause the terms to depend on each other:

$$y = b_0 + b_1x_1 + b_2x_2^{\gamma_2} [b_3x_3 + b_4x_4 + \dots] \quad (4.4)$$

Here both b_i and γ_i are model parameters that need to be estimated.

If x_2 is removed, then any insights regarding the parameters b_3 and b_4 are invalidated. Thus, similar to the above described treatment of interactions, either the terms $b_3x_3 + b_4x_4 + \dots$ are removed or their stability needs to be retested.

The BOGER model – developed for this study (see section 6.2.2 on page 180) – uses multiplicative terms in order to allow for modeling AND-relationships (i.e., an effect on the outcome is only caused if factor A AND B are both strongly present with a large numerical value). Hence by-variable selection is only acceptable in a screening phase for the BOGER algorithm.

Summarizing ... In summary, algorithmically selecting a model by assessing the stability of individual terms (by-variable or by-parameter model selection) leads to optimal models (in terms of model fit and robustness) only under the following conditions: 1.) The independent variables are only weakly collinear, and 2.) the model is composed of mutually independent terms¹⁹.

¹⁸The only exception is the constant offset term b_0 , which is likely to change upon removal of another term. Yet the offset term is frequently not very relevant for interpretation.

¹⁹Unless a special term adding/removing strategy is used, as described before in this section with regard to interactions.

If these conditions for by-variable or by-parameter model selection are violated, a full model search strategy needs to be used in order to search the large space of all possible models, consisting of all possible variable combinations. To improve computational efficiency, by-variable selection with weak criteria may nevertheless be useful in a screening stage that precedes a full model search.

4.2. Practical Challenges with Algorithmic Model Selection

Previously, section 4.1 on page 103 discussed theoretical issues regarding model selection. Section 4.1.6 on page 111 further pointed out that many model selection algorithms are not robust, and thus there is an unacceptably high likelihood that the results are strongly biased. This section will first illustrate how and why the statistical estimators for model fit can become severely biased – dependent on the data and the fitted model. Next, it will discuss the effect of biased estimators on model selection. And, finally, it will present a number of robust model fit (i.e., predictive error) estimators as well as examples of robust algorithmic model selection algorithms that utilize these estimators and are thus much less vulnerable to the various model-fit estimation biases.

4.2.1. Measures for Model Fit with Reality (R^2)

As discussed in section 4.1.7 on page 114, there are two categories of model fit criteria: *predictive error* and *model error*. In the following, a formal and general mathematical definition of model fit with “reality” will be presented. Hence the focus will be on predictive error measures, given their suitability for practical applications.

As discussed before in section 4.1.6 on page 111, robust algorithmic model fitting critically depends on a low level of bias in the model fit estimation procedure. One reason for this is the use of the model fit criterion as an objective function for automatic model fit optimization. Therefore, as an illustration, model fit – as measured by the popular coefficient of determinance R^2 – is defined in a general (non-approximate) mathematical form followed by an analysis of the most common estimation procedure for R^2 .

A principal challenge in low-bias estimation is to obtain an estimate that robustly generalizes beyond the collected sample to the rest of the entire population. As outlined in section 4.1.3 on page 105, statistical models serve as approximations for the “true” underlying statistical process P driving “reality”, and they are used to understand the statistical process in a quantitative way and/or to make predictions. In both cases, the model resulting from this analysis needs to have predictive power beyond the collected survey sample in order to model the underlying mechanism (“reality”) in a sufficiently accurate manner. Thus the model should be based on a sample that is sufficiently representative of the total population in addition to a good model fit (with the fitting data

sample). Fulfilling these two requirements, the model would have a low *sampling bias* – i.e., building another model based on another representative sample would lead to similar results and represent the underlying “real” mechanism well. A model with sufficiently high predictive power is a prerequisite for further analysis of the model and interpretation of the results, e.g., for getting low-bias estimators for the variable effect strengths.

The model’s fit with the collected sample data, measured, e.g., by the coefficient of determinance R^2 , is commonly used as a measure for predictive power, based on the assumption that the collected sample is approximately representative for the entire population and thus the underlying stochastic process P ²⁰.

In a general and mathematically exact (i.e., non-estimated and non-approximate) form, the *coefficient of determinance* R^2 is defined as²¹:

$$R^2 = \frac{\text{explained variance}}{\text{total variance}} = 1 - \frac{\text{unexplained variance}}{\text{total variance}} = 1 - \frac{E_{y,\hat{y}}^P [(y - \hat{y})^2]}{E_{y,\bar{y}}^P [(y - \bar{y})^2]} \quad (4.5)$$

where y is the random outcome variable driven by the stochastic process P , \hat{y} is the prediction for the dependent variable using the model and \bar{y} is the average of the dependent variable²². The residuals $y - \hat{y}$ are squared, which disproportionally penalizes larger deviations of the model from the sample data compared to smaller deviations – in principle, a desirable property. However, penalizing large deviations more than small makes this measure also less robust to the impact of outliers that have large deviations, even though they might represent only a small fraction of the sample.

Thus an alternative measure would be a coefficient based on the absolute values of the residuals, which in the following will be referred to as R the *absolute sum coefficient of fit* or R_{abs} . Its general and precise mathematical definition is:

$$R_{\text{abs}} = 1 - \frac{E_{y,\hat{y}}^P [|y - \hat{y}|]}{E_{y,\bar{y}}^P [|y - \bar{y}|]} \quad (4.6)$$

This measure treats all data points equally and is thus more robust to outliers.

R^2 has nevertheless become the de facto standard to measure model fit, since squaring the residuals has one important convenience advantage: in ordinary linear multivariate regression, the square residuals are minimized²³ instead of the absolute residuals, which

²⁰Citing (Backhaus et al., 2006, p. 64): “Nachdem die Regressionsfunktion geschätzt wurde, ist deren Güte zu überprüfen, d.h. es ist zu klären, wie sie als Modell der Realität geeignet ist.” (“After the regression function is estimated, its goodness must be assessed, i.e., it needs to be investigated, how well it [the regression function] serves as model of reality.”) Backhaus et al. (2006) then proceed to argue for the use of R^2 and by-parameter tests such as the student t-test.

²¹See also (Backhaus et al., 2006, p. 66)

²²Strictly speaking, \bar{y} , the true mean of the population, is unknown as well and thus must be estimated, e.g., by the mean of the sample.

²³The solution to ordinary linear multivariate regression minimizes the quantity: $S = (Y - \mathbf{bX})(Y - \mathbf{bX})$

allows for a closed-form solution for the coefficients. This can be calculated very quickly using a simple and computationally very efficient matrix operation²⁴.

Following this de facto standard, model fit will be measured by both R^2 and R_{abs} in this study. In contrast to the common use of R^2 , however, this study will use multiple different methods to estimate the true R^2 for the whole population in order to reduce estimator bias – as will be detailed later.

At this point, it is worth noting the way in which R^2 and R_{abs} are commonly estimated. The challenge in evaluating the expressions in equations 4.5 and 4.6 is, in both cases, to estimate the expected value $E^P[\cdot]$ of the deviation between model and data with little bias. In the general definitions, R^2 and R_{abs} differ only in the deviation term, which is defined for the following arguments as function $g(\mathbf{x}, y)$: For R^2 the deviation function $g(\mathbf{x}, y) = (y - \hat{y}(\mathbf{x}))^2$, while for R_{abs} the deviation function $g(\mathbf{x}, y) = |y - \hat{y}(\mathbf{x})|$. In both cases, y refers to the sample data and where $\hat{y}(\mathbf{x})$ refers to the current model, which functionally depends on the independent variable vector \mathbf{x} .

In general, $E_{\mathbf{x},y}^P[g(\mathbf{x}, y)]$, the expected value of any function $g(\mathbf{x}, y)$ for a stochastic process P , is given by:

$$E_{\mathbf{x},y}^P[g(\mathbf{x}, y)] = \int_{\text{any } \mathbf{x}, y} \mathfrak{D}_P(\mathbf{x}', y') g(\mathbf{x}', y') d(\mathbf{x}', y') \quad (4.7)$$

where $\mathfrak{D}_P(y', \mathbf{x}')$ is the multi-variate distribution describing the stochastic process P in the form of a multi-variate probability density function (PDF) (Lee, 1997, p. 13) over random variables y', \mathbf{x}' (the prime indicates that the variables are helper variables for the integration). The integration is performed over all possible combinations of \mathbf{x}, y that are valid in the population generated by the process P . Unlikely combinations of \mathbf{x}, y will be multiplied with the very low probability density value and thus will hardly affect the end result of the integration.

Since accurately estimating $E_{\mathbf{x},y}^P[g(\mathbf{x}, y)]$ is the key to estimating the model fit²⁵, the challenge in calculating the model fit is to obtain the true multi-variate representation \mathfrak{D}_P of the stochastic process P and to evaluate the integral over all possible values. Since in most cases the true distribution \mathfrak{D}_P is unknown²⁶, statisticians commonly estimate the

where the data \mathbf{X} and \mathbf{Y} has been transformed linearly to have zero mean (Backhaus et al., 2006, p. 115).

²⁴The R^2 -maximizing solution of the regression coefficient vector \mathbf{b} is $\mathbf{b} = \arg \min_{\mathbf{b}} S(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} \cdot \mathbf{X}'\mathbf{Y}$ (Backhaus et al., 2006, p. 116). Since R^2 is directly proportional to S for a given and fixed sample, R^2 is the natural model fit measure for linear regression and only distinct from S by a normalization that allows comparison across datasets.

²⁵See equation 4.7.

²⁶The true distribution \mathfrak{D}_P completely describes the random process P if it is multi-variate over all relevant variables. Thus if the researcher knows this distribution, he or she has a more detailed description of the process than an expected value model (section 4.1.4 on page 108). For example: Pearl (2003) suggests going beyond linear expected value models in structural equation modeling by using multi-variate distributions. Thus when the distribution \mathfrak{D}_P is known, there is no point in further

model fit using a representative sample generated by P – as is described in the following:

When a random sample of size n from a larger population of size N is available and sufficiently large²⁷ compared to N , it can be assumed that it is approximately representative of the entire population. It follows that the distribution of the sample is an approximation of the distribution of the population. Then a practical method to estimate the value of the integral in equation 4.7 is to perform a Monte-Carlo integration (Robert, 2005) by first randomly drawing samples $(\mathbf{X}, y)_{\text{gen}}$ from the distribution of the collected sample $\mathfrak{D}_P(\mathbf{x}, y)$. Next, artificial samples g_i for $g(\mathbf{x}, y)$ are calculated using the previously generated artificial samples $(\mathbf{X}, y)_{\text{gen}}$. Finally, the g_i samples are averaged to give an estimator for the integral in equation (4.7)²⁸. The simplest method of drawing samples from the distribution of the survey sample is to simply use the samples contained in the sample, rather than modeling the distribution (e.g., by kernel fitting), and then generating samples from this model of the distribution. Using the simple method, the Monte-Carlo integral simplifies to the frequently used *mean estimator* for the expected value $E_{y, \mathbf{x}}^P[g(y, \mathbf{x})]$:

$$\tilde{E}^P[g] = \frac{1}{n} \sum_{i=1}^n g_i \quad (4.8)$$

where the samples g_i have a distribution that approximates \mathfrak{D}_g^P .

The simple estimator in equation 4.8 has the appropriate convergence property of statistical estimators: as the sample size n converges to the size of the entire population N , the estimator $\tilde{E}^P[g]$ will also converge to the true value of $E^P[g]$ ²⁹.

With this estimator, equations 4.5 and 4.6 for R^2 and R_{abs} simplify to:

$$R^2 = 1 - \frac{\tilde{E}^P[g^2]}{\tilde{E}^P[(y - \bar{y})^2]} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.9)$$

and

$$R_{\text{abs}} = 1 - \frac{\tilde{E}^P[g]}{\tilde{E}^P[|y - \bar{y}|]} = 1 - \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i - \bar{y}|} \quad (4.10)$$

While the previous paragraphs may be a rather complicated explanation for the simple mean estimator, it sheds light on how the estimator can be biased: the estimate of the multivariate distribution \mathfrak{D}_g^P based on a sub-sample with size $n \ll N$ will be biased, since the sub sample is hardly ever perfectly representative. It is this bias in the estimation process of the distribution that also biases the estimator itself. Thus the choice of the data

statistical modeling.

²⁷Sufficiency of sample size depends on the type of data-generating process and in particular how much noise it adds, as will be detailed in this and following sections.

²⁸Alternative and more accurate discrete integration methods are *Simpson's rule* or the *Trapezoidal rule*.

²⁹For another statistical estimator van der Laan (2006) includes a mathematical discussion of estimator convergence on p. 10.

used by the estimator is pivotal for its accuracy. For any small sample size $n \ll N$ the estimator will likely be biased, which is the reason that other estimators are considered later – see section 4.2.4 on page 129.

The meaning of the different possible values for R^2 is illustrated in table 4.2. The interpretation of different values of R_{abs} and the limits ($-1 < R_{abs} < 1$) are the same as shown in table 4.2, with the only exception that the absolute values of R_{abs} will generally be lower.

$R^2 = 0$	The model is as good for prediction as simply using the mean of Y for prediction without regard for the independent variables X – i.e., the model has no value.
$0 < R^2 < 1$	The model is a better predictor than the mean. The value of R^2 is the fraction of the variance around the mean and explained by the model.
$R^2 = 1$	The model is a perfect predictor for the data Y when X is known.
$-1 < R^2 < 0$	The model is a worse predictor than the mean – i.e., the model has no value. This never happens for ordinary linear regression or when estimating R^2 using the data the model was fitted to, since minimizing S always leads to $R^2 \geq 0$. Hence $R^2 < 0$ is usually a sign of a defunct fit algorithm.
$R^2 > 1$ or $R^2 < -1$	Not possible.

Table 4.2.: Meaning of Different Values of R^2

In summary, this section presented a general mathematical definition of R^2 and R_{abs} and derived the most commonly used estimator for both model fit measures. This description forms the theoretical basis for the discussion in the following sections on how this simple model fit estimator can lead to biased results.

4.2.2. Biased Model Fit Estimation and Overfitting

The theoretical details of the previous section gain practical relevance when the interaction between the model fit estimation method and the model fitting process itself is considered. This section will illustrate how the bias in the model fit estimator used in ordinary regression may negatively affect the model fitting process and lead to *overfitting*. Note that for illustration of the effect, it is assumed in this section that the statistical model is specified before the model fitting step. The following section will cover the additional effect of model selection.

Note that the effect described below is small and negligible when the sample size is large

and few variables are considered. Unfortunately, many common social science applications are complex problems with many variables, and sample sizes are often relatively small. Empirical evidence for overfitting due to a high model flexibility stemming from a relatively large number of variables can be found in sections 6.1.4 on page 177 and A.6.3 on page 310.

In ordinary linear multivariate regression, the following quantity is minimized to obtain the parameter estimates $\hat{\mathbf{b}}$ (using the commonly used simple mean estimator for the expected value from equation 4.8 on page 122)³⁰:

$$\mathbf{b} = \arg \min_{\mathbf{b}} E_{y,x}^P [(y - \hat{y}(x))^2] \quad \text{for the entire population} \quad (4.11a)$$

$$\hat{\mathbf{b}} \cong \arg \min_{\mathbf{b}} \int_{\text{any } x,y} \hat{\mathfrak{D}}_{x',y'}^P (y' - \hat{y}(x'))^2 d(x', y') \quad \text{with an estimated distribution} \quad (4.11b)$$

$$\cong \arg \min_{\mathbf{b}} [(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})] \quad \text{over the collected sample} \quad (4.11c)$$

$$\cong \arg \min_{\mathbf{b}} [(\mathbf{Y} - \mathbf{bX})(\mathbf{Y} - \mathbf{bX})] \quad \text{over the collected sample} \quad (4.11d)$$

$$\hat{b}_j \cong \arg \min_{b_j} \sum_{i=1}^n (y_i - \hat{y}_i(b_j, x_i))^2 \quad (4.11e)$$

$$\cong \arg \min_{b_j} \sum_{i=1}^n \left(y_i - \sum_j b_j x_i \right)^2 \quad (4.11f)$$

where equation (4.11a) precisely describes the unknown true optimum, while the other equations describe the estimation procedure in various equivalent forms. Any estimation procedure will in some way estimate $\hat{\mathfrak{D}}_{x,y}^P$ – equation (4.11b). The estimation of the distribution is most commonly performed by using the collected samples as described in the following equations: equations (4.11d) and (4.11c) are simply the matrix versions of equations (4.11f) and (4.11e).

Note that the quantity $E_{y,x}^P [(y - \hat{y}(x))^2]$ is directly proportional to R^2 and thus the minimization process shares the same challenges with low-bias estimation with the simple R^2 estimators. Thus the model fit optimization described above is equivalent to minimizing R^2 .

The completeness of the coverage of the distribution \mathfrak{D}^P by the collected sample is essential for a low-bias model fit estimate. The coverage becomes especially sparse, and thus the estimate more strongly biased, if the sample size n is relatively small compared to the number of independent variables m : In order to estimate the model fit, the distribution

³⁰ $\arg \min_{\mathbf{b}}$ refers here to a minimization of the quantity after the $\arg \min$ operator by adjusting \mathbf{b} during the optimization. The entire expression evaluates to the optimized values of \mathbf{b} – labeled here as $\hat{\mathbf{b}}$.

\mathfrak{D}^P needs to be estimated over a variable hyperspace of dimensionality m – i.e., a multi-variate PDF for \mathfrak{D}^P with m -dimensions needs to be estimated. When the sample size n is small compared to m , there will be a very sparse point cloud of sample data in the vast variable hyperspace, which leads to inaccuracies and bias in the estimation of \mathfrak{D}^P and thus also to a bias in the model fit estimate.

This sparseness effect will only lead to a minor bias for R^2 model fit estimates that are estimated based on an *independent* data sample, which was not used during model fitting. This “*natural*” bias is amplified, however, if R^2 or a proportional expression³¹ is used as an optimization goal function during model fitting – as the following illustration will show.

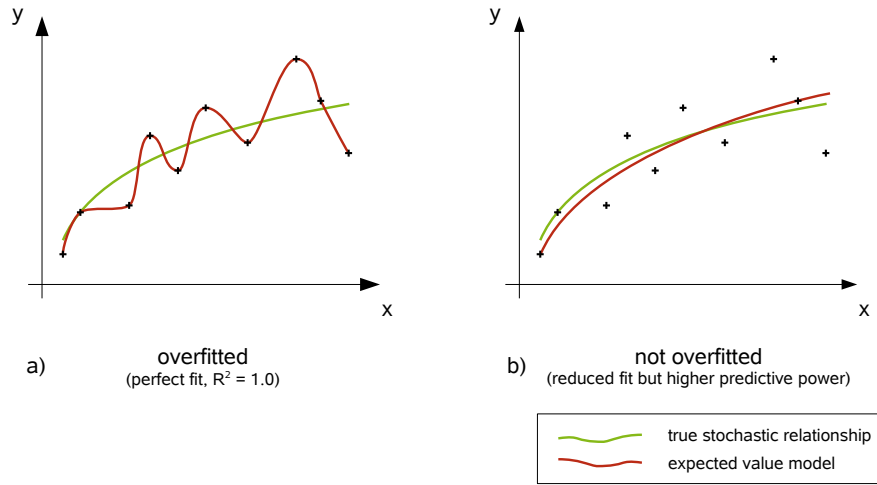


Figure 4.1.: Overfitting - A Graphical Example

The effect of this sparseness on model fitting is illustrated in figure 4.1 showing the same – slightly noisy – data in part a) and b) of the figure. When the algorithm fits an expected value model by aggressively maximizing the R^2 based on the sparsely modeled distribution – as shown in part a) of the figure – then the model (in red) is *overfitted*, i.e., the fit to the collected sample is much better than the real relationship (in green)³² (Breiman, 1996b, p. 2351). In contrast, figure part b) shows a non-overfitted model (in red), which is much closer to the real relationship (in green), but gets a lower R^2 model fit estimate, since the true randomness of the underlying process reduces the R^2 model fit estimate.

This is rather disturbing, since the conventionally estimated R^2 suggests a very good

³¹As shown in equation (4.11f).

³²In figure 4.1, a non-linear model is used for illustration. In the case of ordinary linear regression, the overfitting effect would look different but is still present and similar, especially when the model uses many variables and thus also has many parameters, giving it a high degree of freedom.

fit, but the overfitted model actually fits reality much worse than what the R^2 value suggests, and thus it also fits new samples³³ much worse than a simpler model would³⁴ – see part b) of figure 4.1. This coincides with common intuition that suggests that the model in part b) will be much more reliable for predictions of new samples collected after model building than the model in part a). Citing Chatfield:

“Although a more complicated model may appear to give a better fit, the predictions from it may be worse. The dangers of overfitting are ‘well-known’, particularly in multiple regression [...], but these dangers are not always heeded.”,

Chatfield (1995, p. 429)

Hence the combination of many variables with comparatively few samples leads to an overly sparse modeling of the statistical process’ true distribution \mathfrak{D}^P . The use of this sparse sample within the simple mean estimator, in combination with aggressive optimization, leads to a strong overestimating bias for the R^2 and an overfitted model. The optimization and the weaknesses of the R^2 estimator interact in a disadvantageous manner that leads to a bias far exceeding the abovementioned “natural” bias of the conventional R^2 estimator (Efron, 1986). The principal problem is that estimator for R^2 and the fitted model become dependent on each other during the model fit optimization process. Thus, conversely, an independent estimator for model fit (even when slightly biased) will yield results much closer to the model fit with reality (rather than the sample data). Examples of such independent model fit estimators are presented next in section 4.2.4 on page 129.

While this problem with the ordinary R^2 model fit estimator is central to overfitting, it only occurs when a number of conditions are present simultaneously:

- **little data** – too few samples compared with the number of variables, leading to the distribution sparseness and model fit estimation issues discussed above.
- **noisy data** – the true relationship is not entirely deterministic or some relevant variables are missing from the model – possibly since they have not been or could not be collected³⁵.
- **unevenly distributed data** – even if the absolute number of samples is sufficiently large, the samples may be distributed unevenly over the range of an individual variable – see dataset 2 in figure 4.2 on the next page and the explanations below.

³³New samples are either additional samples from a larger population or are generated by the process in the future.

³⁴With reference to the discussion from section 4.1.7 on page 114, the simpler model has a less than perfect prediction error but has a model error of almost zero – implying that it is close to the best possible expected value model, given the noise in the data.

³⁵See also the discussion regarding spurious correlations and hidden latent variables in section 4.1.5 on page 109 on causality.

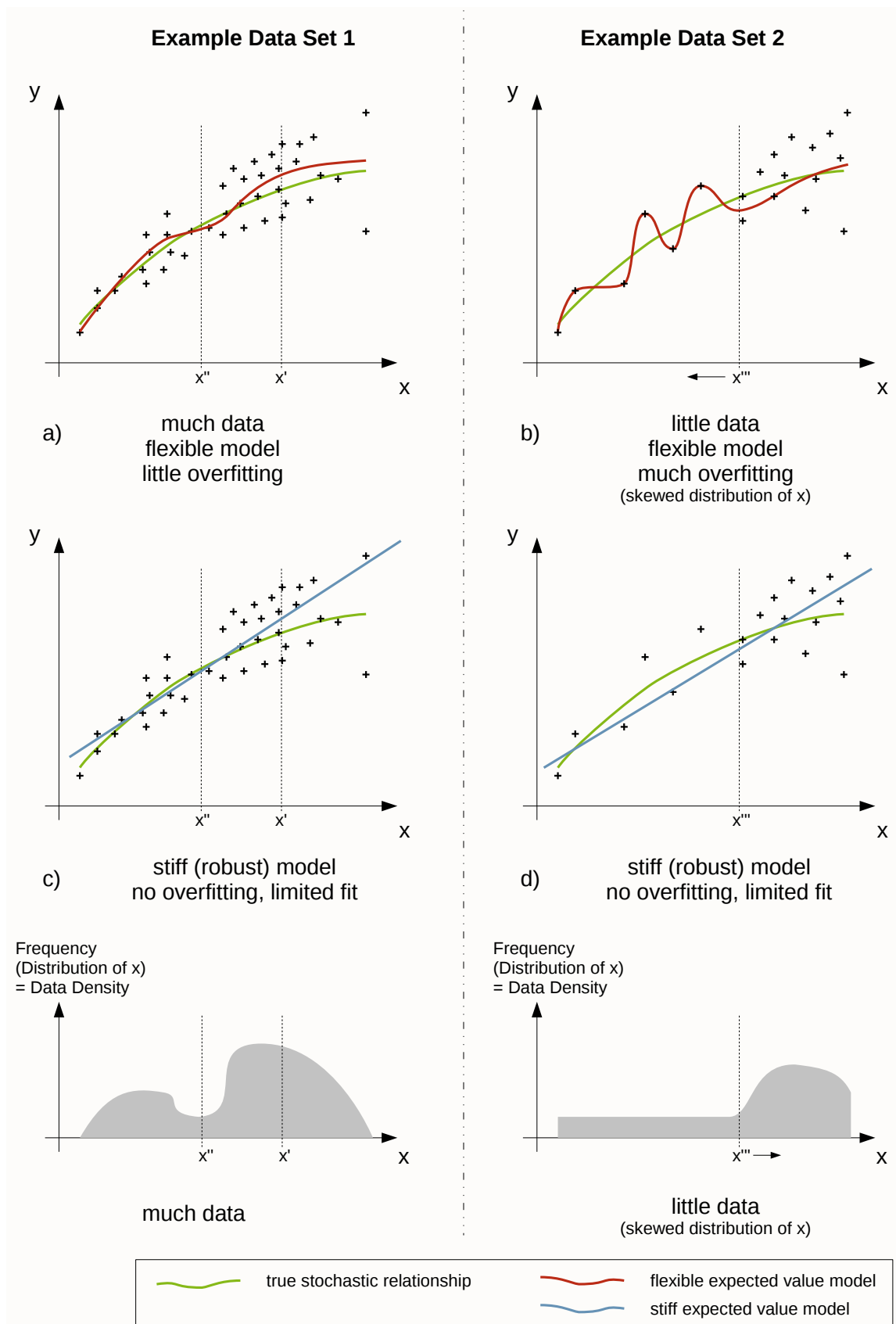


Figure 4.2.: Overfitting Depending on Data Density and Model Flexibility

- **model flexibility/stiffness** – the shape of the model has too many degrees of freedom compared to its dimensionality (i.e., number of variables) – see details below.
- **aggressive optimization** – the algorithm aggressively searches for the mathematically best (i.e., optimal) result, not just any good result – without compromise or a stopping rule to avoid overfitting³⁶.

Besides the absolute number of collected samples, uneven data density over different values of a variable may also lead to overfitting – as figure 4.2 on the preceding page, parts a) and b) illustrate. Note that “Example Dataset 1” in the left column has more data than “Example Dataset 2” in the right column. In addition, dataset 1 is also more evenly distributed, as the graphs at the bottom of the figure show. The data in the right column is the same as in the left column – just with many points removed. The low data densities in part b) for $x < x'''$ lead to strong overfitting for a flexible model (in red). The same model fitting technique performs much more stably in part a), when sufficient data is available and sufficiently uniformly distributed over all possible variable values.

Aside from the data density, the flexibility of the model shape also plays an important role in susceptibility to overfitting. In figure 4.2 on the previous page part a), the red model is rather flexible and already disturbed by a local drop of data density around x' . A stiff linear model (in blue) – shown in part c) – would be less disturbed by such a local drop in data density, at the expense of a lower fit of the true shape of the stochastic process. Thus with increasing stiffness, robustness can be traded for model fit.

Yet linear models can also overfit (i.e., the regression coefficients or slopes become biased) if the sample size is so low that some spaces within the hyperspace spanned by the independent variable are left empty³⁷ or if strong noise requires a high data density throughout the entire hyperspace.

To summarize this section, model fit with “reality” is the aim of all model-building efforts. The model fitting process interacts with the biases in the conventional model fit estimators. Thus, depending on the data and the flexibility of the model, the model may strongly overfit the data – leading to a spurious model that is a poor approximation of “reality” and an overly optimistic model fit estimate.

4.2.3. Challenges in Model Selection

The previous section (4.2.2) highlighted how overly flexible models tend to overfit, driven by aggressive optimization of a biased R^2 model fit measure. The discussion was based

³⁶Almost all optimization algorithms (including the closed-form solution from ordinary regression) have this property, since it would be difficult to determine a stopping point before overfitting occurs.

³⁷This is comparable to using a fractional design instead of a full-factorial design (Devor et al., 1992).

on a model selected a priori. The degree of model freedom is restricted to the flexibility in the model's shape.

If the statistical algorithm also selects the independent variables in the model, then the degree of freedom of the model building process has effectively been increased substantially. Also considering different options for data pre-processing (e.g., filtering out outlier values or variable transformations) adds another degree of freedom to the model building process (Chatfield, 1995, p. 427). Similarly, algorithms, with the option to automatically introduce different transformations of the same variable, have a higher degree of model freedom.

Analogous to the discussion in section 4.2.2 about model flexibility, increasing model freedom while keeping the sample size constant increases the likelihood of overfitting and thus introducing severe biases in the conventional model fit estimators. Since most model selection algorithms also maximize model fit (by adding and removing independent variables), a model selection result on a biased estimator will also be biased. Biased by-variable selection is commonly even less acceptable for interpretation than inaccuracies in the model-shape parameter values³⁸.

Thus model fit estimators with low bias are even more important in algorithmic model selection than in theory-based model fitting – based on an a priori frozen model structure with a frozen set of independent variables and variable transformations.

In addition, Breiman (1996b) has shown by a simulation study that stable model selection also may require very high precision in the data: By the conventional R^2 estimate, the best³⁹ 5-variable sub-model (out of 30 variables in total) has been found (Breiman, 2001b, p. 206). He observed, however, that within the 1% of the best R^2 estimate, there were three other models with very different independent variables.

Thus precision of the model fit estimator, in addition to a low bias, does matter. Both the effect from sampling, i.e., the sampling bias, and measurement inaccuracies may lead to spurious models. Therefore an assessment of the robustness of the solution from a model selection algorithm would be valuable.

4.2.4. Estimates for Predictive Error

Section 4.2.2 on page 123 suggested the independence of the model fitting process as one way to avoid the bias in the model fit estimate caused by overfitting. The previous section (4.2.3) further emphasized why unbiased estimators are a key to robust model selection. Section 4.1.7 on page 114 outlined the difference between model error and predictive error. The conclusion was that for practical applications, where the ideal model is unknown,

³⁸A more detailed discussion of the robustness of by-variable model selection strategies will follow in section 4.1.8 on page 116.

³⁹Since the number of variables in the sub-model is fixed to 5, comparing and maximizing R^2 is equivalent to maximizing more modern estimators, such as AIC or C_p – see section 4.2.4.

predictive error is the model selection and fitting criterion of choice. This section will present various statistical estimators for *predictive error*.

An important property of statistical estimators is convergence towards the true value as the sample size increases (asymptotically unbiased) (Breiman, 1992; Kobayashi and Sakata, 1990; Shao and Wu, 1989). Van der Laan even formally defines the concept of a “*asymptotically linear*” property (van der Laan, 2006), which, aside from no- or low-bias convergence towards the true value, also puts limits on the rate of convergence. This rate is of practical interest because in many applications with many variables the sample-size requirements are high and tough to meet, and thus low-bias behavior of an estimator becomes important for small sample sizes (below the convergence limits) as well. For an example investigation, see Zhang (1993), in which he compares the performance of different cross-validation estimators for small sample sizes. The estimators, which are presented below, will also be evaluated according to their performance for small sample sizes.

Two general estimator design strategies can be observed:

1. **Bias-Correcting Estimators:** A number of estimators (e.g., AIC, BIC⁴⁰ and Mallows’ C_p) are based on the conventional R^2 model fit estimate using the sample data – but the respective authors have found different biases for the limit case (as sample size $n \rightarrow \infty$), and corrected for them. For example, the *Akaike information criterion (AIC)* Akaike (1974) is based on a solid information theoretical foundation (the Kullback-Leibler information quantity), and its estimator is simplified to:

$$\text{AIC} = n \log(\hat{\sigma}^2) + 2K \quad (4.12)$$

where $\hat{\sigma}^2$ is the variance of the model residuals (and thus $\hat{\sigma}^2 \propto R^2$), and K is the number of parameters in the model – serving as bias correction (Anderson and Burnham, 2002, p. 268).

Another example is Mallows’ C_p (Mallows, 1973, p. 662), for which the estimator simplifies to:

$$C_p = \frac{1}{\hat{\sigma}^2} R^2 - n + 2p \quad (4.13)$$

where $\hat{\sigma}^2$ is the estimate of the variance using the current model, n is the sample size and p is the number of variables in the selected sub-model.

Zhang (1992b) even argues that for ordinary linear regression and for the measures AIC, BIC and C_p :

“... all of these can be shown in one way or another to be asymptotically

⁴⁰Burnham and Anderson (2004) show that for large samples, the *Bayesian Information Criterion (BIC)* converges to the same results as AIC.

equivalent to minimizing (with respect to k) [the final prediction error FPE:]”

$$C(k, \lambda) = R^2(k) + \lambda k, \quad 0 \leq k \leq K \quad (4.14)$$

where $R^2(k)$ is the residual sum of squares based on the sample data and for the model with only the first k covariates (out of K covariates in total). In the limit and for ordinary linear regression, λ , the penalty for overfitting, is the only difference between AIC, BIC and C_p .

2. **Bias-Reducing Cross-Validation Estimators:** Other estimators avoid using the conventional R^2 estimate and instead use estimation techniques that are more independent of the model fitting process and thus do not suffer from the model fitting bias (as detailed in section 4.2.2 on page 123). These estimation algorithms are commonly referred to as *cross-validation* techniques, or in Breiman’s terms as *resampling* techniques (Breiman and Spector, 1992).

While theoretically sound for the limit case with large samples, the bias-correcting estimator group has received much criticism in its application for model selection (Breiman and Spector, 1992; Zhang, 1992b). Since model selection is commonly an issue when many potentially relevant variables are involved, and thus the sample-size requirements are also hard to meet, these estimators’ weak performance for small sample sizes causes instable model selection behavior. Yuan and Yang (2005) even use a bootstrapping⁴¹ technique to assess the instability of AIC or BIC due to the limited size of the sample. When the instability, measured by the index PIE, is too large ($\text{PIE} > 0.5$), they suggest specifying the model by other means, e.g., by combining the predictions of multiple models into an average and thus more robust prediction (also called *bagging* or *ensemble building*).

As an alternative to correcting and controlling the bias in the conventional R^2 model fit estimate, *resampling* or *cross-validation* techniques have been suggested by various authors (Breiman, 1996b; Burman, 1989; Efron, 1986; Zhang, 1993). A particularly data-efficient resampling technique is *bootstrapping* – further details are presented in text box 4.2.1.

The term *cross-validation* is apparently used somewhat loosely in the literature. The actual estimation algorithms stretch from simple dataset separation (see below) to sophisticated and computationally intensive bootstrapping methods. All estimation algorithms have in common that they calculate the model fit using the usual R^2 or R_{abs} formulas (equations (4.9) and (4.10)), though based on different datasets – either a sub-set of the collected sample or a completely separate test (validation) dataset. By using independent data sets, these estimates do not suffer from the bias introduced by the interaction of

⁴¹Bootstrapping is a particular cross-validation or resampling technique – see also text box 4.2.1 on the next page.

Text Box 4.2.1 Bootstrapping

Cross-validation techniques commonly split the sample data into two or more sub-datasets, of which some are used exclusively for model building and others are used for validation. In contrast, *bootstrapping* (Breiman, 1992; Breiman and Spector, 1992; Efron and Tibshirani, 1997) in several iterations fits multiple models to different sub-samples of the sample data in order to assess the robustness of the model-building process with respect to the sampling bias. Each model-building iteration commonly begins with splitting the sample data into two sub-samples: one for model building (training data) and one for validation (test data). Then a model is fitted against the training data and the model quality (e.g., model fit R^2) is assessed against both training and test data. After running through multiple iterations, a set of models is created and fitted. The size of the variations in model fit and in the parameters across the iterations provide a good measure for assessing model (and parameter) robustness.

model fit estimation and model fitting (as described in section 4.2.2 on page 123).

The following list gives an overview of the most common cross-validation algorithms:

- **Simple Dataset Separation Cross-Validation** – The collected data is split into a training dataset and a test dataset. The training dataset is used for model fitting, while the test dataset is used for calculating a model fit independent estimate of R^2 or R_{abs} (Stone, 1974, p. 111). While this method is simple and delivers a truly independent estimate, the available data is split into two smaller samples, leading to a higher sampling bias for the model fit estimates on both the test and training data. Thus this technique is useful and simple if enough data is available for splitting. Since cross-validation becomes especially important when relatively little data is available, other more data-efficient techniques have been developed.
- **Jackknife Cross-Validation** – Shao and Wu (1989) developed Jackknife cross-validation as a very data-efficient estimation method:
 1. One or more data points are removed from the collected sample;
 2. a new model is fitted to this new data sub-set; and
 3. a model fit estimate is calculated based on the data sub-set and the new model.
 4. Steps 1-3 are repeated many times, resulting in a distribution of R^2 or R_{abs} estimates.

With the distribution of R^2 or R_{abs} estimates, the variability of the model fit estimate can be assessed. In some variants of jackknife, square correlations between the complete data and the subset are used to estimate a new, less-biased model fit. While this estimation algorithm is very data efficient (almost the complete dataset is used for model fitting), the models are fitted on different but very similar samples, and hence the estimate is not very independent of the model fitting process.

- **Breiman’s Little Bootstrap** A bootstrapping-based estimator for the predictive error in the X-fixed case, i.e., the experiment is repeated in a controlled manner with the same dependent variable settings X. Simulation experiments in [Breiman \(1992\)](#) show that the estimator has very low bias and is very data efficient – i.e., does not require a large and separate dataset.
- **Efron’s Bootstrap Cross-Validation** Efron suggests a bootstrapping-based cross-validation estimator that has a low level of bias that is comparable to the simple data separation cross-validation, but is more data efficient – i.e., has lower variability than simple cross-validation for a given sample size ([Efron and Tibshirani, 1997](#)).
- **Breiman’s Predictive Error Cross-Validation Estimate** The estimator works only with *bagged* or *ensemble methods*, such as the *random forest* algorithm by [Breiman \(2001a\)](#), in which a number of individual models are averaged to give a more robust prediction. The estimator works as follows:
 1. From a part of the collected sample, a training sample is generated by bootstrapping⁴². All samples not used in the training dataset make up the test dataset. Many such training and test dataset pairs are generated.
 2. A new “individual” model is fitted to each training and test dataset.
 3. A model fit estimate for each test dataset and the respective individual model is calculated.

The result is a distribution of what Breiman calls “*out-of-bag*” model fit R^2 estimates. The average of these out-of-bag estimates is an estimator for the predictive error. For details, see: [Breiman \(1996b\)](#); [Breiman and Spector \(1992\)](#).

[Zhang \(1993\)](#) confirms that Breiman’s multifold cross-validation estimate outperforms any other cross-validation algorithm in terms of bias and data efficiency – albeit at high computational cost.

As the development of further cross-validation techniques is still in progress, this list is not complete. [Zhang \(1993\)](#) and [Burman \(1989\)](#) list and compare additional variations of cross-validation.

The more sophisticated bootstrapping-based cross-validation techniques have an additional benefit: the estimator is the mean (or another aggregation) of a number of results for different bootstrapping samples (and models). Thus if the aggregation step is skipped, one gets information about the distribution of the individual results that make up the aggregated estimator. If the mean is used for aggregation, and the number of individual

⁴²One option is to use bootstrapping with replacement, using about 63.2% of the data in the training sample ([Strobl et al., 2007](#), p. 4). Then the training dataset has the same size as the original sample, and the test dataset has about 36.8% of the original sample size. Alternatively, bootstrapping without replacement but with a limited training data size (e.g., 70%) can be used.

results aggregated with the mean is sufficiently large, the central limit theorem⁴³ applies to the mean – i.e., the mean estimator will be normally distributed, and thus one can calculate confidence intervals for it. Hence the information about the distribution of the individual results allows one to assess the robustness of the estimator towards the sampling effect. It furthermore helps to reduce and control the high variability that these low-bias estimators often exhibit (Efron and Tibshirani, 1997).

In addition, confidence intervals around the estimators for two different models allow the robust judgment of whether one model is superior to the other or whether the predictive error estimates for the two models are too close and too inaccurate to distinguish. Such judgments furthermore allow an assessment regarding how stable the ranking of candidate models is towards biases in the measurements.

In summary, during recent decades a number of estimators with much less bias from model fitting became available as an alternative to the conventional R^2 or R_{abs} model fit estimates. Two classes of estimators emerged: estimators that aim to correct for the bias in the conventional R^2 model fit estimate and estimators that use independent datasets for cross-validation. Recent simulation studies (Breiman, 1992; Efron, 1986; Zhang, 1993) have shown that cross-validation approaches show better performance, especially for smaller samples, when the bias-correcting estimators are not yet close to convergence. Additionally, the bootstrapping-based cross-validation measures offer distributions of individual model fit estimates. These distributions allow an assessment of the variability of the estimated value in a particular case, e.g., by means of confidence intervals. With this information they allow an assessment regarding robustness towards the sampling effect and measurement errors.

Breiman’s multifold (bootstrapped) cross-validation estimate for the prediction error (Breiman, 1996b) outperforms any of the other methods and provides results with minimal bias – even for smaller sample sizes. Therefore it is used in the BOGER algorithm developed for this study – as described in detail in section 6.2.7 on page 195.

4.2.5. Examples of Robust Algorithms and their Properties

The conclusion from the theoretical part of this discussion was that model selection based on the sample data is possible in principle (section 4.1.6 on page 111) and that predictive error is a suitable criterion for both model selection and parameter fitting (section 4.1.7 on page 114). Next, various practical challenges with model fit estimators were discussed (section 4.2.2 on page 123) and more robust alternative cross-validation estimators were

⁴³The *central limit theorem*, also known as the “*law of large sums*”, states that the distribution of a large sum of independent and identically distributed random variables (the summands) is distributed normally with a mean equal to the expected value of the individual summand random variables and their variance divided by the number of summands ($\sigma_{\text{sum}}^2 = \sigma_{\text{summands}}^2/n$) (Lee, 1997). Thus the mean of the sum is a more accurate estimator for the true mean of the process that generated the individual summands.

presented (section 4.2.4 on page 129). With these robust estimators, the practical model selection challenges, described in section 4.2.3 on page 128, can be overcome. After these mostly theoretical arguments, this section provides examples of robust model selection algorithms and highlights their common properties.

The following algorithms are referred to as “*robust*”, which means more specifically that they are more robust than others, when run with the same sample size, or, conversely, that these algorithms need less data to attain the same level of robustness.

Examples of robust model selection algorithms are:

- **Zhang’s Cross-Validation Model Selection Method** – Zhang (1993) uses different cross-validation estimators for the predictive error in order to rank and then manually select from a small set of regression models.
- **Yuan and Yang’s ARMS Algorithm** – Yuan and Yang (2005) propose an algorithm with a model-screening feature, i.e., first a large number of models are screened and ranked by the simple and biased AIC and BIC criteria. Only a certain fraction of good models (e.g., the top 25%) are used in the next stage. The final model is a weighted average (a *bagged* model), with the weights based on a more elaborate cross-validation estimate of the relative predictive error. Similar to a bootstrapping approach, the original sample data is split into many pairs of equally sized training and test datasets. Next, the internal regression algorithm is used to fit individual models to each training dataset. Based on the respective test dataset, a predictive error estimate for each individual model is calculated. Aside from the usage of a robust predictive error estimator, much of the algorithm’s robustness is derived from the combination of many models – commonly referred to as *ensemble building* or *bagging*⁴⁴.
- **Breiman’s Random Forest Algorithm** – Breiman (2001a) proposes a non-parametric ensemble-building algorithm based on binary decision trees, which can be used for classification but also regression with continuous outcome variables. In the same spirit as Yuan and Yang’s split data method, a number of artificial training and test dataset pairs are generated by randomly drawing data rows⁴⁵ from the original sample (bootstrapping). While the datasets are all random and independent of each other, as an effect of bootstrapping, they share the same distribution with the original sample. CART decision trees are fitted to the training part of the bootstrapped datasets. Since each node (or split) of the tree is built only on a

⁴⁴In the statistical literature, there is a whole string of discussion regarding the class of ensemble-building algorithms, e.g., arcing (Breiman, 1998) or boosting (Lutz and Buhlmann, 2006).

⁴⁵The term “data rows” here refers to, e.g., the survey data of a single participant or the data for a particular event – including the outcome variable. Thus the bootstrapping procedure never mixes data (e.g., between survey participants) and thus ensures that the new random sample has the same multi-variate distribution as the original sample.

single variable, this algorithm does not need a separate variable-screening or feature-selection method to deal with large numbers of variables compared to the sample size (Strobl et al., 2007, p. 3). Finally the individual trees are combined (*bagged*) to a single random forest model by simple averaging of the individual tree predictions. Breiman (1996a) calls this ensemble-building procedure *bagging* (*bootstrap aggregating*). The algorithm includes but does not actively use Breiman’s estimator for predictive error (see section 4.2.4 on page 129), which in the context of random forest, Breiman calls *out-of-bag estimate*.

According to Breiman’s simulation studies (Breiman, 2001a,b), the random forest algorithm performs well even with relatively few samples compared to the number of variables. Breiman shows mathematically that in the limit case (for large samples) the algorithm does not overfit and generalizes well beyond the sample as the number of individual trees grows large.

An open-source implementation for \mathbb{R} is available as `randomForest` package (Liaw and Wiener, 2002)⁴⁶

- **Strobl’s cForest Algorithm** – Strobl et al. (2007) have improved Breiman’s Random Forest algorithm by using an improved version of CART trees. The split at each node is not just determined by minimizing the misclassification rate (the analog of the conventional R^2 model fit estimate in decision trees), but by a significance test. The user may choose from simple tests, such as the student t-test, to more sophisticated procedures, such as Monte Carlo – see also the \mathbb{R} package `party` (Hothorn et al., 2008). Thus Strobl’s algorithm is even more conservative in preventing overfitting than Breiman’s original random forest algorithm.

In summary, this list is far from complete and instead illustrates that robust model-selection algorithms are feasible and have been implemented.

Given the special properties of the survey dataset collected for this study (in terms of number of variables, variable types and level of noise – see section 5.12 on page 163), the BOGER algorithm – as a more suitable alternative for this dataset – has been developed for this study based on the design principles underlying the algorithms presented above (see section 6.2 on page 179).

⁴⁶As with any other add-on package for \mathbb{R} , it is available on CRAN either via the web or through the `install.packages()` function.

5. Quantitative Stage - The Survey Instrument

Chapter Contents

5.1. Overview	138
5.2. Pre-Survey Qualitative Pilot Interviews	139
5.3. Covered Constructs in the Survey	141
5.4. Quantifying On-The-Job Learning – Learning Index	145
5.4.1. Learning Index Survey Tool	146
5.4.2. Learning Index Definition	150
5.5. Survey Pilots	152
5.6. Survey Design Goals	153
5.7. Survey Algorithm	153
5.8. Survey Conduction and Resulting Sample	158
5.9. Actual Performance of the Interactive Survey	159
5.10. Data Pre-Processing	161
5.11. Validity Investigation of the Learning Index	162
5.12. Properties of the Data Set	163
5.12.1. Multi-Variate Relationships / Collinearity / Correlations	163
5.12.2. Noise	166
5.12.3. Non-Linear Relationships	168

5.1. Overview

The initial literature research (chapter 2 on page 21) and the insights from the qualitative stage (chapter 5.2 on the next page) lead to a large set of organizational and personal factors that potentially affect on-the-job learning. Since the aim of this research is to find the *most important* factors affecting on-the-job learning (section 2.7.2 on page 80), a quantitative study became necessary – as discussed earlier in the methods section 3.2 on page 97.

Given the large number of variables, and with it a requirement for a sufficiently large sample size, a strongly structured online questionnaire was chosen as a quantitative research tool, as will be described in this chapter.

The 60-minute online survey was fielded in the summer of 2007 at the shipyard *Meyer Werft* in Papenburg, Germany, and yielded 329 samples for further statistical analysis (in chapter 7 on page 205).

Meyer Werft specializes in building cruise ships and other high-value vessels. The yard builds highly customized ships in small series of two to six, using a flow-line principle from smaller assemblies (sections and blocks). Since every manufactured piece is different, Meyer Werft is no typical mass-production yard¹ The survey covered all departments, including a wide range of diverse tasks and working environments, such as automatized steel cutting and block assembly, purchasing and technical design, IT and administration.

The following sections illustrate:

- insights from a few pre-survey qualitative pilot interviews (section 5.2 on the facing page),
- the content of the survey (section 5.3 on page 141),
- the design and validation of a quantitative measure for on-the-job learning activity (the learning index),
- the design of the interactive survey mechanism² that led the participants through the survey to allow for a wide spectrum of questions within a comparatively limited time frame (sections 5.4 on page 145 and 5.11 on page 162),
- a description of how the survey was piloted and fielded (sections 5.5 on page 152 until 5.9 on page 159),
- the data pre-processing and filtering steps used (section 5.10 on page 161),
- and the properties of the resulting data (section 5.12 on page 163).

¹For a company profile of Meyer Werft, see appendix section A.3 on page 289.

²Section 5.9 on page 159 includes a quantitative assessment of the performance of this survey algorithm.

5.2. Pre-Survey Qualitative Pilot Interviews

From a first literature research iteration, the following factors affecting on-the-job learning came in focus for this study:

- **Biographical factors** (e.g. age, education, years in the firm)
- **The Nature of the Work Environment** (e.g. leadership)
- **Properties of the Example Task** (e.g. interdependence of this task with tasks of other people, autonomy in working on the task)
- **Learning Behavior and Learning Success** (e.g. learning used strategies)
- **Personality** (e.g. self-efficacy, Big-5)
- **Personal Network** (e.g. number of personal contacts)
- **General Climate** Working, Learning and Age Climate

Given the goal and the method choice of this study (see section 3.2 on page 97), these factors needed to be quantified in a time economical and standardized manner – as required for data acquisition with a fully structured survey.

Hence, the first literature search iteration also included a search for suitable standardized and tested surveying tools. As will be detailed in section 5.3 on page 141 some standardized surveying tools were found to be suitable for the task and context of this study. Yet for a number of important aspects of the problem no suitable existing surveying tools could be found. This most prominent example is the outcome variable of this study: **on-the-job learning effect** for dissimilar tasks and dissimilar learning episodes (further details in section 5.4.1 on page 146).

Furthermore, many standardized surveying constructs used in literature have been developed, tested and applied with groups of participants that do not represent the population average (e.g. university students). Therefore it was necessary to interactively test existing surveying tools in interviews, whether the questions were understood in the intended way by the participants.

Thus with the aim to test existing and if necessary develop new surveying tools for this study, a series of 6 qualitative interviews was conducted. While semi-structure interview guides were used for the interviews, the interview guides were evolving with each interview and thus no standardized data, that is comparable across participants, was collected³.

³Analysis methods for unstructured interviews, such as transcript coding methods, were thus not used either, since the aim was not to generate scientific evidence but rather to test and refine the line of questioning. Hence there is also no extensive documentation of the interviews at this point.

5.2. Pre-Survey Qualitative Pilot Interviews

One of the interviews was lead by Polina Isichenko with a stronger focus on innovation, which was the focus of her research.

Using the method of evolving interviews the following insights – regarding practical and suitable surveying methods – emerged:

- It is easy to talk about knowledge in general terms but very **hard to** ask people **to list what knowledge they have** in different thematic areas. Hence it is difficult to have the participants categorize their knowledge in abstract functional categories such as 'embedded knowledge', 'procedural knowledge', 'event knowledge' as suggested by [von Krogh and Venzin \(1995\)](#), which is in-line with the arguments from the theory section [2.5.2 on page 73](#). This challenge becomes practically infeasible, when additional restrictions from the research method become relevant: For example, if a structured survey is chosen in order to obtain a sizable sample – as proposed later in section [5.3 on the next page](#) – the survey instrument allows only for very limited (and before-hand scripted) interaction in addition to the time limitations of a survey.
- **Using concrete examples** from the interviewees' work, proved as very effective facilitator to get more details on the interviewees use of their knowledge. This matches with findings in the psychological literature, as for example [Schwarz and Bienias \(1990\)](#) have observed in their study comparing general and episode specific questionnaire results.

These two insights were the trigger to link the question items for the learning index, which is used in the survey of this study as the outcome variable, to specific and concrete learning episodes – as experienced by the individual participant (details in section [5.4.1 on page 146](#)).

Eventually the insights gained in this evolutionary interviewing process lead to the design of the fully structured survey (described in [5.3 on the next page](#)) and further testing of the fully structured survey in interview form followed before fielding it (as described in the survey pilot section [5.5 on page 152](#)).

As mentioned before, the before mentioned practical insights regarding the surveying methods, were not derived in a particularly systematic (i.e. scientific) manner. Yet the insights were confirmed in the survey pilots (section [5.5 on page 152](#)) and the successful application of the fully structured survey (section [5.9 on page 159](#)).

Furthermore and in the spirit of the iterative research approach (from section [3.1.6 on page 93](#)), the qualitative interviews inspired further literature search iterations during the weeks the interviews were conducted.

5.3. Covered Constructs in the Survey

The aim of this survey was not only to provide data for this study, but also to provide data for an innovation study by Polina Isichenko and a work motive study by Christian Roßnagel. Furthermore, the consulting company DNV⁴ also had a few general questions on learning and knowledge as a rerun from an earlier survey.

Hence the selection of constructs⁵ of contained in the survey is based on the initial literature review for this study but also contains constructs that serve the aims of the other parties. Aside from a common general part, the questions for the different research aims are triggered in a partly random fashion by an algorithm described in section 5.7. The survey is usually not run with the complete set of question items listed below, and therefore all participating researchers got a large data set – albeit with a substantial fraction of missing values that will require filtering and imputation (see appendix section A.5.2 on page 303).

The following table lists all included (but algorithmically activated) constructs and question items. All questions items were originally posed to the participants in German; they have been translated here into English. Unless otherwise noted, the questions were created by the author of this study.

Table 5.1.: Survey Constructs

<i>Construct Purpose</i>	<i>Official Name/Source / Comment</i>	<i>No. of Factors</i>
► Biographical Information		8 (Section Total)
Sex		1
Age	In number of years not age groups	1
Department	By top-level department (cost account) number	1
Education Level	Highest completed degree and current enrollment in practical training (“Lehre”)	2
Years – in the company & in the current dept.	In number of years	2
Years experience in other companies	In number of years	1

⁴Det Norske Veritas <http://www.dnv.com>

⁵The *constructs* mentioned here are batteries of question items that all probe in a similar direction. The combined (usually summed or averaged) results of the question items provide a single reliable measurement in a general / broad direction.

Table 5.1.: Survey Constructs

<i>Construct Purpose</i>	<i>Official Name/Source / Comment</i>	<i>No. of Factors</i>
► DNV Questions		25 (Section Total)
Knowledge Distribution & Access	DNV's custom questions, already used in only one previous survey at Meyer Werft	12
Personal Motivation	DNV's custom questions	13
► Work Environment – non-task specific		9 (Section Total)
Leadership (perceived behavior of the superior) <ul style="list-style-type: none"> • Encouragement for employee initiative • Clarity of responsibilities • Constructive feedback • Result focus • Group climate focus • Trust in the employees 	Adapted from van de Ven et al. (2000)	6
Formal Continuing Education	Seminar Offerings and Participation	2
Speed of change in the work environment		1
► Innovation / Project / Work – Task Properties		33 (Section Total)
<i>Depending on the chosen branch: an innovation project, a non-innovation project, the normal work in the last 4 weeks.</i>		
▷ Innovation Participation		7 (15 in detail) (Subsection Total)
<i>Depending on the survey path, a short or a detailed set of innovation questions is included.</i>		
Nature of participation in innovation projects	Idea Creator, Supporter, ...	5 (7 in detail)
Involvement in the Project	Time involvement	2 (5 in detail)
Innovation Novelty	Incremental / radical	(1 in detail)

Table 5.1.: Survey Constructs

<i>Construct Purpose</i>	<i>Official Name/Source / Comment</i>	<i>No. of Factors</i>
Innovation Impulse	New requirement, external idea, competitive pressure, ...	(1 in detail)
▷ Project Properties		2 (Subsection Total)
Project Duration		1
Work (Time) Input		1
▷ Work Process Classification – for the example project / work of last 4 weeks		31 (Subsection Total)
Task Effectiveness	perceived results / performance (self-rating)	2
Task difficulty level		1
Change in Working Processes	Changes in the department within the last 3 years.	2
Job/Task Characteristics <ul style="list-style-type: none"> • Skill Variety (AV) • Task Identity / Closeness of Task Definition (AG) • Task Significance (BSK) • Autonomy (Auto) • Feedback (FB) 	Job Diagnostic Survey (JDS) a Job Characteristics Model Survey Tool (see Oldham & Hackmann 1975 in Kulik et al., 1988), translated into German by Schmidt and Kleinbeck	5
Comparison with the project and the participant's regular work	For projects only	1
Task Interdependence	Question on whether this task depends on other people's works or vice versa	2
Procedural detail	On detail level of work instructions	1
Level of systematization for solving problems		1
Resource scarcity	Innovations only	1

Table 5.1.: Survey Constructs

<i>Construct Purpose</i>	<i>Official Name/Source / Comment</i>	<i>No. of Factors</i>
Level of personal communication	Number of contacts by contact group (own dept., supplier, superior, ...) + frequency of communication	14
Level of Routine	Frequency of similar tasks in the past	1
► Learning Component		20 (Section Total)
“Learning Frequency”	Frequency of consciously perceived (and recalled) learning events	1
Learning Strategies	High-level individual learning strategy (by experimentation, reading, discussion, investigations of past incidents, ...)	6
Perceived Learning Barriers & Opportunities	Mostly organizational barriers: lack of information, contact persons, time, ...	9
General learning intensity impression	“How much in total and compared with other tasks did you learn in project/task XYZ?” (see appendix section A.4.3 on page 297)	1
Individual novelty of the subject	Level of surprise about the learning events	1
Personal interest in the topic	“The topic was also interesting to me personally and independent of my tasks.”	1
Feedback intensity (from colleagues)		1
► Person specific variables		14 (Section Total)
Task-specific self-efficacy	Loose adaption from Ralf Schwarzer’s scale for teacher job-specific self-efficacy (Schwarzer and Jerusalem, 1999)	1
Self-Regulation	Self-regulation scale by Schwarzer (2000) [Original in German]	1

Table 5.1.: Survey Constructs

<i>Construct Purpose</i>	<i>Official Name/Source / Comment</i>	<i>No. of Factors</i>
Epistemological Beliefs (EÜ) <ul style="list-style-type: none"> • Objectivity of Facts • Subjective learning aims 	Based on Bauer, Johannes, Festner, Dagmar, Harteis, Roßnagel.	5
Job Self-Efficacy	“Berufliche Selbstwirksamkeit (BSW)” by Abele et al. (2000)	1
Job Involvement	Survey instrument from Frone and Russell (1995), translated into German by author	1
NEO FFI Big Five	10-item short version (NEO FFI) of the Big-Five personality dimensions Rammstedt and John (2007)	5
► Work Motives		212 (Section Total)
An instrument by and for Christian Roßnagel		212 (short items)
+ free text field for comments		
Overall Total	excl. short Work Motive Items	157

All questions that were included in the final model, as a basis of the interpretation of results, are quoted (in an English full-length translation) later in the discussion of statistical results in chapter 7 on page 205.

5.4. Quantifying On-The-Job Learning – Learning Index

The aim of this survey instrument is to quantify the on-the-job learning effect. For that purpose, I developed and tested (section 5.11 on page 162) a novel survey instrument that makes it possible to measure the on-the-job learning effect: the *learning index*. Aside from the employed learning strategies, it is the core outcome variable. As a back-up and cross-checking instrument, a much simpler question on the general learning impression is included in the survey (see appendix section A.4.3 on page 297).

A new learning construct was designed for this study because an existing standard construct – with acceptable quality and applicability for this research effort – was not found in the literature.

5.4.1. Learning Index Survey Tool

A challenge with quantifying on-the-job learning is that the learning content is not standardized. Hence there is no standardized exam that allows one to measure the learning effect of the survey participants⁶. Non-standard learning situations might therefore call for more qualitative methods, with on-the-job observation and in-depth interviews of individual participants by a team of researchers. But given the large number of potentially relevant factors and thus the high required sample size (> 200), such an individual qualitative research approach with a case-by-case analysis is not feasible for this study. Therefore a perceived, i.e., self-reported, learning effect is quantified by the learning index survey tool in multiple steps – as presented below.

Self-reports have a number of challenges – especially when they involve a recollection of past events (Loftus, 2003). Therefore this survey tool is designed to reduce the biases associated with recollection:

Following the arguments from theory section 2.3.6 on page 47 on iterative learning, the learning index survey tool is designed to directly address the episodic nature of learning: learning occurs during episodes of experience and remains connected to these episodes (D'Eredita and Barreto, 2006b; Racsmany and Conway, 2006). Hence it is not surprising that Schwarz and Bienias (1990) found that relating survey questions to actual episodes of experience reduces bias in the recollection of facts (see also Ji et al. (2000)).

That is why this survey tool relates questions on past learning experiences to actual work episodes of the individual survey participant in two ways: 1.) The questions about learning activity are linked to an actual task from the participant's working experience from the recent past. To further facilitate the recollection of the task's context, the participant is asked to divide the task into 2 – 3 work steps. 2.) For each of these work steps, the participant is asked to *name* a particular learning situation rather than simply providing an unspecific learning intensity.

This survey strategy requires an interactive (i.e., computer automated) survey system, for which the web-based software system Unipark was chosen (details are provided in section 5.7 on page 153). This software system can steer the participant through different paths within the survey based on pre-programmed rules and the individual participant's

⁶In work settings, measuring the learning effect is substantially more difficult than in school or university settings, since it is generally much less clear what the participants already knew about the problem and its solution before they encountered it. Hence when using a standardized exam, the researcher would need to assess the participant's knowledge before and after the learning episode – without giving hints towards the problem or its solution with the exams.

answers. For the learning index, the loop feature of the online survey system was used, which generates question pages based on templates and previous answers on-the-fly.

This process sends the participant through two cascaded loops. Figure 5.1 illustrates it in further detail. This process is a central component of a larger survey algorithm, which is described later in section 5.7 on page 153.

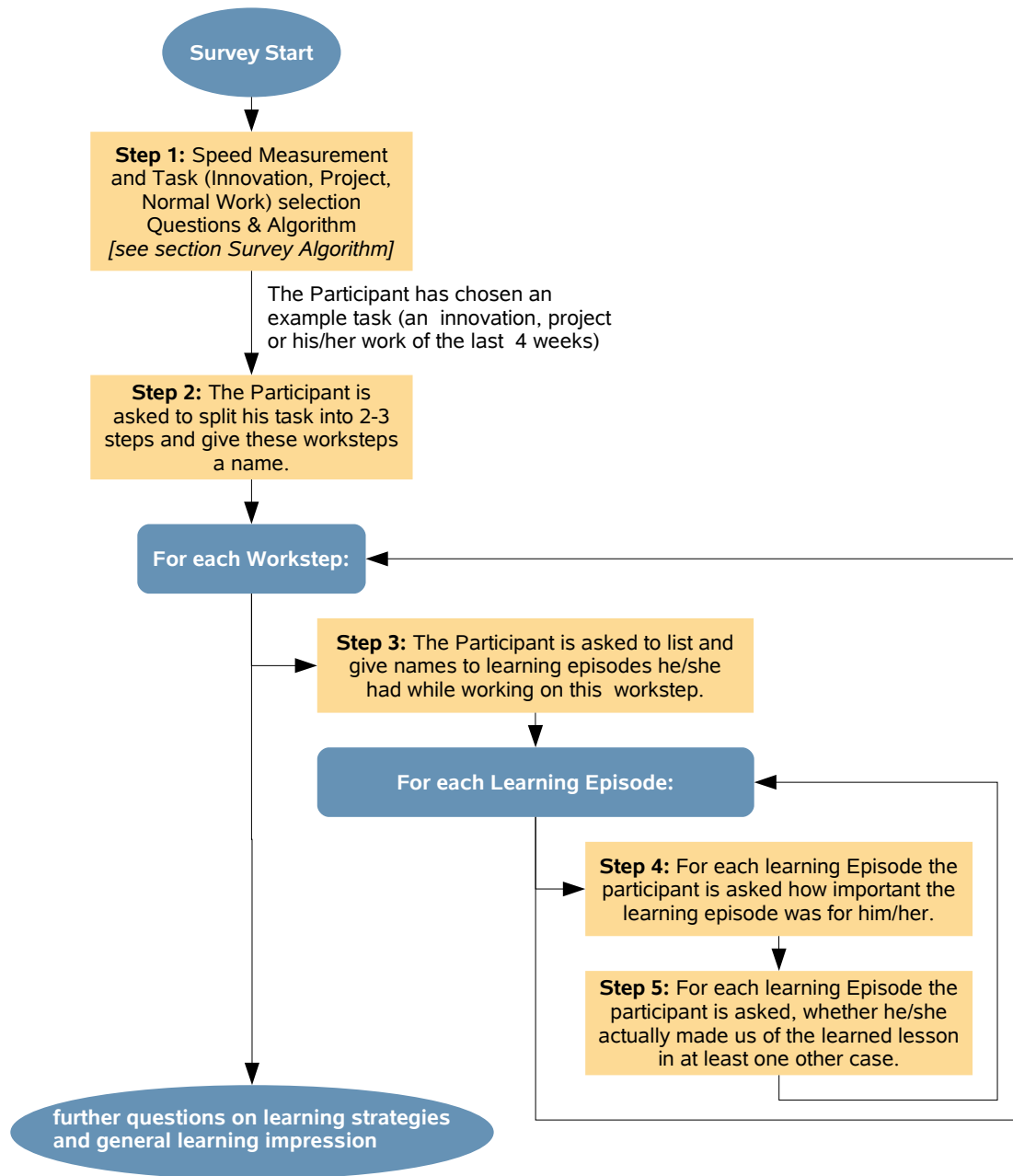


Figure 5.1.: Learning Index Surveying Tool – Flow Chart (Source: Author)

1. In **step 1**, the participant is asked to choose a project he or she has recently participated in: a sample innovation, a non-innovation project, or his or her work in the last four weeks. In order to prevent participants from changing their example project during the rest of the survey, and to encourage participants to think seriously about this step, the participant is asked to enter a name for the example project in a text field⁷.

Note that the instructions ask for *any* project – explicitly including projects and tasks without exceptionally high learning effects. In the survey introduction, however, it was communicated to the participants that this survey is about on-the-job learning. Hence the participants will more likely⁸ select learning-rich rather than non-learning projects and tasks.

2. For the chosen example project, the participant is asked in **step 2** to split his/her innovation or normal project into two or three different work steps; in the case of the last four weeks of normal work, the participant is asked to split his/her work into two or three tasks. Again, the participant must enter names⁹ for the task steps.

All following questions regarding the example project, the work step or the short task, include the wording entered by the participant, which increases the context-setting effect and speeds up the answering of related questions.

3. In **step 3**, the participant is asked to name any learning situations (i.e., learning episodes) that he or she can consciously recall within each work step. As with the projects and work steps, the participant is asked to enter a name for the learning situation – encouraging the participant to think of actual learning episodes instead of just entering a rough number.

The number of learning episodes is limited to three, with the option to enter a number for additional learning episodes if more than three occurred. Furthermore, there is an explicit check-box for the case that somebody did not learn anything during this work step¹⁰.

4. For each of these learning situations, the participant is asked to self-assess the learning effect with a question on the lesson's importance (**step 4**) and another question on the lesson's usefulness (**step 5**) in another actually experienced occasion. The

⁷No name is required for the work of the last four weeks.

⁸This self-selection bias for learning-rich projects and tasks is not a problem because the research questions of this study can also be answered with a slightly but systematically biased learning index – as was discussed in section 5.4.2 on page 150.

⁹To maintain anonymity, an instruction offers the participants the option to use acronyms instead of recognizable names for the work steps or the example project.

¹⁰This check box is used as a data consistency check – triggering a warning message to the participant if violated.

aim is to get a second assessment of the lesson's importance and to reduce bias in this assessment by linking it to another concrete situation (context) where the lesson became useful. For the full question texts and answering scale description, see the text box [5.4.1 on the next page](#).

In the qualitative pre-pilot interviews (section [5.2 on page 139](#)), it became clear that the two questions are understood by the participants in very similar yet different (i.e., partly interchangeable) ways. This observation is further supported by a mild Pearson correlation of 0.36 in the survey data (for further details and a graphical representation, see appendix section [A.4.1 on page 295](#)).

Therefore the two question items were merged into a single construct by addition – in order to gauge the lesson's value. In four cases (1.3% of the survey data), the survey reduction skipped the question on learning usefulness to save time (see section [5.7 on page 153](#)). In those few cases, the lesson value was estimated from the learning importance question item alone – as a proxy for both. Using learning importance as a proxy for both questions is only an approximate replacement for the result from both question items and thus may be challenged. Yet in the case of this survey, the proxy method came to application in such a small fraction of all cases (1.3%) that the effect of using this approximation method is in any case negligible.

Learning usefulness, when measured as an expectation *before* the learning task is engaged, may also be related to learning motivation, as predicted in the 'expectancy theory' by [Vroom \(1964\)](#). This survey, however, collected the learning usefulness *after* the learning episode – as experienced in an actual instance – and therefore the data is not suitable to test the expectancy theory. Yet it might be valuable to test the expectancy theory in the context of on-the-job learning in future research – as detailed in section [8.5 on page 282](#) on future research.

After the participant completes all iterations of the double-loop, the bulk of other general questions follows. Hence this difficult part of the survey is located close to the beginning, when attention is still fresh. As a positive surprise the results in section [5.9 on page 159](#) indicate that the attention of the majority of participants stays high during all iterations of the double-loop. There is no detectable degradation of the results by a loss of attention – as might have been expected.

The result of these loop iterations are a collection of learning situations that are linked to a specific work experience¹¹, which the individual participant chooses and names, as well as assessments of the importance and usefulness of the lesson learned for each learning situation.

¹¹This experience can be either an innovation project, a longer normal (non-innovation) project or specific work tasks from the previous four weeks, which the participant has to name.

Text Box 5.4.1 Learning Importance and Usefulness Questions in Full Text

Learning Importance is assessed for each learning situation by the question:

“How important for you was what you learned?”

“Wie wichtig war das Erlernte für Sie?” [German Original]

A five-step scale from “not at all important” to “very important” over was used without further intermediate anchoring. In order to get a true interval scale, this scale is only anchored at the ends and the intervals in between are not labeled. The web-based survey tool visually suggests equal distance between the steps by the equal spacing of the radio buttons.

Learning Usefulness is assessed for each learning situation by the question:

“Was what you learned after <Project Name> <Workstep Name> also useful for you in at least one specific case?”

“War das Erlernte für Sie auch nach <Projekt Name> <Name des Arbeitsschrittes> in mindestens einem konkreten weiteren Fall für Sie nützlich?” [German Original]

The participant rates the question on a similar five-step scale, from “not useful at all” to “very useful”. The placeholders <Project Name> and <Workstep Name> are both replaced with the respective entries by the participant – in order to make the connection to the participant’s context directly visible in the questions.

5.4.2. Learning Index Definition

The previous subsection illustrated the data collection for the learning index. This section describes how this data is used to calculate a single scalar index for the on-the-job learning effect of a particular participant linked to a particular work context. In the following chapters, this **learning index** is used as the primary outcome variable of this study.

The learning index is defined as follows:

The learning index is the sum of the learning situations weighted by the respective learning importance and usefulness.

The primary underlying assumption is that the actual learning effect strongly correlates with the number of learning situations that a person can remember.

Mathematically this definition translates to equation 5.1:

$$\text{Learning Index}_{\text{Person } j, \text{Workstep } k} = \sum_i^{\text{all valid Learning Situations}} \left[\frac{1}{2} \left(\text{Learning Importance} + \text{Learning Usefulness} \right) \text{Learning Situation } i, \text{Person } j, \text{Workstep } k \right] \quad (5.1)$$

In the four cases that the learning usefulness was not available (due to an automatically reduced survey), only the learning importance is used as a proxy for the average of both (similar) items – as shown in equation 5.2 (see also the discussion of the construct for the lesson’s value in section 5.4.1 on page 146.). Since both learning importance and usefulness indicate the value of a particular lesson, the results of equations 5.1 and 5.2 can be and are mixed¹² in the final outcome variable vector (containing the learning outcomes for all participants). Since the questions on importance and usefulness are similar and related, the only major difference between the two equations is that the learning index values on both learning importance and usefulness are based on two rather than just one question item per learning situation and thus are likely to have a slightly higher quality. More details on the correlation of learning importance and usefulness are provided in appendix section A.4.1 on page 295.

$$\text{Learning Index}_{\text{Person } j, \text{Workstep } k} = \sum_i^{\text{all valid Learning Situations}} \left[\begin{array}{c} \text{Learning} \\ \text{Importance} \end{array} \text{ Learning Situation } i, \begin{array}{c} \text{Person } j, \\ \text{Workstep } k \end{array} \right] \quad (5.2)$$

Like psychometric constructs, both variants of the learning index are based on many question items, which increases reliability. The data from the text-entry fields even allow a manual (and thus intelligent) quality inspection of the entries.

Despite the bias-reducing features of the learning index (such as the context reference), a number of biases should be expected. Social desirability might lead some participants to report about their learning experience in an overly optimistic fashion. Given that the survey was introduced to the participants as a study on workplace learning, the participants will most likely choose a sample project that involves an above-average learning effect. Some participants might want to make a good impression by presenting themselves as more active learners than they actually are. And since the importance or usefulness of the learning effect is a subjective judgement, it might be subjectively biased. Moreover, the learning index measures the consciously experienced learning effect, but there may also be an unconscious or tacit learning effect that has effect on employee productivity. While conscious and unconscious learning frequently occur in conjunction (see section 2.3.8 on page 52), we do not directly measure the unconscious part with the learning index.

As will be detailed later in appendix section A.4.2 on page 295, the distribution of the surveyed learning index, with many participants at learning index zero, suggests that the bias due to social desirability is small.

¹²Depending on the survey-reduction level, a participant will or will not get questions about learning usefulness; thus either the full or the reduced version of the learning index will be available. Hence the results never overlap and can be merged into a single outcome vector.

Although these effects bias the learning index, the biases are systematic and uniform for a larger sample. Hence it is unlikely, e.g., that the effect of social desirability depends on independent variables such as education level – which would also bias the fitted statistical model. By a similar argument, the learning index is also seen as a sufficiently accurate relative proxy for total learning, including unconscious learning¹³.

The aim of the learning index is not to measure the true and absolute average learning effect for any project or task. To gain insights regarding the research questions, it is sufficient to compare the relative learning effect under the effect of different organizational factors, such as the working environment. Biases that do not depend on these organizational factors are therefore not a problem.

As section [5.11 on page 162](#) further confirms, the design of the learning index meets the quality requirements for this research purpose and is cost effective (section [3.1.8 on page 96](#)).

5.5. Survey Pilots

This section will describe the results of the survey piloting sessions – as a basis for the description and origin of the survey design goals next in section [5.6 on the facing page](#).

Based on the insights generated in the qualitative stage, a draft version of the online survey was constructed and tested in three stages with one, two and 12 Meyer Werft employees, respectively. In the pilot phase, it was possible to streamline the wording and the flow of the survey significantly for better and quicker understanding by the participants.

Yet the time requirement, combined with the large number constructs necessary for reasonably extensive models for learning and innovation, proved to be a formidable challenge. For a certain revision of the survey, most people took around 80 minutes to finish, but about 20% took much longer and were only done with about 30% of the survey after 80 minutes. Thus the survey was still too long, and, more importantly, different participants had very different speeds in completing it. Hence the participating researchers realized that further shortening of the survey (at great cost in model sophistication) would not solve the speed variability problem and therefore decided on another solution: The online survey was equipped with a program logic (the automatic flow control survey algorithm) that selects questions only if relevant to the person's task example and based on the participant's initial speed, in addition to some randomness (details follow in section [5.7 on the next page](#)).

The data from the pilot stage was not used in the statistical analysis.

¹³Supported by the literature cited in section [2.3.8 on page 52](#), the underlying assumption regarding unconscious learning is that conscious learning is a good predictor (i.e., is highly correlated) for unconscious learning and thus also correlated to total learning, combining both effects.

5.6. Survey Design Goals

With experiences from the pilot stage, the online survey was designed for the following design goals (in addition to the quality goals described in [chapter 3 on page 83](#)):

- The questions should refer to actual examples of the individual employee's work in order to ensure the connection to the actual organizational context and in order to reduce subjectivity and the influence of general attitudes about the survey participant (as already discussed in [section 5.4.1 on page 146](#)).
- The survey should not take longer than one hour on average – within reasonable bands, e.g., a minimum of 40 minutes and a maximum of 80 minutes.

The participants of this survey have very different backgrounds and thus will show very different speeds when working through the survey. Hence to meet the time requirement from above for all participants, the survey needs to adapt to the participant's speed.

- The survey should include a wide range of sub-factors describing the major factors listed in [section 5.3 on page 141](#), including a part on innovation (for Polina Isichenko) and a part on work motives (for Prof. Christian Roßnagel).
- The questions in the survey should come from standardized question item batteries forming validated constructs – as far as possible. In addition the constructs should also closely fit the study question and be applicable within the context. When these requirements were contradictory and compromises were necessary, preference was given to the context fit criterion – in line with the aims of the study and the methodology ([section 3.1.7 on page 95](#)).

5.7. Survey Algorithm

This section describes the overall survey algorithm, which steers the participant through different paths within the survey based on pre-programmed rules – in order to cover the wide spectrum of questions and in order to automatically adapt to the individual participant's speed. It includes the learning index survey tool described in [section 5.4.1 on page 146](#).

For its flexibility and simplicity of deployment (given the web-based client application), the *Unipark* online survey tool was chosen for this study. Technical details on Unipark's features and the implementation of this survey can be found in [text box 5.7.1 on the next page](#).

For the purpose of the algorithm, the survey can roughly be split into three *content branches* and one common section:

Text Box 5.7.1 Unipark – Features of a Flexible Web-Based Survey System

Unipark is the survey product by Globalpark AG, of Cologne, Germany (see also <http://www.unipark.de/>), used in both academic and commercial settings.

Out-of-the-box Unipark features:

- Participants can participate in the survey simply with their web browser and a special link to Unipark's web server.
- A web-based editor to quickly create large dynamic online surveys, complete with rules that check the consistency of the answers and, if necessary, force the participant to correct inconsistent answers (very important for the quality of the data).
- Unipark allows automatic generation of question pages for loops. In these (and other) pages, it can dynamically display the contents of variables within the question text – an example is mentioned in text box 5.4.1 on page 150.
- There is also some limited built-in support for survey flow control with conditional rules based on previous answers or other variables of the survey (except time, unfortunately).
- For greater flexibility, the researcher can add advanced features by embedding Java Script code in the survey pages and executing this code in the participant's web browser (see <http://developers.sun.com/scripting/javascript/>). Able to use any common Java Script function, the researcher can then perform any calculation, define any condition and feed the results back into the survey variables (which reside on the Unipark server).
- In addition, there is support for anonymous and non-anonymous surveys (with e-mail invitations). In the end, the researcher can download a single data file with the data from all participants.

Nevertheless, to realize the complex adaptive flow control mechanism of this survey, the built-in functionality was not sufficient, and thus embedded Java Scripts were used heavily for measuring time and pre-process variables for complex flow control decisions (including controlled random decisions).

- a common component (including, e.g., biographical information),
- learning component – (a),
- the innovation component – (b) and
- the work motives component plus the DNV knowledge questions – (c)

Since the content branches are not directly related, the algorithm selects only a single branch (a, b or c) plus the common section for each participant. Since a single branch plus the common section was still too long for some participants, the algorithm would select one of five different survey-reduction levels (reducing the number and scope of questions

of each branch) for each participant individually, depending on the participant's speed in answering questions at the beginning of the survey.

In addition, participants would classify their normal tasks by answering a few questions, and the algorithm would then ask them to select and name an example from one of the following three *task branch* categories:

- an innovation project,
- a normal (non-innovation) project, or
- the participant's normal work during the previous four weeks.

The remaining questions were then asked in the context of the work example chosen by the participant.

The pilot phase showed that the terms 'project' and 'innovation' had very different meanings for different participants. Thus a simple question asking them to choose between the three task branches was not effective and led many participants into the wrong branch – invalidating their entire dataset. Therefore a system involving multiple questions and consistency checks was used to reduce the number of participants getting into the wrong task branch – described in further detail below.

Using these three filtering stages (content branch, task branch and survey-reduction level) opened up 35 different ways to proceed through the survey¹⁴ and maximized the number of questions that could be answered within a reasonably predictable and limited time frame. With this algorithm, roughly 95% of the participants completed the survey within 50 to 70 minutes.

Since some of these mechanisms are non-random, there is a danger of undesired self-selection effects. Employees with less education, for example, may be systematically slower in filling out the survey and thus may only be represented with lower reduction levels. In that case, certain factors would be collected with higher frequency for those with higher levels of education.

Nevertheless, the algorithm has been designed following the aims of the methodology to keep these undesired and unavoidable self-selection effects within reasonable limits. As the following description will illustrate, the random elements of the flow control mitigate the self-selection effects, and controlling for self-selection¹⁵ makes it possible to detect certain types of self-selection.

The following algorithmic steps are executed at run-time, while the participant fills out the survey. The steps are individualized for each participant (see also figure 5.2):

¹⁴Not all permutations of content branch, task branch and reduction level make sense – thus the number of paths is fewer than 45.

¹⁵Using variables such as the “education level” see section 5.12.1 on page 163, it is possible to control for these self-selection effect and if necessary by means of sample weighting correct for them – which was not necessary for this survey.

5.7. Survey Algorithm

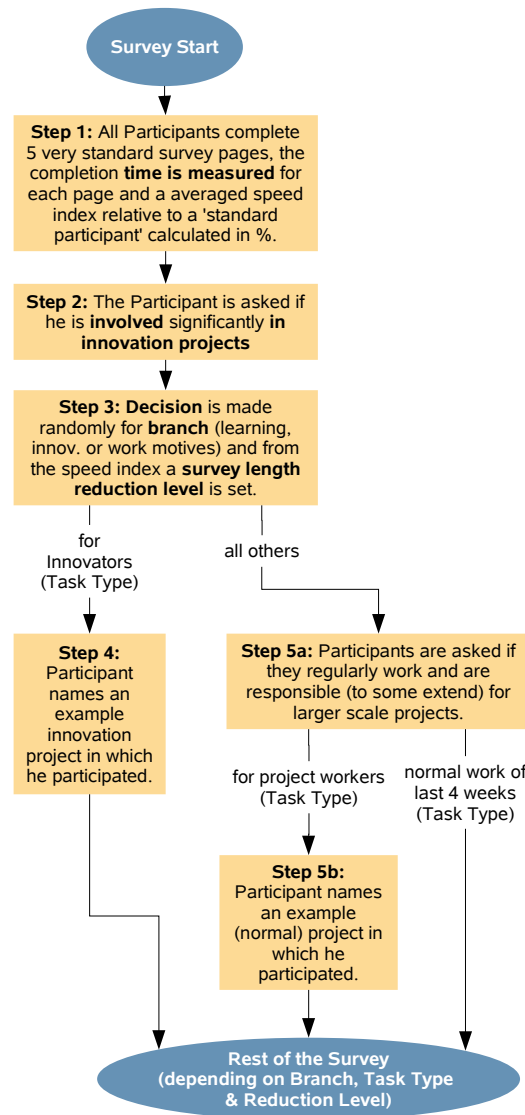


Figure 5.2.: The Interactive **Survey Algorithm** – The Example Task Selection Flow (Source: Author)

1. **Step 1:** For each of the first five survey pages, the participant's completion time is measured individually. Each measurement is then compared to a standard benchmark completion time, which was defined a priori based on the results of the pilot study. The result is a **speed index** measuring the participant's speed as a percent of the benchmark speed for each page. For use in the following decision stages, these page-speed indexes are averaged by an exponential moving average.
2. Next, in **step 2**, a series of questions are used to assess whether the participant was

substantially involved¹⁶ in an innovation project in the recent past. This pivotal question page was designed in a failsafe way: if the answers are inconsistent, either ‘no innovation participation’ is assumed or the participant receives an error message for some standard cases – asking him or her to rethink the answers before proceeding.

3. With this information collected, the central **decision-making component** of the survey is triggered. It behaves differently for innovation and non-innovation participants:

- **If** the participant was significantly involved in an **innovation** project, a random number is generated and used for randomly setting the content branch with the probabilities:
a) learning – 30%, b) innovation – 55% and c) work motives – 15%.
- **If** the participant was **not** involved in an **innovation**, a random number is generated and used to randomly set the content branch with the probabilities¹⁷:
a) learning – 67% and c) work motives – 33%.

In addition, the averaged speed index is used to determine the survey-reduction level, which further triggers or deactivates questions within the content branches:

Average Speed Index Range			Reduction Level
0 %	-	34 %	1
35 %	-	49 %	2
50 %	-	74 %	3
75 %	-	99 %	4
100 %	-	∞	5

4. **If** the participant has an **innovation** involvement and he/she is in **content branch b** (innovation), the participant is asked to think about this project and enter a name for it. This name is not used in the analysis but makes it possible to automatically use this participant specific name in all following questions, in order to ensure that the participant answers the question specifically about the chosen innovation and not in general¹⁸. Next, there are more detailed questions about this particular

¹⁶The questions probe for the type of personal participation, duration, size and novelty of the project. The involvement is substantial if the participant was a contributor – not just a user – of the innovation, and if he or she worked on it for more than 15 hours.

¹⁷Note that the innovation content branch is not possible for non-innovation task branches, which is the reason that the number of paths through the survey is limited to 35.

¹⁸During the pilot phase, I observed while talking with some participants during the survey that some people were strongly tempted to change their mind about which project they had chosen. Naming the project in the beginning reduces the risk of participants changing the project during the course of the survey.

5.8. Survey Conduction and Resulting Sample

innovation, and the following step about **selecting a project is skipped**.

Alternatively, **if** there was **innovation** participation but content branch **a** (learning) was randomly set, the participant is automatically forwarded to the next step.

5. Similar to the question page on innovation participation, the participant is asked whether he or she frequently works on larger-scale projects over a time span of at least three weeks. If the series of probing questions is answered in a particular way, the participant is assigned the task type **‘Project’** and then asked to select an example problem from his or her work and to assign it a name. If not, the assignment is **‘normal work’** from the previous four weeks. Again, additional questions for a consistency check with direct feedback to the participant were used.
6. At this point (about one-quarter into the survey), all variables determining the further shape of the survey for this participant are set:
 - Content Branch – a) learning, b) innovation and c) work-motives
 - Task Branch – 1) Innovation, 2) Project or 3) Normal Work, last four weeks
 - Reduction Level – an estimator on five levels for how many questions can be asked of this participant in order for the person to be done in about 60 minutes. For example, for each content branch there is a stripped-down package of questions (level 2), a normal package of questions (level 4) and a complete/luxury package of questions (level 5).
7. ... It **follows** a long series of questions that are either enabled or not, depending on the state of the above three variables and the survey-reduction level. Unless the most extreme survey reduction (level 1) is activated, the **learning index survey component** (figure 5.1 on page 147) follows next.

All meaningful permutations of these three variables lead to 35 different paths through the survey. However, not all of them have been equally frequently used – as will be described in section 5.9 on the next page. Fortunately, many participants were fast enough to get into reduction level 4 or above.

5.8. Survey Conduction and Resulting Sample

In parallel to the survey pilots (from section 5.5 on page 152), the survey was advertised to all 2,400 employees at Meyer Werft in Papenburg as a voluntary and anonymous survey. Participation was offered in two ways: either at a fixed time in a computer room session or over the web directly from the person’s work PC.

Within 2.5 months from the end of May 2007 to the beginning of August 2007, 446 employees participated in total.

As with any voluntary study, adverse selection effects may be present – i.e., a group sharing a particular property may systematically choose not to participate. For example, participation among white-collar employees was higher, at about 20%, compared with about 10% participation in the production departments. Yet since there was substantial participation across all departments, this difference is regarded as acceptable here.

5.9. Actual Performance of the Interactive Survey

This section shows some statistics about how the interactive survey was actually used in the field – as an indicator of how well the survey algorithm (section 5.7 on page 153) and the questions worked.

Task Branch	No. of Participants	% of Participants
innovation projects	89	27 %
large projects	96	29 %
short tasks during the last 4 weeks	144	44 %

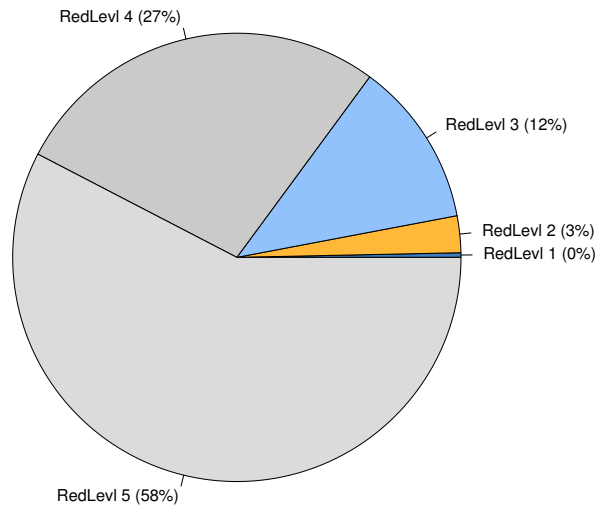
Table 5.2.: Usage of the 3 Different Task Branches

Table 5.2 shows the fraction of participants¹⁹ who went into each task branch. Given that most people’s work at the shipyard is dominated by tasks that take less than four weeks, it is not surprising that 44% of all participants chose a learning episode from such a shorter task. Nevertheless, the two other branches (innovation and large non-innovation projects) are also well represented, with almost 30% each.

As described in section 5.7 on page 153, the number of survey questions is automatically reduced, unless the participant was fast enough during the first few questions. The reduction is performed on five levels, where level 5 has the least reduction and thus the widest coverage of questions – yielding the most complete datasets. At the other extreme, level 1 is a very minimal set of questions that hardly allows a broad analysis. During the pilot stage, the questions and explanations during task branch selection and the learning frequency survey components were especially fine tuned – also with the aim of allowing the participants to understand and go through the questions more quickly.

As figure 5.3 on the following page and table 5.3 on the next page show, this fine tuning paid off. Most of the participants experienced no survey reduction at all. The general reduction of questions was not too strict but rather well balanced – considering that the next 39% of participants are spread over levels 3 and 4.

¹⁹Based on the mostly filtered data with n=329 (after filtering out inconsistent answers and before filtering out apprentices or due to missing values – see section A.5.1 on page 302).

Figure 5.3.: **Reduction levels** 5 and 4 are used most frequently.

Reduction Level	No. of Participants per Reduction Level	% Participants per Reduction Level
5 (no reduction)	174	58 %
4	83	27 %
3	36	12 %
2	8	3 %
1 (max reduction)	1	0 %

Table 5.3.: Reduction Level Usage

Figure 5.4 on the facing page further shows that the time participants needed to get through the survey stayed below the aim of 60 minutes in most cases²⁰. Moreover, the distribution of the survey duration does not change much across reduction levels, which indicates that another design goal for the interactive survey was achieved: the reduction-level mechanism automatically reduced the number of questions in order to reach a roughly equal survey duration for all participants.

²⁰Some of the very far outliers are due to a problem with measuring the duration when the participant does not finish the survey properly after the last page (though this is not a problem for the collection of data).

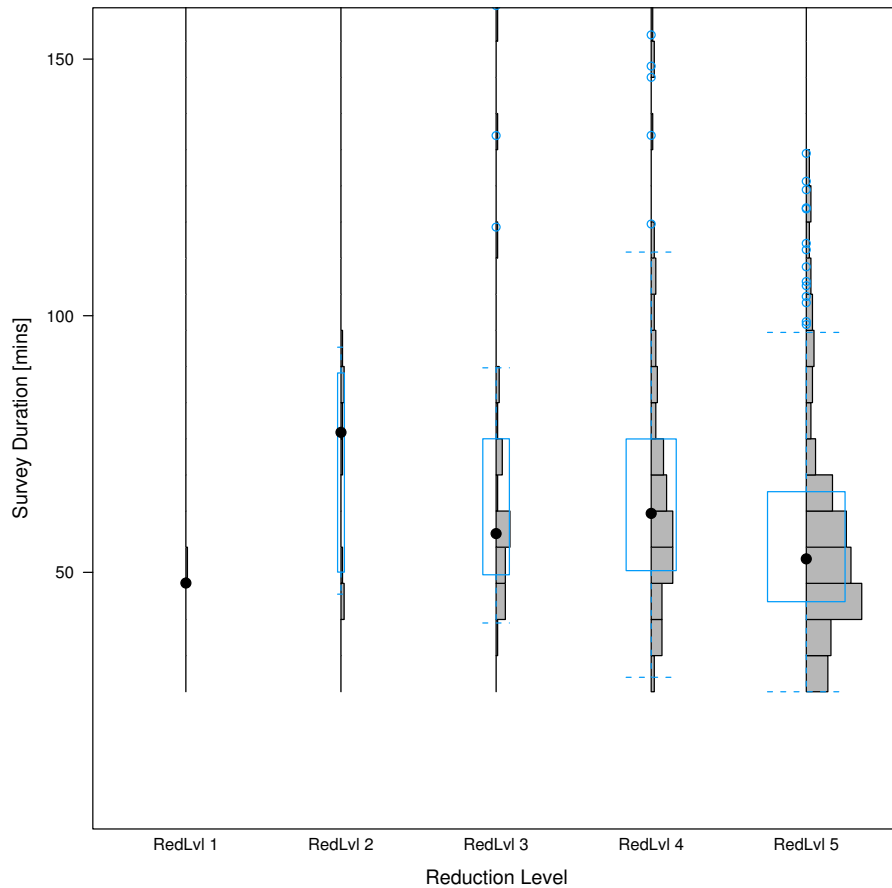


Figure 5.4.: The **Survey Duration Distribution** hardly changes across **Reduction Levels**

Hence, the collected statistics about how the survey algorithm worked in the field show that the design goals regarding time and question breadth were fulfilled.

5.10. Data Pre-Processing

Given the many paths that a participant can follow in the interactive survey, and given the many different types of questions and answer formats, extensive pre-processing of the raw survey data was necessary.

This pre-processing is performed by 3,800 lines of \mathbb{R} -code (see section A.6.2 on page 307) and involves operations such as:

- **Filter** the data in multiple stages – yet refrain from classical outlier removal. Details are provided in appendix section A.5.1 on page 302.

5.11. Validity Investigation of the Learning Index

- Correctly **assign** the **data types** of all scales.
- **Rescale** the numerical values from, e.g., the coded scale data (1, 2, 3, 4, 5) to (0, 0.25, 0.5, 0.75, 1). Note that this is not ordinary *normalization*, which depends on the mean and standard deviation of the data. The intention was to perform data pre-processing in such a way that the transformation steps are independent of the input data.
- Depending on the path actually chosen, mark all invalid or missing values correctly as **missing values (NA)**.
- Generate (i.e., calculate) some **new variables**, such as blue or white collar or department category (technical design, administration, production) based on other variables – here, the department of the participant.

In particular, the **calculation** of the **learning index** is included as an aggregation of information from multiple variables from the learning frequency survey tool (section [5.4.1 on page 146](#)).

- Imputation of the missing values by one of the simplest, most predictive, robust and conservative strategies: **mean imputation** – see appendix section [A.5.2 on page 303](#).

5.11. Validity Investigation of the Learning Index

Given the novelty of the learning index, its reliability and validity needs to be verified. Thus in appendix section [A.4 on page 290](#), the learning index's consistency – both internally and with other question items and constructs – is verified by the following tests:

- Statistics about the input data for the learning index, i.e., the learning situations, learning importance and learning usefulness data, were generated – and showed the expected (thus good) results.
- The distribution of the learning index was inspected and showed similarity of a log-normal distribution – a common distribution of many natural processes. Not surprisingly, there were also many participants who had a learning index of zero – who could not recall any specific learning episode – which is a good indication of low bias caused by social desirability (section [A.4.2 on page 295](#)). Moreover, a comparison of these distribution results with the answering behavior for the much simpler question on the general learning impression (in section [A.4.3 on page 297](#)) indicates that the link to concrete learning episodes has led to the desired effect of bias reduction for the recollection of past events (see section [5.4.1 on page 146](#)).

- The survey contains three questions that are expected to correlate with the learning index. For example, there is a question after the learning index survey tool asking the participant to give a general impression (i.e., rating) of the learning effect of the example project or task.

The correlation of the learning index with these variables was therefore investigated with distribution graphs in addition to the ordinary Pearson correlation. The results clearly show a visible correlation of all investigated variables with the learning index.

In summary, the learning index passed all mentioned validation tests.

5.12. Properties of the Data Set

The choice of statistical method depends largely on the type of data collected for analysis. In the case of this survey, the data properties made it necessary to design and implement the BOGER algorithm – see section 6.2 on page 179.

Therefore the properties of the survey dataset are presented in this section.

5.12.1. Multi-Variate Relationships / Collinearity / Correlations

From figure 5.5, it appears that the learning index strongly depends on the participant's education level. The two variables correlate with 0.21. Most readers (including the author) will quickly be able to think of a plausible theoretical explanation for this relationship.

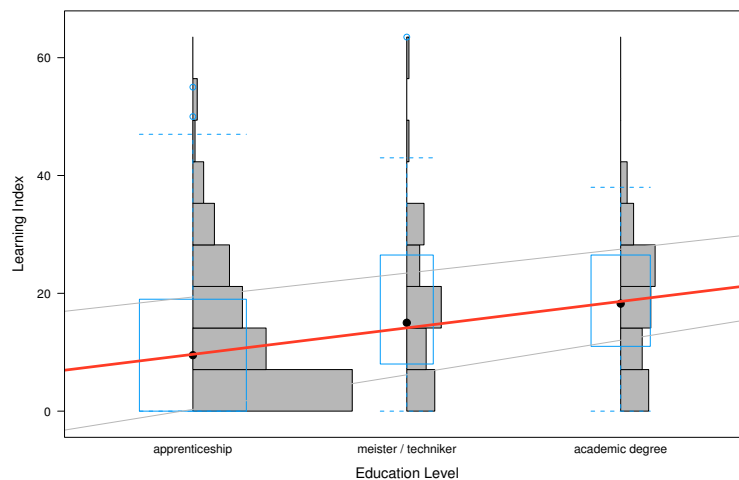


Figure 5.5.: Apparent Positive **Dependence** of **Learning Index** on **Education Level**.

Yet a few additional facts disturb this pretty picture:

At an early stage of the statistical analysis, a 140-page report was auto-generated²¹, listing the top 150 correlations of all 15,000 possible correlations of this dataset with about 120 variables. In line with the observations by [Starbuck \(2004\)](#), these top 150 correlations have a Pearson correlation of 0.36 or higher. Hence a correlation of 0.21 is not even high enough to be in the top 150.

In addition, education level appears to correlate with -0.37 even stronger to another variable: the task branch of the survey – i.e., the type of the example project or task. Figure 5.6 further underlines that there appears to be a relationship between education level and the task type²² (task branch).

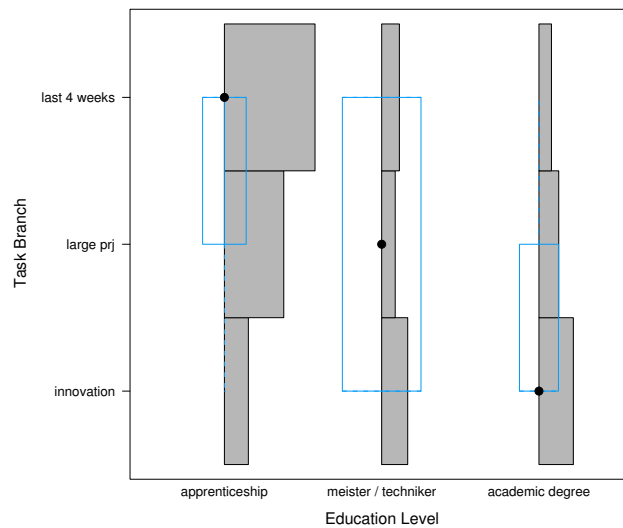


Figure 5.6.: An even **stronger Relationship: Task Branch and Education Level**.

Furthermore, the task branch also correlates with the learning index. The correlation is with -0.26 even stronger than the correlation between education and the learning index. See also figure 5.7.

These three pieces of evidence suggest that the information allowing a prediction of the learning index is shared by education level and task branch. Thus these two variables are collinear – as discussed in section 4.1.8 on page 117.

Moreover, this evidence on statistical association allows for all of the following conclusions:

²¹With the report generation capability of \mathbb{R} – see appendix section A.6.2 on page 307.

²²Strictly speaking, the task branch is an ordinal (i.e., category) variable without an interval sequence – hence in the actual analysis, this variable is broken into three dummy variables, which are referred to as the taskType group – see section 7.3.7 on page 250. For simplicity, the argument is made as if the task branch were an interval variable, which is acceptable because the relationships happen to behave the same as if they were.

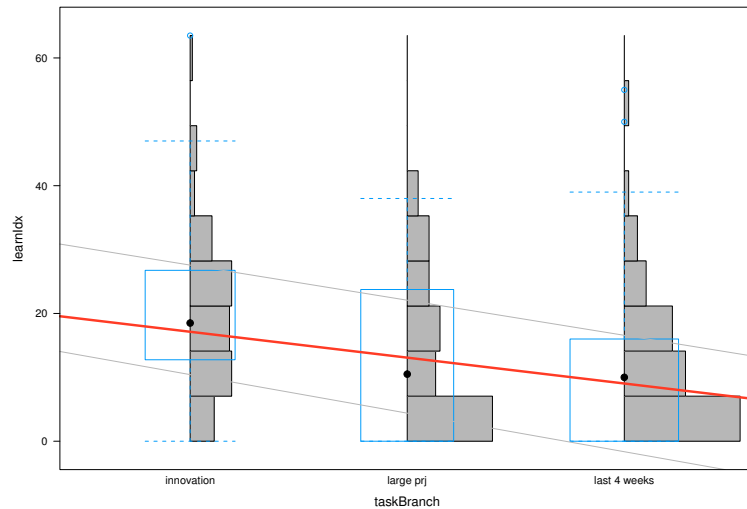


Figure 5.7.: Distribution of **Learning Index by Task Branch** – Showing a stronger Correlation than in Figure 5.5 on page 163.

1. Education level acts through the task type on the learning index – i.e., the task type mediates the effect of the education level.
2. The task type happens to correlate with the education level, which is plausible since higher-educated employees are more likely to work on innovation or larger projects than employees with less education are.

Task type has an effect on learning index, but education level has no direct effect on the subjectively rated relative learning effect. The correlation between education level and learning index is simply due to the correlation between task branch and education level.

3. The task type happens to correlate with the education level, and education has a direct effect on learning. The task branch has only a very weak effect on learning, since most of the correlation between the learning index and task branches stems from the relationship between education level and task branch.
4. Education level acts directly on learning. In addition, the education level is related to the task type, but this is not related to the relationship between task type and learning index.

With a simple correlation analysis or a uni-variate graphical analysis with figures such as the conditional distribution figure 5.7, the researcher cannot gain any further insights for determining which of the above hypotheses holds true. Assumptions about causality cannot be justified either, since all of the above variations appear plausible.

There is, however, one method to test the above hypotheses without further assumptions: fitting a suitable **multi-variate** statistical model.

As a brief preview of the statistical results chapter [7 on page 205](#), I will note here the results from fitting the multi-variate BOGER model to the data: when both independent variables were in the model, the term for education level was not stable (significant), nor was an interaction between task type²³ and education level. Hence education level not linked with the task type and does not directly act on the learning index, but a direct and stable effect of task type on the learning index could be detected. Thus hypothesis 2 could be confirmed.

In summary, the many correlations indicate that the survey dataset is substantially collinear. Due to these collinearities and many plausible explanations from theory, a suitable multi-variate statistical modelling approach should be pursued.

5.12.2. Noise

Figure [5.8 on the facing page](#) shows the dependency of the learning index on the personality trait “openness to new experiences” from the psychometric standard scale: Big-Five (see section [7.3.11 on page 257](#)). The various features of figure [5.8](#) are explained in text box [5.12.1](#) and with more background and an annotated figure in section [7.1.3 on page 213](#).

Text Box 5.12.1 Dependent Distribution Figures Explained

This ‘*dependent distribution*’ figure type, used extensively for this study, is a more information-rich alternative than scatterplots to visualize dependencies of a continuous variable on a category (ordinal) variable (continuous independent variables are automatically converted to category variables).

The data of the dependent variable (on the y-axis) is grouped by the independent variable (on the x-axis).

Instead of scatters, the dependent variable’s distribution for each group is plotted by independent variable. The distribution makes visible any dependencies *and* the noise.

To make dependencies even more obvious, boxplots of the grouped dependent variable data are overlaid. The range inside the box contains 50% of the data. In addition, the median of the grouped data is plotted as a thick black dot. An ordinary linear regression line – the thick red line – is fitted through these median points. This fit is weighted by the number of samples in each group. The thinner grey lines indicate similar weighted regression fits through the 25% and 75% quantile points (the hinges of the boxplots) – giving an indication of the behavior of the variance.

See also figure [7.2 on page 215](#) for an annotated example. More detailed background explanations are given in section [7.1.3 on page 213](#).

²³In the actual analysis, this the task branch variable is broken into three dummy variables (with values either 0 or 1). The variable group is referred to as the taskType group – see section [7.3.7 on page 250](#).

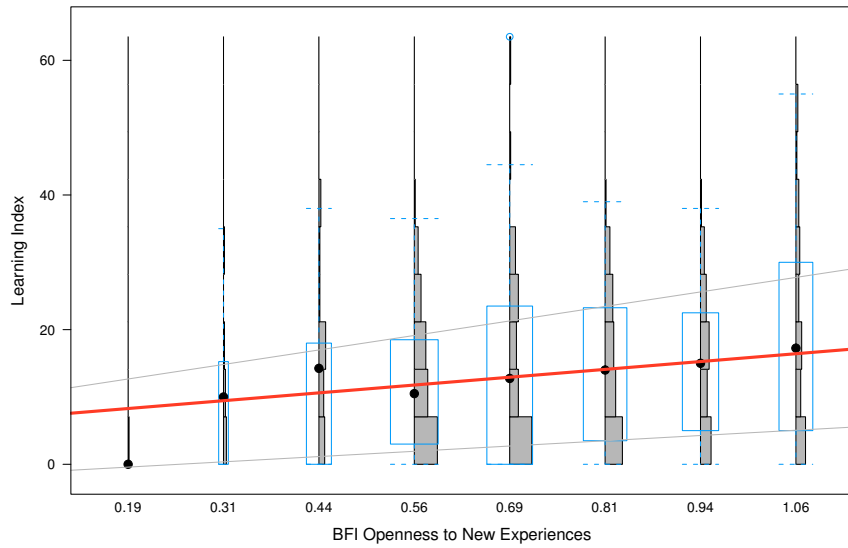


Figure 5.8.: A Noisy Relationship: Distribution of **Learning Index** by **BFI Openness** level.

What becomes directly evident from figure 5.8 is that the effect of openness is masked by a lot of noise, which appears as a large variance on the learning index over the different levels of openness (large compared to the effect of openness on learning). This noise or variance can stem from truly random noise²⁴ in the process, but most of it will stem from variance explained by other variables. Some of these variables are included in this survey. If they have explanatory power, they are included in the multi-variate statistical model (described in chapter 6.2 on page 179), which then increases the explained variance of the model and reduces the level of noise in the predictions.

However, by the nature of the survey subject, there will also be many latent variables that affect learning but that have not been included in the survey. There will be a large group of variables related to the situation of the learning episode, e.g., detailed and non-recurring properties of the task or the learning opportunity; the participant's current mood; the involved colleagues, superiors, suppliers (including the participant's personal relationship with these collaborators) etc. Hence even if there is a learning-supportive organizational environment, there are many situational factors that may hinder or completely block learning in a particular situation. This also explains the large amount of zero-learning cases in the survey data – see appendix section A.4.2 on page 295.

These situational variables have not been included in the survey since they are difficult or impossible to survey in a standardized form, and since they cannot be influenced

²⁴Random noise that cannot be explained by any variable – even not a latent variable.

by the management of the organization and thus do not represent organizational levers for supporting learning. While this approach makes it possible to keep the survey at a reasonable length, it limits the maximum achievable explained variance by the model, since latent variables – by the lack of data on them – cannot be included in any multi-variate model. Therefore a substantial and irreducible fraction of noise will remain for the analysis (see also section [6.3.2 on page 199](#) on model fit).

Aside from the noise, the fit of the medians (the red line) suggests that the personality trait of “openness” leads to an increased learning effect. The boxplots are a little less assuring of this relationship. The Pearson correlation of only 0.17 also suggests that there is at best a very weak relationship between openness and learning.

Yet as will be described in the statistical results in section [7.3.11 on page 257](#), this is a factor that is stable (i.e., significant) in the final statistical model. The magnitude of the effect is also weak in the statistical model, since openness is one of the weaker factors in the final model, and large fractions of the variance are explained by other stronger independent variables.

In summary, there is a high level of noise in the data, which for the analysis can be reduced to some extent with a multi-variate model. Thus the noise that masks effects in the uni-variate analysis can be reduced by a multi-variate model. However, given that a substantial part of the stochastic process is driven by latent variables that describe the specific learning situation, a substantial fraction of noise will remain – even in the final analysis with a multi-variate model.

5.12.3. Non-Linear Relationships

As will be presented in the statistical results in section [7.3.3 on page 238](#), one of the strongest factors driving learning is personal interest in a topic, assessed by the question item:

“The topic was also interesting to me independently from my tasks.”

“Mich hat das Thema auch persönlich unabhängig von meinen Aufgaben interessiert.” [German Original]

Yet figure [5.9 on the next page](#) shows an unusual relationship. The red-line median fit suggests a strong linear relationship, in line with the strong linear correlation of 0.294. However, the conditional distributions and boxplots instead suggest that the learning index is acting on the variance like a fade-in amplitude function acts on a sinusoidal.

An amplitude function, however, makes ordinary linear regression models non-linear²⁵ when the amplitude function is multiplied with a group of ordinary linear regression terms. Hence this strong cone-shaped relationship in figure 5.9 suggests that some of the stronger effects in the data are non-linear – which is later confirmed in the statistical results in section 7.3.3 on page 238.

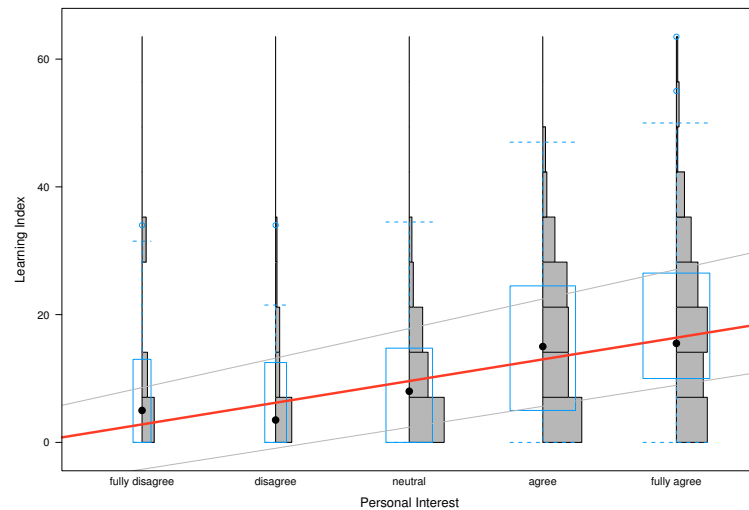


Figure 5.9.: A **Non-linear Relationship**: An increasing Personal Interest is associated with an increasing mean *and* **variance** of the Learning Index Distribution.

In summary, even before the modelling effort with BOGER there was evidence that some of the stronger effects would be non-linear (and more specifically multiplicative), which led to the non-linear design goal for BOGER (section 6.2.1 on page 179).

²⁵Strictly speaking, multiplying the sum of ordinary linear regression terms (in parentheses) by a variable does not make the regression model non-linear, since this multiplication is equivalent to replacing all original terms with the respective interaction with the multiplied variable. Multiplying a function with shape parameter that needs to be fitted to linear regression terms, however, does make the regression model non-linear.

6. Statistical Analysis with BOGER

Chapter Contents

6.1. Performance of Existing Algorithms	172
6.1.1. Choice of Existing Algorithms	172
6.1.2. Comparison Method	174
6.1.3. Comparison Results – Existing Algorithms	175
6.1.4. The Process of the Analysis	177
6.2. Design of the BOGER Algorithm	179
6.2.1. Design Goals	179
6.2.2. The Mathematical Model	180
6.2.3. Genetic (Non-linear) Function Fit Algorithm	183
6.2.4. Overview on Model Building and Robustness Testing	185
6.2.5. The Fully Automatic Screening Stage	187
6.2.6. The Final Stage of Model Selection	192
6.2.7. Robust Model Fit Estimation in BOGER	195
6.2.8. Implementation in \mathbb{R}	197
6.3. Design Requirement Validation & Performance of BOGER .	198
6.3.1. Model Selection Progress	198
6.3.2. Model Fit & Predictive Power Estimate	199
6.3.3. Design Requirement Validation Summary	202

6.1. Performance of Existing Algorithms

Previously, in [section 5.12 on page 163](#), the properties of the survey data set were described as multi-variate, collinear, noisy and non-linear.

Initially the plan was to use existing statistical analysis methods to build a multi-variate model of the data in order to obtain a ranking of the most important factors acting on individual learning. After various trials, first with conventional and later with state-of-the-art algorithms, it became evident that no existing algorithm was suitable for building a statistical model for the survey data with an acceptable level of predictive power and robustness ([section 6.1](#)).

Therefore, drawing on ideas from various state-of-the-art algorithms, a new statistical model-building algorithm was developed for this study, which handled the specific properties of the survey data significantly better than any of the existing algorithms. The design of this new algorithm, called BOGER¹, is described in [section 6.2 on page 179](#), followed by a performance evaluation in [section 6.3 on page 198](#).

6.1. Performance of Existing Algorithms

[Section 4 on page 101](#) laid out the theoretical requirements for a statistical analysis algorithm suitable for this application, based on existing literature. This section will illustrate with the model fitting performance of different algorithms, how these requirements have practical relevance to this study and why none of the tested existing algorithms satisfactorily met the requirements.

After the description of a systematic comparison method and its application to the presented algorithms, qualitative descriptions of the individual algorithm performance follow, with discussions of the reasons for the unsatisfactory behavior.

6.1.1. Choice of Existing Algorithms

While attempting to find an adequate statistical model of the survey data, a number of existing algorithms were tested. The insight gained during this experimentation process led to the formulation of the requirements for a suitable statistical algorithm, given the dataset at hand (details will follow in [section 6.2.1 on page 179](#)).

During the experimentation process, a number of algorithms (e.g., structural equations modeling, or SEM) could be excluded from the trials. Further details on the “story of the analysis”, explaining how the algorithms were chosen during the course of the analysis, are provided in [section 6.1.4 on page 177](#).

The following lists and briefly describes the tested algorithms (test results follow in [section 6.1.3 on page 175](#)):

¹BOGER is short for *BOotstrapped GENetic Regression*

- a) **Static Multi-Variate Linear Regression** An ordinary multi-variate linear regression (Backhaus et al., 2006, Chpt. 1) was run with an a priori defined model based on the state of knowledge from the literature research at the time of the analysis².
- b) **Breiman's RandomForest** The original random forest algorithm by Breiman (2001a). RandomForest is essentially a classification algorithm (section 4.1.4 on page 108). When used with a metric outcome variable, it divides the outcome variable into a number of intervals. These intervals become a set of unordered³ classification categories, against which the internal and native RandomForest classification algorithm is run.

The RandomForest classification works by first searching for the independent variable that is the best predictor for the outcome categories. This variable, which is either categorical or metric (at an optimized 'split' threshold), becomes the first node in the tree. The data is split into two branches depending on the threshold of the independent variable. Then, for each branch of the tree, the next-best variable that predicts the outcome of the branch's data sub-set is selected as the next node and branch point. By repeating these steps recursively for the sub-branches until no variables are left (or a stopping criterion is reached), a decision tree leading to the outcome categories based on the independent variable settings is created⁴.

Next, the entire tree-growing process is repeated based on different bootstrap samples (similar to the mechanism used in BOGER – see section 6.2.4 on page 185), yielding a whole ensemble of trees, i.e., a forest. Breiman and others (e.g., Chatfield (1995, p. 428) and Yuan and Yang (2005)) claim that this bagged ensemble predictor is more robust to sampling biases than building a single statistical model. The only downside is that a bagged model cannot be easily inspected for statistical inference without more sophisticated methods such as Breiman's variable importance measure (see section 7.1.1 on page 206). For more details, see Breiman (2001a,b); Strobl et al. (2007).

Since model selection and model fitting are combined in each step of the tree-growing process, RandomForest integrates model selection and fitting in an inseparable manner.

- c) **cForest** cForest by Strobl et al. (2007) is an improved version of Breiman's RandomForest algorithm, with some minor differences in the bootstrapping method and tree-split selection mechanism (offering different choices for the branch split definition method: significance testing or Monte-Carlo). Strobl et al. claim that their improve-

²See also section 3.2 on page 97.

³Since RandomForest loses the order of different outcome variable levels, it loses a significant part of the information contained in the data set. Thus RandomForest is less data-efficient for a regression problem than for a classification problem.

⁴Note that this is an expected value model.

6.1. Performance of Existing Algorithms

ments increase the robustness of the tree-growing process – i.e., the robustness of combined model selection and fitting.

- d) **Genetic Function Fitter** The Genetic Function Fitter is an early version of the non-linear function fit component used within BOGER (see section 6.2.3 on page 183) without many of BOGER’s other important features, such as algorithmic variable selection, bootstrapping and bagging. It was run with an a priori selected statistical model, using the same variables as the static linear multi-variate regression model, with the addition of non-linear terms⁵ for the same variables. Like RandomForest and cForest, the genetic function fitter is able to fit a non-linear model – yet in contrast to the two forest-building algorithms, it fits a *parametric* and thus much less flexible non-linear model.
- e) **Static Regression with RandomForest Variables** Another ordinary multivariate linear regression fit with a model consisting of the variables as selected by Breiman’s random forest.
- f) **Genetic Function Fitter with RandomForest Variables** Another genetic model fit with a model consisting of the variables as selected by Breiman’s random forest.

6.1.2. Comparison Method

A number of statistical model fitting algorithms were tested on the survey data. In line with the discussion on mode fit estimation bias in section 4.2.2 on page 123, the models were not only compared by the ordinary R^2 model fit (section 4.2.1 on page 119) but also by a rough estimate for the prediction error, indicating the model’s predictive power for future samples.

The prediction-error estimation was performed with the following method, which is equivalent to a simple cross-validation prediction-error estimate (as presented in section 4.2.4 on page 129): First, the data-filtering and simple mean imputation method were frozen – i.e., they remained unchanged for all tests mentioned below. Then the data points were randomly assigned labels from 1 to 10 – hence the data was separated into 10 different randomly selected slices. Different combinations of slices make up three different (yet overlapping) data sets A, B and C – as shown in table 6.1.

Next, each algorithm was run on the three different data sets A, B and C – each consisting of:

- 204 samples, or 70%, “training data”, which the algorithms used for model fitting, plus

⁵The non-linear terms were the same as described in the BOGER math model section 6.2.2 on page 180.

Data Set	Training Data Slices	Test Data Slices
A	1, 3–6, 8, 10	2, 7, 9
B	2–5, 7, 9, 10	1, 6, 8
C	1, 2, 5–9	3, 4, 10

Table 6.1.: Simple Cross-Validation Data Sets A, B, C

- 88 samples, or 30%, “test data”, which the algorithms did not use. Instead, it was used exclusively for cross-validation to get a model fit estimate, which is independent of the model fit and thus free of the bias due to overfitting⁶ (section 4.2.2 on page 123).

As the final step, the training and cross-validation test fit performance, measured by the simple R^2 estimator⁷, were averaged over the data sets A, B and C (for training and test datasets separately) – as shown in table 6.2 on the next page. A high-quality model would thus show both a high training and test model fit average and have little difference between the training and test fit – indicating low overfitting.

This method is a crude threefold version of Breiman’s cross-validation estimator of the prediction error – see section 4.2.4 on page 129. Since only three test sets were used, and strong fluctuations of the individual test fit results are evident, the precision of the averaged test fit result is limited. Judging from the fluctuations alone, I would estimate the accuracy of the test fit results to be within $\pm 3\%$ (absolute R^2 deviation).

6.1.3. Comparison Results – Existing Algorithms

Using the model fit measures defined in the previous section, table 6.2 on the next page shows the model fit results for the different algorithms. The labels a) – f) correspond to labels in the algorithm description list from section 6.1.1 on page 172.

The a priori frozen statistical model used for the ordinary multi-variate linear regression model achieved a reasonably good training model fit but almost no cross-validation model fit on the test data – which indicates that this regression model appears to be sound when using the conventional tests, but in fact is completely useless for prediction. Trial e) with a different a priori model suggests that this problem is not only due to the choice of model.

RandomForest achieves the best training dataset performance but with 6% only a very low cross-validation model fit (i.e., a low predictive power) – even if this cross-validation model fit is the second-best of the tested algorithms. Hence RandomForest is strongly

⁶As mentioned in section 4.2 on page 119, the cross-validation model fit estimate is free of the overfitting bias but not necessarily of other biases, e.g., the sample bias due to a sample that is too small. Nevertheless, it is useful in comparison with the training data model fit estimate.

⁷For the definition of the R^2 model fit, see section 4.2.1 on page 119.

	R ² Model Fit to ...						Average R ²	
	Data Set A		Data Set B		Data Set C		Train.	Test
	Train.	Test	Train.	Test	Train.	Test		
a) Static Regression (with variables as above)	26 %	−1 %	21 %	−3 %	22 %	6 %	23.0 %	0.7 %
b) Breiman Random Forest	81 %	10 %	81 %	2 %	82 %	6 %	81.3 %	6.0 %
c) Strobl's cForest	31 %	8 %	32 %	13 %	34 %	9 %	32.3 %	9.8 %
d) Genetic Function Fitter	24 %	9 %	20 %	−8 %	24* %	14 %	22.5 %	5.2 %
e) Static Regression (with variables from Breiman random Forest)	30 %	−6 %			24 %	4 %	27.0 %	−1.0 %
f) Genetic Function Fitter (with variables from Breiman random Forest)	26 %	−1 %					26.0 %	−1.0 %

Table 6.2.: Model Fit Comparison of Existing Algorithms

overfitting (section 4.2.2 on page 123), which implies that any insights about the model (e.g., variable importance) will not reflect reality beyond the training data sample.

The cForest algorithm showed the best, but still low, predictive power, with almost 10%. It still significantly overfits (32% training data fit) but much less than RandomForest does.

Of all algorithms, the genetic function shows the least overfitting but yields a lower predictive power than the two tree-building algorithms. Trial f) indicates an even lower performance and higher overfitting. Hence these results appear to depend on the chosen model.

In summary, all model fit results are very low – even by social science standards for

noisy data⁸. This is particularly problematic since low model fits cause the parameter estimates and variable selection results to become unstable (non-robust) – not surprisingly, given that the fitted model only very crudely represents the real underlying stochastic process (see section 4.1.3 on page 105).

6.1.4. The Process of the Analysis

This section describes the sequence of events (the actual path of the analysis process) that led to the selection of the algorithms in the comparison presented in the preceding section 6.1.3. In addition to describing the rationale behind the final algorithm choice, it presents further details on the behavior of the tested algorithms.

The experimenting started with simple uni-variate graphical analyses and a correlation analysis, which led to results as described in section 5.12 on page 163 as well as the primary insight that a multi-variate statistical inference model is necessary for this dataset (as discussed in section 5.12.1 on page 163).

Despite the indications for non-linear relationships, the statistical inference model building effort began with an ordinary multi-variate linear regression and the tweaking of the imputation method (appendix section A.5.2 on page 303).

The first results were encouraging and somewhat matched the theory and other plausible explanations (e.g., ‘*education level*’ was part of the model – see sections 5.12.1 on page 163). However, after further experimenting with the imputation method, I noticed that the effect strengths and significance of the regression parameters would vary with different imputation methods. More disturbingly, the regression results also changed with the same imputation method but different filtered sub-datasets⁹. The regression results (parameter effect strength and significance) even changed between fitting the model to a larger dataset A and fitting the model to a smaller dataset A’, which was only a subset of A. This behavior suggests that even the inflexible non-linear regression model was subject to overfitting (section 4.2.2 on page 123), which is confirmed by the close-to-zero cross-validation results from table 6.1.3 on page 175. Further experimentation with different a priori frozen models showed that the algorithm model fit performance results would be affected by the choice of a priori model but that the cross-validation results generally remained very poor.

The linear regression model’s low predictive power due to overfitting led to the decision

⁸In engineering, the researcher frequently has the opportunity to collect a sufficiently large sample with little effort and cost by automatic means. In these cases, the R^2 model fit frequently exceeds 0.7. However, in the social sciences, R^2 model fits around 0.2 are not uncommon – given large numbers of variables, relatively small sample sizes, hidden latent variables and stochastic effects.

⁹Reducing the need for imputation, I experimented with various filtering settings to get a mostly complete dataset for an a priori reduced set of variables, with the original intention to get a first regression model without the effects of imputation. For the actual analysis, the filtering was later relaxed again – as described in appendix section A.5.1 on page 302.

6.1. Performance of Existing Algorithms

not to test other, more sophisticated algorithms that a) use a linear model¹⁰ and b) do not feature an algorithmic variable selection.

The low model fit and the indications for non-linear relationships led me next to try two non-parametric algorithms, which have features to mitigate overfitting: Breiman's RandomForest (Breiman, 2001a) and a variant called cForest (Strobl et al., 2007). Unlike the regression trials that were performed with an a priori fixed model¹¹, both Forest building algorithms include the automated building of a decision tree – i.e., fully algorithmic model selection and fitting¹².

As discussed in section 6.1.3 on page 175 on the algorithm results, both RandomForest and cForest are overfitting, with RandomForest much more so than cForest. Both algorithms show a low cross-validation model fit below 10%. While the cross-validation model fit performance is not particularly good, the model selection behavior of the two algorithms is even less acceptable for statistical inference – for the data set of this study. The variables that the two algorithms identified as the most important ones¹³ varied widely for different data sets (A, B, C).

The low performance of the two forest algorithms, which in literature with classification problems showed superior performance, led me to understand that a) using a classification algorithm for a problem with a metric outcome variable is not very data efficient¹⁴, which is a severe disadvantage given the number of variables and the relatively small sample size; and b) given the level of noise in the data, non-parametric algorithms are too flexible and thus led to an unacceptable risk of overfitting (section 4.2.2 on page 123).

Therefore the genetic function fitter (section 6.2.3 on page 183) was developed as a model that is more flexible than the ordinary linear regression and less flexible than a non-parametric model. My (unfulfilled) hope was that a higher fit (but little overfit) would mitigate the non-robustness of the ordinary linear regression. Yet although the genetic function fitter actually did show comparatively good overfitting performance, its cross-validation fit and thus predictive power results fell short of the two forest algorithms (see table 6.1.3 on page 175).

In summary, the low performance of the tested algorithms and their model selection

¹⁰This includes a decision against structural equations modelling (SEM), which is based on the matrix of Pearson correlations (Backhaus et al., 2006) and thus on the assumption that the relationships are at least approximately linear (Guyon et al., 2006; Hothorn et al., 2008).

¹¹The were also some less serious attempts with step-wise regression, which led to the expectedly very non-robust results (Miller, 1984).

¹²Given the tree-building mechanism, there is no differentiation of model selection and fitting. Instead, model selection and fitting are integrated in a single tree-growing process.

¹³The order of most important variables was identified by using Breiman's variable importance measure – see section 7.1.1 on page 206.

¹⁴Both RandomForest and cForest convert regression problems with metric variables into classification problems by slicing the metric outcome variable into several intervals, which the native classification algorithm uses as non-ordered categories. Given the loss of the order of different outcome variable levels, a significant part of the information contained in the dataset is lost. Thus RandomForest and cForest are less data-efficient for a regression problem than for a classification problem.

behavior led to the following insights regarding the requirements for a suitable algorithm:

- use a non-linear but parametric model
- feature a variable-selection procedure
- use the metric outcome variable natively and thus in a data-efficient manner, retaining its sequence information
- show low overfitting
- yield an acceptably high (cross-validated) model fit – i.e., a high predictive power

These requirements, and further requirements listed in detail in section 6.2.1, led to the development of the BOGER algorithm – based on the genetic function fitter – as detailed in the next section.

6.2. Design of the BOGER Algorithm

The particular nature of the survey data (section 5.12 on page 163) and the unsatisfying performance of existing model-selection and fitting algorithms (section 6.1.3 on page 175) led to the development of BOGER. This section describes the design goals and details of the BOGER algorithm – addressing the algorithmic modeling challenges described in the statistical theory chapter 4 on page 101 – followed by details on BOGER’s implementation. Section 6.2.4 on page 185 provides an overview of BOGER’s model selection mechanism, while section 6.2.2 on the next page provides details on the underlying statistical model in mathematical form.

BOGER is short for *BOotstrapped GENetic Regression*.

6.2.1. Design Goals

Given that even modern existing algorithms (as discussed in section 6.1.1 on page 172) do not adequately meet the challenges of this dataset (as described in section 6.1.3 on page 175 and 6.1.4 on page 177), the BOGER algorithm was designed to meet the following challenges:

- Build a **multi-variate model** – given the collinear nature of the survey data (see also section 5.12.1 on page 163).
- **From a large set of potentially relevant variables, semi-automatically select a much smaller set of variables** for inclusion in the statistical model. Given that the number of potentially relevant variables is far too large compared to the sample size of the survey dataset, the number of variables needs to be drastically reduced in order to build a robust statistical model.

- Test the **model robustness** by a state-of-the-art and robust method: **bootstrapping**¹⁵ – in order to directly and robustly assess the effect of the sampling error by statistical simulation.
- Yield robust results with **multi-collinearity** in the data – i.e., deal with explanatory information that is shared by multiple variables (as discussed in section 4.1.8 on page 116).
- **Restrict the complexity of the underlying mathematical model**, but retain sufficient model flexibility for a good model fit, which may require **non-linear functions and interactions**. The model should not, however, be fully non-parametric, in order to **mitigate** the risk of **overfitting** (as discussed in section 4.2.2 on page 123).
- Given the large number of variables relative to the sample size and the noise in the data, **make maximum use of the available information** inherent in the data – i.e., achieve a **high data efficiency** by directly using the data rather than an intermediate, such as the correlation¹⁶, and by retaining the sequence and distance information embedded in the **metric scales** used in this survey¹⁷.
- **Yield a higher explained variance** with a robust and non-overfitted model **than any of the other algorithms** presented in section 6.1.3 on page 175 – given the nature and size of the dataset of this study (see section 5.12 on page 163).

Section 6.3.3 on page 202 will summarize how BOGER's design features and resulting performance (section 6.3.2 on page 199) fulfill these design goals.

The following sections will first illustrate the core components of BOGER before describing the model selection algorithm, which leverages the core components. The core components are the parametric mathematical model, the genetic algorithm used to estimate the parameters of the mathematical model, and the parameter-instability estimation method involving bootstrapping.

6.2.2. The Mathematical Model

In order to reduce overfitting (as observed for the non-parametric RandomForest algorithm in section 6.1.4 on page 177 and the survey data), the BOGER algorithm uses a parametric

¹⁵For more on bootstrapping and alternative resampling techniques, see 4.2.4 on page 129.

¹⁶Some algorithms, such as factor analysis or structural equations modelling (SEM), build their models only on statistical intermediates (SEM uses the correlation) (Backhaus et al., 2006; Pearl, 2003). Using an intermediate for statistical analysis may be acceptable if a large sample size is available to build a model with few variables.

¹⁷Many modern algorithms are internally classification algorithms (e.g., random forest), which do not internally make use of the sequence and distance information inherent in metric scales. To use these algorithms, the metric data is converted into category (class) data – with the loss of the sequence and distance information.

mathematical model, which restricts the flexibility of the model. Making the model less flexible is effectively a trade-off between achievable modeling accuracy (i.e., the sample data's fit with the model) and model robustness due to reduced overfitting (see also section 4.2.2 on page 123). Yet a model that most accurately models the sample data but then generalizes poorly to the entire population (i.e., has poor robustness and little predictive power) is of no value¹⁸. Therefore a simple robust model is more valuable for this study than a complex, accurate but non-generalizable model.

The simplest model would have been a simple linear one, such as $y = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n$. A linear model has a number of advantages, mainly: due to its simplicity there is little risk of overfitting, and the parameters fitting the model optimally to the data can be calculated with a simple, very computationally efficient and fast matrix operation¹⁹. Furthermore, a linear model can be extended without losing these advantages with linear interaction terms (e.g., $\alpha_{12} x_1 x_2$). These are the reasons that linear models form the basis for many popular statistical methods, including multi-variate regression, factor models, structural equation models and logistic regression (Backhaus et al., 2006).

However, many relationships in organizations are not linear – e.g., if two factors are connected by a logical AND relation. A relationship such as “*employees need to be qualified (x_q) AND motivated (x_m) to perform well (y)*” implies for the negative outcome that not being qualified *and* / *or* not motivated leads to bad performance. This is markedly different from a linear relationship, which implies that *either* qualification or motivation both separately *or* in combination lead to some degree of good performance – e.g., being motivated but not qualified still leads to a fair performance.

The AND relationship translates mathematically into a multiplicative function such as $y = \alpha f_1(x_g, s_1) f_2(x_m, s_2)$ where $f_1(\)$ and $f_2(\)$ are reshaping functions with shape parameters s_1 and s_2 – e.g.: $f_1(x_g, s_1) = x_g^{s_1}$ with $0 < s_1 \leq 1$. If the variables are appropriately reshaped, the relationship simplifies to $y' = x_g' x_m'$ (where the symbol ' means reshaped). With this simplification, the term $x_g' x_m'$ is a linear interaction term, if the shape parameters s_1 and s_2 are held fixed during the model fitting process. However, in most cases the reshaping is unknown and needs to be adjusted in the model fitting process to optimize the model fit. Yet more difficult are feedback systems, which may be described by a linear differential equation, but their directly observable solution function is highly non-linear (Senturia and Wedlock, 1993).

To model at least overall AND-relationships in addition to simple linear terms and interactions, the following mathematical model (eq. 6.1) is fit to the sample data by a non-linear function fit algorithm leveraged by BOGER:

¹⁸For a discussion of fit and overfitting, and their relation to predictive power, see sections 4.1.7 on page 114 and 4.2.1 on page 119.

¹⁹Most algorithms minimize the square error between the data and the model. This makes it possible to calculate the optimal parameter values with a simple matrix operation, as described in section 4.2.1 on page 119.

$$y = \left(\frac{x_1 + \alpha_1}{1 + \alpha_1} \right) \cdot \left(\frac{x_2 + \alpha_2}{1 + \alpha_2} \right)^{\gamma_2} \cdot \dots \cdot [\beta_0 + \beta_1 x_1 + \beta_3 x_3 + \dots] \quad (6.1)$$

This model is designed for $x_i \in [0, 1]$ ²⁰.

As in linear regression²¹, there are ordinary linear terms $(\beta_0 + \beta_1 x_1 + \beta_3 x_3 + \dots)$ with parameters β_i and a linear intercept β_0 . In addition, there are also special multiplicative terms $\left(\frac{x_i + \alpha_i}{1 + \alpha_i} \right)$, which have a principal parameter α_i and an optional power parameter γ_i – similar to the shaping parameters s_i from the example above. One may think of the linear terms as an analog signal and the multiplicative terms as an amplitude-shaping function.

Note the differences from a multi-variate linear model with interactions: 1.) The shape parameters α_i and γ_i are not known a priori and therefore are automatically adjusted (optimized) during function fitting, together with the linear coefficients β_i . 2.) The multiplication terms apply to all linear terms (note the parentheses around the linear terms).

The linear terms behave like linear terms in a multi-variate regression. With the coefficients β_i , the slope (or effect strength) of a variable x_i in the model can be set. Negative β_i denote negative effect strength (here, effects that with increasing variable x_i reduce learning levels).

The particular arrangement of the multiplicative terms in the mathematical model stems from the simplifying assumption that if AND-relationships exist for any of the variables, then the simplest case applies that such a variable affects learning in an AND-relationship with all other learning relevant factors. As will be shown in [section 7.2.3 on page 224](#), personal interest in the learning topic enhanced the effect of all other variables relevant to learning performance, such as learning strategy, leadership support etc. Hence personal interest in the topic AND all other factors driving learning (including learning-supportive leadership, learning strategy etc.) drive learning. If all other possible combinations of personal interest with any other variable (or variable group) would have been included in the model, the model's flexibility would have increased significantly, which in turn would have led to a much higher risk of overfitting ([section 4.2.3 on page 128](#)). Hence the design decision here was to create a model that is only slightly more flexible than a linear model. In case the assumption of the simplest case is violated, the effect will be a reduced model fit but not a reduced model robustness. Thus, as long as the overall model fit is at an acceptably high level, this possible reduction of fit due to the violated

²⁰The scale data of this survey has been rescaled to the interval from 0 to 1 – e.g., the coded scale data (1, 2, 3, 4, 5) has been rescaled to 0, 0.25, 0.5, 0.75, 1. See also [section 5.10 on page 161](#).

²¹Compare also the ordinary multi-variate regression model in [equation 4.4 on page 118](#).

assumption poses no bigger problem, since any model will only be an approximation of reality.

Furthermore, the multiplicative terms $\left(\frac{x_i + \alpha_i}{1 + \alpha_i}\right)^{\gamma_i}$ have been designed to have the following properties: the shape parameter α_i can effectively set a lower offset for the effect of x_i . If $\alpha_i = 3$, $\gamma_i = 1$ and x_i is somewhere between 0 and 1 (due to the data scaling during pre-processing), then the multiplicative term will take a value between 0.75 and 1. Thus, with $\alpha_i = 3$, x_i has a multiplicative effect – yet its strength is reduced. For $\alpha_i = 20$, the multiplicative effect of x_i is limited between 0.95 and 1 and begins to become negligible. For simplicity of the model, $\alpha_i = 20$ has been chosen as the (almost) neutral and maximum²² α -parameter setting.

With the shape parameter γ_i between 0 and 1, further non-linearity can be introduced with the effect of a root function on the multiplicative term. γ_i between -1 and 0 allow an inversion of the multiplicative term. $\gamma_i = 1$ is the neutral position leading to no deformation or inversion.

In addition, the effect of all terms can be neutralized in the process of fitting the model by the function fitter via special parameter settings – as shown in table 6.3.

Parameter	Label	Neutral pos.	Effect
β_i	linTerm	0	no linear effect of x_i
α_i	multOffs	20	hardly any multiplicative effect of x_i
γ_i	multExp	1 or -1	no deformation of inversion of the multiplicative term

Table 6.3.: Parameter Labels

In summary, a non-linear yet still fairly simple and thus inflexible parametric mathematical model (eq. 6.1 on the facing page) is fit to the sample data by the BOGER algorithm.

6.2.3. Genetic (Non-linear) Function Fit Algorithm

As in linear multi-variate regression, the parameters in the BOGER mathematical model (equation 6.1 on the preceding page) that lead to an optimal model fit with the data need to be estimated from the sample data. However, due to the non-linearities of the BOGER mathematical model, there is no algebraic way to solve for the fit-optimizing parameter values in order to directly calculate the optimal values – similar to the matrix operation used in ordinary multi-variate regression (see footnote in section 4.2.1 on p. 121).

²²Strictly speaking only $\alpha_i = \infty$ completely neutralizes the multiplicative effect. Therefore somewhat arbitrarily but based on approximation, $\alpha_i = 20$ was chosen as the maximum α -value for the genetic model fit optimizer, which causes multiplicative terms with higher α -values to become deactivated.

The parameter optimization problem instead belongs to the most difficult class of numerical problems – NP-complete problems. Finding the perfectly optimal solution to such problems is not feasible within a reasonable time frame, and thus they need to be solved heuristically (Rivkin, 2000). A heuristic algorithm searches and finds a good solution that is close to the optimum.

Genetic algorithms, mimicking Darwinian evolution involving many generations of mutation and selection, offer robust heuristic optimization results for arbitrary mathematical forms of target functions. Many alternative algorithms that can handle arbitrary target functions, such as simulated annealing or gradient descent, are much less robust, since they can be trapped in *local* minima, while genetic algorithms get close to the *global* minimum with a high likelihood²³.

Genetic algorithms have only one major disadvantage: they are very computationally inefficient, as the following brief description of genetic algorithms illustrates.

The aim of the genetic algorithm is to find the combination of input-variable settings that leads to a good (i.e., close to optimal) result when used in the evaluation of the target function. In the case of BOGER, the genetic algorithm leads to a high and almost maximized model fit based on a good input parameter vector.

1. Based on user-defined starting values, a set of input-independent variable vectors is artificially generated using a number of different techniques, such as perturbation of the starting values, random generation or a combination of the two. This initial set of candidate-independent variable vectors is referred to as generation 1.
2. For each solution candidate, the target function (in this case, the fit between the model and the sample data²⁴) is evaluated.
3. Ranked by the target function values, the top fraction of the solution candidate vectors is selected. This set of best candidates forms the basis of the next generation.
4. To complete the next generation to its full size, further solution candidates are generated by mutation of features from the selected best candidates and/or by random generation.
5. Steps 2 and 4 are repeated until the improvements of the target function from generation to generation stagnate and yield successive improvements below a stopping threshold value or a maximum number of generations is reached.

²³Unlike algebraic solutions for the optimum, most heuristic algorithms provide close to optimal results with a high likelihood but without guarantee. This remaining risk needs to be balanced with other risks of getting false results, such as getting trapped in a local minimum.

²⁴In the case of this study, the target function is the mathematical model's fit against the sample data. The model parameters form the independent variable vector, which is adjusted successively over many generations to lead to a good model fit result. (The sample data, consisting of an independent and dependent variable matrix, remains unchanged.)

Hence a genetic algorithm is effectively a random search procedure with mechanisms that guide the random process towards a good solution. With large sizes of each generation and a large number of generations (compared to the number of parameters that are optimized), the likelihood of finding a solution close to the global optimum becomes large as well.

Yet, not surprisingly, even systematic trial-and-error is a computationally intensive process.

For the BOGER algorithm, the **genoud** genetic optimization algorithm from the **rgenoud** package (v5.4-7) for \mathbb{R} was used (Mebane and Sekhon, 2007). The package features parallelization of the computations to multiple CPUs²⁵, and it permits various adjustments (such as the size of the generations, the fraction of best candidates retained in the next generation, the generation and mutation strategies and the stopping threshold). Depending on the setting, optimizing around 25 parameters takes 5-10 minutes for a single data set with about 300 data rows (on a single CPU core²⁶ of an Intel Pentium D 945 with 3.4 GHz).

In summary, automatically fitting the parameters of the non-linear mathematical model (eq. 6.1 on page 182) in order to minimize the error between model and data requires a non-linear optimization algorithm. BOGER uses the genetic optimization algorithm **genoud** for the task, which robustly finds the global optimum but is computationally intensive.

6.2.4. Overview on Model Building and Robustness Testing

This section gives an overview on how the BOGER algorithm semi-automatically (i.e., algorithmically²⁷) builds the statistical model from a set of candidate terms and how the model (and parameters) are tested for robustness.

As discussed in section 4.1.8 on page 116, by-variable or by-parameter selection strategies are only suitable for non-collinear data and mathematical models with mutually independent terms. Two conditions that do not apply in the case of the survey data and chosen mathematical model: Firstly, the BOGER mathematical model contains mutually *dependent* multiplicative terms (see equation 6.1 on page 182). Secondly, some independent variables in the dataset of this study were found to be collinear (see section 5.12 on page 163).

Thus, as suggested by Guyon et al., BOGER uses by-variable model selection only during a screening stage – in order to reduce the number of independent variables con-

²⁵The computational load can be spread across multiple CPU cores or computers with **genoud** – leading to a much faster completion time. Since BOGER’s model selection procedure is parallelizable with less overhead, this **genoud**’s parallel feature was not used.

²⁶The Intel Pentium D 945 is effectively a two-core variant of a Pentium IV. Each core is an improved version of the Pentium IV.

²⁷For more theory on algorithmic model building, see section 4.1 on page 103.

6.2. Design of the BOGER Algorithm

sidered in the preceding full model search and selection stage. This approach reduces the computational effort by several magnitudes²⁸ (Guyon, 2007; Guyon et al., 2006). In the final full model selection stage, the predictive power for all unique candidate models (i.e., unique combinations of the reduced variables and terms set) needs to be robustly estimated.

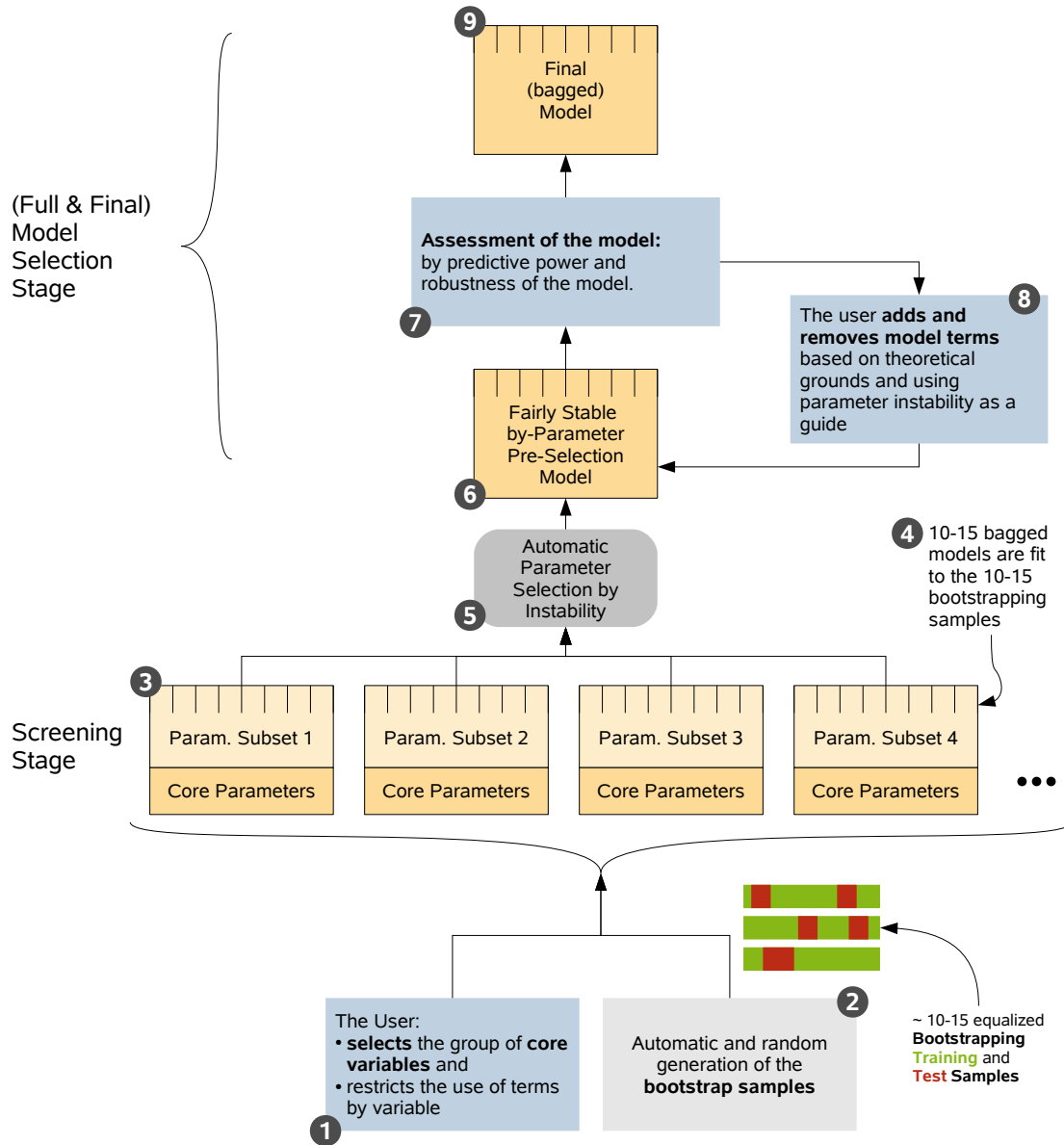


Figure 6.1.: BOGER Algorithm – Flow Chart (Source: Author)

²⁸For example, a full model search with 100 potentially relevant terms requires robust testing and comparison of $1.26 \cdot 10^{30}$ unique combinations of terms (i.e., unique mathematical models).

Hence BOGER allows the user to build the model semi-automatically in two stages:

1. First, in a fully automatic **screening stage**, BOGER selects a set of promising candidate terms from a larger set by using a by-parameter selection strategy – as illustrated by steps 1–6 in figure 6.1 on the preceding page. In contrast to non-robust algorithms such as step-wise regression, which rely on Student’s t-test (Anderson et al., 2000; Chatfield, 1995; Lukacs et al., 2007; Miller, 1984), the statistical significance of each parameter in BOGER is estimated by a more robust bootstrapping method (Breiman, 1992). The user can restrict the use of variables within the model (e.g., a particular variable may enter only as a linear term) – based on theoretical considerations (step 1 in figure 6.1).
2. Finally, in a **model selection stage**, starting with the reduced set of terms with fairly stable parameters, the user manually creates a number of candidate models. BOGER tests the different models so the user may select a good one using the criteria of predictive power and robustness of the entire model (step 7 in figure 6.1). This interactive (partly manual) procedure allows the user to make use of theoretical insights (e.g., about causality²⁹) and steer towards a model that has a high (but not perfect) predictive power, that is robust *and* that makes (causal) sense – all with an acceptable level of computational effort (iterating through steps 6–8 and leading to step 9 in figure 6.1 on the facing page).

6.2.5. The Fully Automatic Screening Stage

Overview of the Screening Stage As mentioned before, in the screening stage (steps 1–5 of figure 6.1 on the preceding page), BOGER uses a by-parameter selection approach to fully automatically reduce the set of candidate variables for further investigation during the final full model selection stage. A term will be included in the final stage if the associated parameter proves to be sufficiently stable as gaged by a robustly estimated instability measure (details follow in text box 6.2.1 on page 190).

As described in section 4.2.2 on page 123, there are a number of challenges with model robustness (overfitting and fit estimation) in cases where the number of variables is relatively high compared to the number of data samples. Therefore BOGER tests individual parameters in ‘smaller’ groups of 32 parameters or less.

First, in **step 1** of figure 6.1 on the facing page, the user chooses a few (~ 15) variables that he or she expects to have the strongest influence. These make up a ‘*core variable*’

²⁹An ideal (i.e., accurate and robust) and fully automatic model building algorithm would be effective at finding the statistical association (correlation) of two variables. However, it could not determine the causality (and causal direction) of the relationship between two variables. But the researcher must build a model of causal relationships if the intention is to explain a process. Thus fully automatic algorithms are mainly useful when prediction rather than causal understanding is the aim – as is the case in medical diagnosis (Sobel, 2005).

group, which together form a stable foundation for all screening models³⁰. In addition, the user can restrict which types of parameters are to be used in conjunction with which variables. For the survey used in this study, the majority of variables were expected to yield only a weak effect. Thus – for this majority – only a linear and a basic multiplicative term (employing only the α parameter) were included in the model.

As a preparation for robustly³¹ estimating model fit and parameter instability, BOGER – in **step 2** of figure 6.1 – generates n_{boot} sets of training and test samples (referred to as ‘*bootstrapping samples*’). The purpose of bootstrapping in this application will become clearer further below.

In order to test all candidate parameters for stability, BOGER – in **step 3** of figure 6.1 on page 186 – creates n_{screen} **screening models** composed of the user-defined core variables (~ 15 variables) plus ~ 5 – 8 non-core variables. Thus each screening model contains the same core group variables in addition to a selection of non-core variables, which are different for every screening model. Enough variable sets are created in order to have each non-core variable in at least one screening model. The variable sets forming screening models will be referred to as ‘*screening iterations*’.

This strategy of splitting up the problem into many smaller screening model selection problems may seem unusual compared to the strategy of many other algorithmic model building methods. In step-wise regression, for example, the algorithm starts with a large model including all potentially relevant variables and then reduces it down to include only statistically significant terms – which leads to robust results only for sample sizes that are large compared to the number of screened variables³². Screening larger numbers of potentially relevant variables leads to the challenges described in section 4.2.2 on page 123. Fitting any model (even an ordinary multi-variate linear model) that has comparatively many variables will give the model too much flexibility, which leads to a high risk of overfitting. Therefore, in BOGER, a divide-and-conquer approach is used in fitting many smaller models to the sample data, which reduces the risk of overfitting to an acceptably low, user-defined level³³. This approach enables BOGER to robustly screen a very large number of variables compared to the sample size.

Next, in **step 4** of figure 6.1 on page 186, the genetic optimizer (section 6.2.3 on

³⁰The intention behind using a core group is to avoid testing any significant parameter with an otherwise very weak and unstable model composed of other insignificant parameters. A core group gives the candidate model a minimum basis of robustness – even if individual core group parameters later prove to be less stable than expected.

³¹For the underlying theory, see section 4.2.4 on page 129.

³²Step-wise regression is a simple algorithmic and fully automatic model building method. It uses the t-test, F-test or another simple estimator as a criterion to determine which terms to include in an ordinary multi-variate regression model. Despite its inclusion in many statistical software packages, the algorithm is only robust with sample sizes that are very large compared to the number of variables. It has received much criticism for this reason (Breiman, 1992; Miller, 1984).

³³By restricting the number of variables in each screening run, the user can choose a suitable compromise between risk of overfitting and computational effort.

page 183) fits each screening model to all training bootstrapping samples (i.e., $n_{\text{screen}} \cdot n_{\text{boot}}$ model fits are performed). The result is a R^2 and R_{abs} model fit estimate (more details in section 6.2.7 on page 195) and a vector³⁴ of the parameter estimates. All estimates are provided for each screening model and bootstrapping sample. In order to improve the average quality of the models (without compromising robustness), 65%³⁵ of the models with the lowest model fit³⁶ are filtered out for the remaining steps.

The parameter estimates over different bootstrapping samples and screening models³⁷ make it possible to robustly³⁸ estimate the instability of each parameter – as defined in detail in text box 6.2.1 on the following page. Parameter instability is used here similar to how Student’s t-test is used to determine statistical significance in ordinary multi-variate regression – yet with the difference that BOGER’s instability measure is more robust against the sampling bias for this application³⁹.

The parameter instability measures the degree of a parameter’s fluctuation when the model is fit to another data sample. The more accurately a parameter estimate reflects the properties of the entire population, the less it will fluctuate across different samples. Since complete data about the entire population is hardly ever available, statistical researchers have developed a number of cross-validation techniques, which split up or *resample* from the collected data sample in order to more robustly estimate statistical quantities by assessing the effect of the sampling bias (see also section 4.2.4 on page 129). The resampling technique used by BOGER is *bootstrapping*⁴⁰, which was chosen because it uses the limited available data very efficiently compared to many other resampling techniques⁴¹.

³⁴Note that the parameters are never averaged for direct use in the mathematical model, since experiments with BOGER have shown that a model based on an average of parameters has less fit with both training and test bootstrapping samples. Instead, the final model will be used as a *bagged* model – i.e., the prediction will be the average of the predictions of multiple models based on different bootstrapping training samples. See also section 4.2.5 on page 134 and 6.2.6 on page 192 for details.

³⁵This is a default setting that can be modified by the user.

³⁶For model filtering, the sum of the model’s training and test set fit is used as criterion.

³⁷For the core variables, parameter estimates for many different screening models have been generated and thus are used for instability estimation. Similarly, for ‘ordinary’ (non-core) variables there can be results from multiple screening models if screening redundancy is used – as described on page 191.

³⁸BOGER uses a bootstrapping approach to make the estimates of model fit and parameter instability robust against the sampling bias – see also section 4.2.4 on page 129.

³⁹BOGER does not simply use the t-test as well, due to various problems with the t-test when used for by-parameter model selection: Student’s t-test relies on a number of strong and, in this case, unfulfilled assumptions. Furthermore, the usefulness of a claim about a non-zero hypothesis has been the subject of recent debate (Anderson et al., 2000; Lee, 1997; Lukacs et al., 2007; Miller, 1984). And most important for model selection, the t-test is not robust enough against the biases caused by the sampling process – as discussed in section 4.2 on page 119. There is a significant risk that the results of the t-test do not generalize to the entire population.

⁴⁰More theoretical details on *bootstrapping* were presented in text box 4.2.1 on page 132.

⁴¹Ordinary cross-validation, e.g., simply splits the dataset once into a training and a test set, doubling the sample size requirement.

Text Box 6.2.1 BOGER's Parameter Instability Measure

For assessing parameter instability, BOGER generates multiple estimates for a single parameter by fitting the model to different bootstrapping training samples (see appendix section A.6.1 on page 305). The result is a vector of parameter estimates for each parameter.

Parameter instability measures a parameter's fluctuation across different models built on different bootstrapping samples. In some cases, the combination of terms not connected to the parameter may also differ (e.g., when the parameter is included in multiple screening models). The fluctuation is based on a measure similar to the standard deviation: the *single-sided square error (SSSE)* – defined in equation 6.2. This design of the measure is in principle similar to Student's t-test⁴² or an inverse signal-to-noise ratio (with normalization).

$$SSSE = \sqrt{\frac{1}{n} \sum_{f_i} \left[(\text{median}(\text{parameter}_{f_i}) - \text{parameter}_{f_i})^2 \right]} \quad (6.2)$$

where parameter_{f_i} is a filtered parameter vector consisting only of those parameter values that are on the weak side of the median parameter. For a linear parameter with median 0.8 and an original parameter estimate vector $\{0.3, 0.8, 0.5, 1.3, 0.9\}$, the filtered vector is $\{0.3, 0.5\}$ and thus contains only all values that are weak with respect to the median strength of the parameter.

The parameter-instability measure is the single-sided square error normalized with the mean of the variable corresponding to the parameter and the parameter's strength – as shown in equation 6.3. This normalization ensures that smaller parameters, acting on high variable means, get a similar stability score as higher parameter values, acting on variables with smaller means. In addition, the fluctuation is set in relation to the strength of the parameter⁴³.

$$\begin{aligned} \text{parameter instability} &= \frac{SSSE \cdot \text{variableMean}}{\text{absoluteParameterStrength}} \\ &= \frac{SSSE \cdot \text{variableMean}}{\text{abs}(\text{mean}(\text{parameter}_{f_i}) - \text{neutralPos}_{\text{parameter } f_i})} \end{aligned} \quad (6.3)$$

The result of the normalization is an instability measure that allows robust and direct comparison of the stability not only of linear parameters (α_i) but also of multiplicative parameters (β_i and γ_i – see eq. 6.1 on page 182).

This instability measure was developed on the basis of many trials with the BOGER algorithm, using the robust model fit measure as a criterion and graphically visualized distributions for understanding the quality of individual parameters. Yet despite these efforts, as discussed earlier in this section and section 4.1.8 on page 116, this parameter-instability measure (like any other parameter criterion) is suitable only for model screening rather than full algorithmic model selection.

Step 5 of figure 6.1 on page 186 concludes the screening stage by selecting the terms of the mathematical model that will enter the model selection stage. This selection is fully automatic and uses the parameter instability as its criterion. Only those parameters whose instability⁴⁴ is below a user-defined threshold will be activated for the full model selection stage. BOGER chooses the terms of mathematical model for the full model selection stage based on the activation state of the parameters – in step 6 of figure 6.1 on page 186.

The chosen parameter-instability threshold should be strict enough to reduce the variables down to a number that can be handled in an interactive full model selection stage and allow a further reduction of variables. For the dataset of this study, the screening stage was ended with ~ 28 -35 variables and further reduced down to ~ 15 -25.

In summary, the result of the BOGER screening stage is a model with a significantly reduced number of variables. The selection of the variables (and terms) is based on a by-parameter selection approach. The algorithm is designed to also function robustly when the number of potentially relevant screening variables is large compared to the sample size. Yet, given the weaknesses of by-parameter model selection (see section 4.1.8 on page 116), the resulting model is only preliminary and just a basis for a final model selection stage, which is described in section 6.2.6 on the following page.

Details of the Screening Stage The preceding pages have given a general overview of the screening stage. The presentation was detailed enough for a general understanding of the algorithm. The following pages will present two implementation details that make the screening process more robust. Readers who are not interested in these details may skip to section 6.2.6 on the next page.

The first detail is *frequency equalized bootstrapping sample generation*. In order to make use of the original sample in the most data-efficient manner, the bootstrapping samples are generated in such a way that they are random but frequency equalized – i.e., each data point occurs equally frequently in the bootstrapping samples, but in which bootstrapping sample a data point is included is still random. For more details, see appendix section A.6.1 on page 305.

The second detail is *screening redundancy*. In an attempt to mitigate the weaknesses of

⁴²Student's t-value is defined as $t = \frac{b_i}{s_i}$, where b_i is the regression parameter estimate and s_i is the standard error corresponding to the parameter. The t-test then involves looking up the corresponding confidence probability value in a distribution for the t-test, which is based on a number of assumptions, such as normality (Backhaus et al., 2006, p. 74).

⁴³A small fluctuation towards the weak end may be a serious problem for a parameter with a weak effect, while it is not for a strong parameter. Thus strong parameters are not penalized by fluctuations with a medium magnitude.

⁴⁴More precisely: BOGER will estimate multiple instability values for some parameters – either if the parameter is a core parameter or if more screening runs are performed than absolutely necessary (“screening redundancy”. This is described later on p. 191). BOGER then uses the average of the 35% worst instability values of a parameter as the criterion for parameter selection.

by-parameter model selection, BOGER can be run with a screening modeling redundancy: more variable sets and thus screening models are generated for testing each variable in multiple and different screening models. Using this approach, the composition and peculiarities of individual screening models become less important because the non-core parameters are also included in more than one screening model – at the expense of having to fit 2 to n times as many screening models (where $(n-1)$ is the number of redundancies). With the screening redundancy, the user adjusts the number of different screening models in which each variable is included. The result is multiple parameter instability values – of which the worst 35% are used as the parameter selection criteria in by-parameter selection. Furthermore, this procedure provides additional “samples” of the parameter estimate, which is less dependent on (and thus biased by) a particular model. Despite an improvement in robustness, this by-parameter selection still cannot compete with a full model search in terms of robustness (see section 4.1.8 on page 116).

Hence both design details improve BOGER’s accuracy and robustness.

6.2.6. The Final Stage of Model Selection

The aim of the final model-selection stage is to validate and refine the results of the screening stage. There are two principal differences between the two stages: 1.) the selection of the final model is based on a robust estimate of the model fit rather than parameter instability; and 2.) the process is interactive and partly manual, with the possibility for intelligent user input on, e.g., causalities (see section 4.1.5 on page 109).

The final model selection stage starts with the preliminary model (**step 6** in figure 6.1 on page 186), which was created fully automatically with a by-parameter selection approach during the screening stage. This model is then iteratively refined without a predefined search path:

1. For increased accuracy and robustness, the current model is fit to an enlarged⁴⁵ set of bootstrapping samples – similar to how the screening models were fit in the screening phase (**step 6** in figure 6.1 on page 186). Like in the screening phase, the result is a collection of fitted models – one model for each bootstrapping sample.
2. Next, the quality of the model as a whole is assessed (**step 7** in figure 6.1 on page 186). For each individual bootstrapping model, the model fit with the training data and the fit with the cross-validation test data is calculated. As in the screening stage, the average model quality is increased by removing the lower fraction of models based on the sum of a model’s training and test model fit. In contrast to the

⁴⁵For computational efficiency during the lengthy screening stage, most estimations are performed at a normal accuracy. Yet the final stage achieves increased accuracy and robustness by using more bootstrapping samples and a higher screening redundancy.

screening phase, the criterion is slightly stricter, keeping only the best 25%⁴⁶ of the models for inclusion in the bagged model. Empirical evidence for the effectiveness of this model filtering strategy can be found in appendix section A.6.5 on page 314.

The result is a) a selection of the best models, as well as b) a robust overall estimate for predictive power (of this best sub-set). The predictive-power estimate is measured by R_{abs} and is provided in the form of a model fit distribution as well as an average based on the filtered set of bootstrapping models – for details, see section 6.2.7 on page 195. Using this predictive-power estimate as criterion, the researcher decides whether the model needs further improvement by adding and removing terms and variables. When starting with the final model selection stage, the researcher will always try to further improve the model.

3. Based on theoretical insights and the updated parameter instability as a tentative suggestion⁴⁷, the researcher may add and/or remove terms (i.e., parameters) to/from the mathematical model⁴⁸ (**step 8** in figure 6.1 on page 186).
4. Next, **step 6** of figure 6.1 on page 186 (or step 1 in this list) is repeated and the modified model is fitted to the bootstrapping samples. With this step another iteration of the final model selection process has begun. As for the previous model, the next steps will be the model fit assessment (step 7) and further modification of the model (step 8).

The researcher iterates through this process of assessing and modifying models until there are no (or only marginal) improvements of the predictive power compared to the best model tried so far. At that point, the user can use BOGER’s history logging function, which stores all tested models and their performance assessment, in order to select the best one to become the final model (**step 9**) – see section 6.2.8 on page 197.

The result of the entire semi-automatic process is the robust selection of a *good* model⁴⁹ by taking all available information (data and theoretical considerations) into account. The process is performed within an acceptable time frame and starts on a solid foundation: an already improved model from the screening stage. However, since there is no prescribed

⁴⁶This fraction is user-definable (with the parameter `bootSampleFilter.frac`) and should be set in such a way that the final bag still contains sufficiently many individual models. Since the number of bootstrapping samples and thus also the number of models is higher in the final stage, depending on the settings, the total number of models in the final bag may still be larger than in the screening stage.

⁴⁷Note that parameter instability is not used as a criterion in the final model selection stage. Instead the parameter instability, which was updated during the previous model fitting, *may* be used as a tentative *suggestion* for creating the next model.

⁴⁸Terms that were not included in the model from the screening stage may also be reactivated for further investigation. A *history* function saves all models and allows the user to go back to any older model (see section 6.2.8 on page 197).

⁴⁹The result from BOGER is a good *and* robust model selection – instead of a mathematically perfect optimum, which is non-robust.

search path with full coverage of the vast amount of possible models, there is no guarantee of finding the mathematically optimal model (the *best* model)⁵⁰. Following the argument of Breiman (2001b) and Lukacs et al. (2007), I claim that a mathematically optimal model is not necessarily of much greater value, since the mathematically optimal result is based on inaccurate and at least weakly biased estimators. Hence the mathematically optimal model is also only a ‘good’ model with respect to the overall process.

As outlined in chapter 4 on page 101, there are a number of challenges in statistical inference, most notably causality, model selection (including fitting), overfitting, biased model fit estimators and variable vs. full model selection. BOGER was designed to address and mitigate all of the risks resulting from these challenges (except for causality⁵¹) instead of focusing on only some parts of the process and thus on only one type of risk. Hence the risk of getting spurious results is non-zero. Yet, I argue that the BOGER design represents a suitable compromise among the algorithm’s effectiveness in mitigating the various risks of delivering spurious results, time and data efficiency.

For actual prediction, the predictions of the best individual bootstrapping models are averaged. The weaker bootstrapping models are excluded from this average by the combined training and test model fit criterion (see step 7 above). For this approach, Breiman (1996a) coined the term bagging, which is short for “*bootstrap aggregating*”. Other authors refer to this technique as “ensemble building” (Chatfield, 1995; Strobl et al., 2007; Yuan and Yang, 2005) – see also section 4.2.5 on page 134. This approach mimics a board of experts who collaborate on a joint (and thus inter-subjective) expert prediction.

The much acclaimed strength of bagging is its robustness and its flexibility with regard to the model type (Breiman, 2001b; Chatfield, 1995; Strobl et al., 2007). Bagging leads to robust predictions, since it is built on the basis of multiple models, which are fit to different (bootstrapping) data sets. In addition, it is flexible enough to accommodate models with mutually dependent terms (section 4.1.8 on page 116) that would not allow for a simple averaging of model parameters. One can even assess the model’s local robustness over specific parts (intervals) of the independent variable space (see also figure 4.2 on page 127) by measuring changes in the divergence of the individual model predictions over different parts of the variable space.

A downside of bagged models is that they lead to black-box models that the researcher cannot easily interpret. The bagged BOGER model also must be seen as a black-box model, since the parameters of the individual models have very little meaning. As discussed before in this section and in section 4.1.8 on page 116, averaging the parameters can only serve as a rough indication for the importance of a particular term in the bagged

⁵⁰Most statistical model fitting algorithms (ordinary regression, logistic regression, SEM etc.) achieve a mathematically optimal model fit but have no systematic model selection process, let alone a mathematically optimal model selection procedure. See also section 4.1.6 on page 111.

⁵¹Dealing properly with causality either requires strong (and possibly false) assumptions using other (qualitative) evidence or suitable experimental designs, as described in section 4.1.5 on page 109.

model. Therefore alternative model inspection techniques need to be used, ones that do not involve drawing inferences from the internal mathematical structure of the model. A suitable model inspection technique is Breiman’s variable importance measure – see [section 7.1.1 on page 206](#).

Note that the result of the BOGER algorithm is a robust expected value model. If the true data-generating process is stochastic, the prediction of a bagged BOGER model is the expected value for a particular setting of independent variables, rather than a model of the full multi-variate distribution completely describing the stochastic process – see also [section 4.1.4 on page 108](#).

The final model selection stage, in summary, performs a full model search in order to avoid the challenges and restrictions of by-variable (i.e., by-parameter) model selection. The result of the interactive iteration process is a good “bagged” expected value model consisting of multiple yet selected (filtered) bootstrapping models with no or little overfitting.

6.2.7. Robust Model Fit Estimation in BOGER

The previous sections described the model building steps performed by BOGER. As already mentioned in [section 4.2.2 on page 123](#), a robust model fit estimator is a principal component of a robust model selection algorithm. This section will describe in further detail the estimator used by BOGER. This section is relevant for readers who wish to inspect and evaluate the quality of BOGER’s design and for readers who need to understand the meaning of BOGER’s model fit measures in order to assess model quality in detail.

[Section 4.2.4 on page 129](#) presented various modern model fit estimation methods. Evidence from various studies showed that bias-reducing estimators were more robust and accurate than bias-correcting estimators – such as simple bias-correcting estimators for Mallows’ C_P and Akaike’s information criterion (AIC).

For bagging algorithms such as BOGER or Random Forest, which bundle and average predictions of multiple individual models, Breiman’s *Predictive Error (PE) Cross-Validation Estimate* (Breiman, 1996b) showed good accuracy and low bias (Breiman and Spector, 1992; Strobl et al., 2007; Zhang, 1993). In addition, it makes very efficient use of the limited sample data, since models are fitted to all parts of the sample data instead of an a priori separation into a training and a validation dataset – as in simple dataset separation cross-validation (see [section 4.2.4 on page 129](#)). The high computational effort connected to bootstrapping and building a bagged model is the only disadvantage cited (see text box [4.2.1 on page 132](#)).

Analogous to Breiman’s approach, BOGER calculates the ordinary R^2 and R_{abs} model fit with the training and test datasets for each bootstrapping model. The result is two

predictive error estimates (as defined in section 4.1.7 on page 114) for each bootstrapping sample (i.e., for each iteration):

- The **training data model fit** – which is an ordinary R^2 and R_{abs} estimate (section 4.2.1 on page 119) based on the training bootstrapping sample and the corresponding fitted BOGER model.
- The **test data model fit** – based on the corresponding test bootstrapping sample and fitted BOGER model. Breiman and Strobl et al. refer to the test data model fit as an *out-of-bag model fit estimate* (Breiman, 2001a; Strobl et al., 2007).

Since the bootstrapping test samples were not used for model fitting, the test data model fit – i.e., the out-of-bag model fit – serves as an independent validation of the predictive error and allows detection of biases related to the fit process, such as biases due to overfitting (see section 4.2.2 on page 123).

As discussed in section 4.2.2 on page 123 and 4.2.1 on page 119, estimates for the model fit that are based on the training data may be severely biased due to an interaction of the estimation biases and aggressive fit optimization. Nevertheless, the training data model fit should be considered in addition to the test data model fit. Commonly, the user chooses fewer bootstrapping test samples than training samples. In those cases, the test data model fit estimates have a higher risk of being subject to sampling biases (see section 4.2.1 on page 119), given the smaller underlying sample size. See also the empirical comparison in appendix section A.6.5 on page 314. Hence both model fit measures will be biased, but in different ways – which can be exploited to estimate the true predictive power:

A model has a high predictive power if the following two conditions hold:

1. The training and test model fits are both **high**; and
2. the training and test model fits are **close to each other**, indicating that there is little overfitting and thus the model is generalizable beyond the training sample (*true* predictive power).

BOGER calculates⁵² model fit with two measures: R^2 and R_{abs} . R^2 is very common in statistics and penalizes larger deviations from an expected value model more than smaller deviation due to its square function. Thus individual but strong outliers will have a strong effect on the R^2 estimate. Therefore R^2 is a sensible choice for statistical processes that have little or no noise compared to the predicted value (i.e., high signal-to-noise ratio) and thus few outliers. However, for noisy statistical processes, R_{abs} is more robust against individual outliers, since it is based on the simple sum of the (absolute) errors. Empirical

⁵²BOGER uses the simple estimators for R^2 and R_{abs} as presented in a mathematical form in section 4.2.1 on page 119.

evidence for this argument can be found in appendix section [A.6.5 on page 314](#). Since the data from this study is noisy, BOGER is used with R_{abs} as the primary model fit measure that is used for model filtering and predictive power assessment.

So far the discussion has focused on the individual models that constitute the bagged BOGER model. Assessing the individual models is relevant and useful during the iterative model building process. Yet in the end, the predictive power of the final model needs to be assessed. However there is no independent test dataset for the bagged model, since the test samples of one model are the training samples for another model and vice versa. (This is the property that makes BOGER data efficient.)

Therefore BOGER relies on a model filtering strategy that uses only a fraction of the bootstrapping models in the bag – based on the sum of training and test model fit as criteria. Test runs with BOGER indicated that these ‘best’ bootstrapping models have a training and test model fit close to each other. Hence at the end of the iterative model building and filtering process, the bagged BOGER model consists of a number of ‘best’ bootstrapping models that show little overfitting or other biases related to the fit process.

Since averaging many non-overfitting models does not involve a new fit optimization process, no interaction with model fit estimates is possible, and thus the bagged model will not be overfit either. Therefore an ordinary R^2 and R_{abs} model fit estimate based on all available (and thus many) samples is an acceptable model fit measure for the bagged model, since any biases from overfitting will be as low as in the individual bootstrapping model. In addition, the sample bias of the ordinary model fit will be reduced as well – due to the larger sample that the R_{abs} model fit is calculated against.

In summary, the true predictive power of the bagged BOGER model can be assessed robustly by assessing the predictive power of the individual models on the training and test data as well as assessing the simple model fit of the bagged model against the total sample. With these different and redundant assessments of predictive power, the BOGER model’s generalizability to the entire population or into the future can be judged including an assessment of the robustness of this judgment (i.e., the magnitude of the various biases).

6.2.8. Implementation in \mathbb{R}

The BOGER algorithm was implemented in a specialized statistical programming language called \mathbb{R} , which is an open-source system popular in research on statistics and bio-informatics ([R Development Core Team, 2007](#)).

This choice of statistical software package (instead of SPSS, STATA or similar) allowed for much flexibility in algorithm implementation, visualization and analysis, which was much needed to deal with the challenges described in the previous sections of this chapter.

The BOGER algorithm (without data preparation or analysis) was implemented in about 4,000 lines of \mathbb{R} code (that is equivalent to about 200 pages with single-spaced

text). The BOGER implementation features an internal logging mechanism, a command-line interface for user-interaction and the parallelization of the algorithm for execution on multiple CPU cores or computers.

More details on the implementation and \mathbb{R} can be found in appendix section [A.6.2 on page 307](#).

6.3. Design Requirement Validation & Performance of BOGER

Section [6.2.1 on page 179](#) described the design goals for BOGER. To validate these requirements, this section describes BOGER's model fitting performance and summarizes its features that meet the requirements.

Since the performance validation is based on the survey data from this study, it also serves as an assessment of the quality of the statistical model, which is the foundation of the accuracy and robustness of the statistical results presented in section [7.2.1 on page 217](#).

6.3.1. Model Selection Progress

As discussed before, the statistical model building challenge for the survey dataset begins with a fairly large set of variables and associated parameters:

- 139 original survey variables with 284 parameters (see section [5.3 on page 141](#));
- plus 154 interactions with 291 parameters (automatically added based on a correlation analysis⁵³);
- summing up to **293 variables total** (survey variables + interactions) with **575 parameters in total**.

Based on theoretical grounds (the obvious lack of a causal effect on learning – see section [4.1.5 on page 109](#)), some of these variables were disabled a priori and were not included in the search for a suitable statistical model (see step 1 in fig. [6.1 on page 186](#)). Therefore only **502 parameters** (229 parameters from interactions) were fed to the BOGER algorithm.

Given an effective sample size⁵⁴ of $n = 292$, this number of parameters is large enough to lead to significant risks of overfitting (see sections [4.2.2 on page 123](#) and [A.6.3 on page 310](#)). Therefore, in order to improve robustness, a model with a much smaller

⁵³For the 139 survey variables, a full correlation analysis of all possible combinations of two variables is performed. Any combination of two variables with a correlation between 0.32 and 0.98 is added to the dataset as a simple linear correlation (e.g., variable 1 · variable 2). Eleven additional 'custom' interactions were also added based on theoretical considerations.

⁵⁴See the discussion of data filtering in section [A.5.1 on page 302](#).

number of parameters needs to be selected systematically – as described in section 6.2.4 on page 185.

After BOGER’s screening procedure (see section 6.2.5 on page 187) was implemented and debugged, a high-quality screening run was started and ran for ~ 6.5 days on the two CPU cores of an Intel Pentium D 945 with 3.4 GHz.

In the beginning, as a preparation of the screening stage (step 3 in fig. 6.1 on page 186), 32 screening models were automatically generated, each consisting of 25-35 parameters. Of these 25-35 parameters⁵⁵, 12 were contained in all screening models as core parameters (from five core variables). Each screening model was fit to 14 bootstrapping samples. A screening redundancy (see p. 191) of five leads to 192 screening iterations (i.e., $6 \cdot 32 = 192$ partly overlapping screening models). Hence, in total, the genetic optimization algorithm performed $192 \cdot 14 = 2688$ model fit operations (with 192 different math models and 14 different bootstrapping samples). Thus a computational effort of approximately 6.5 days leads to an average model fitting time of $6.5 \cdot 24 \cdot 60 / 2688 \cdot 2 = 7.0$ minutes per model and CPU core.

The result of this parameter screening was a parameter pre-selection (based on the parameter instability criterion) as the basis for the first interactive model, with 31 parameters.

This preliminary model was then iteratively and interactively refined in more than 32 iterations (steps 6 – 8 in fig. 6.1 on page 186) by using the predictive power estimate of the model as model optimization criterion – see also section 6.2.6 on page 192 on the final stage of full model selection.

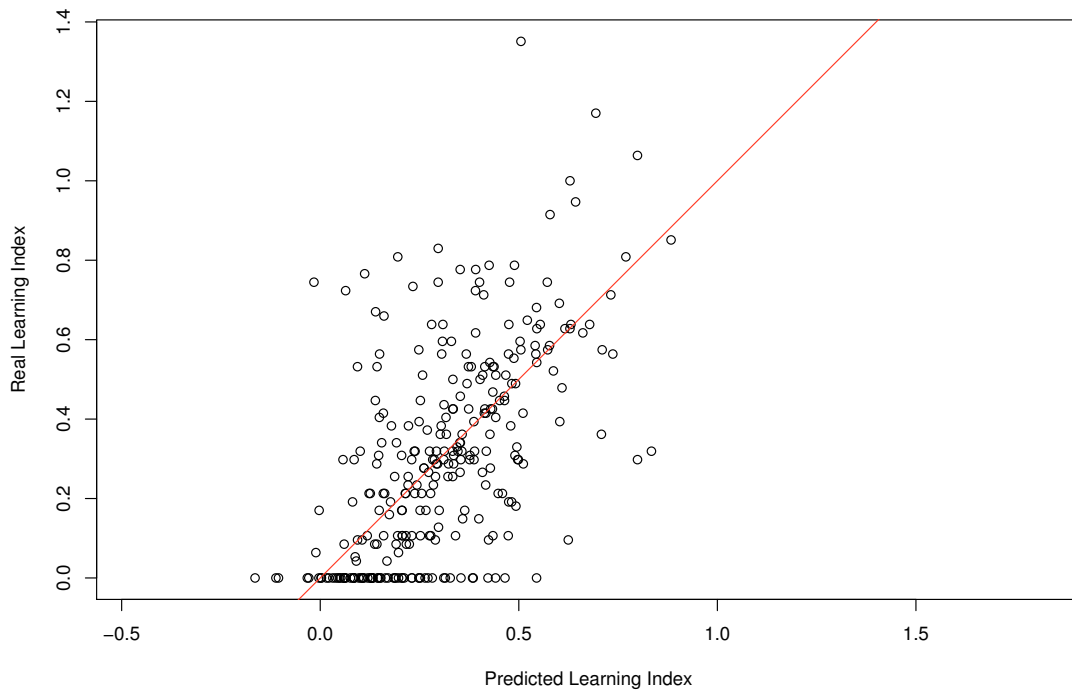
The resulting **final model** consists of ‘only’ **22 parameters**. For quality improvement, the final **bagged model** includes only the **top 25%** of all individual bootstrapping model fits for the same 22-parameter model structure – see section 6.2.6 on page 192). Compared to the preliminary pre-selection model from the screening stage, the final model has improved predictive power – as the model fit results in the following section will show.

6.3.2. Model Fit & Predictive Power Estimate

Figure 6.2 on the next page shows the real learning index data vs. the predicted learning index, giving a graphical impression of how well the BOGER model fits the data. If there was no randomness (noise) in the process, all points would lie on the red line for a perfect model.

Appendix section A.6.4 on page 311 contains an analysis of the residuals, which are approximately normally distributed – indicating that the non-linearities in the data have been treated by the BOGER model with sufficient accuracy.

⁵⁵The automatic generation of screening models is based on variables, not parameters, which is why the number of parameters per screening model varies.

Figure 6.2.: A Graphical Impression of **Model Fit: Real vs. Predicted** Learning Index

	Training Fit	Test Fit
R^2 (bagged model)	34.5 %	-
R^2 (indiv. filtered models)	32.6 %	36.7 %
R_{abs} (bagged model)	28.4 %	-
R_{abs} (indiv. filtered models)	28.1 %	26.1 %

Table 6.4.: Model Fit Summary – Solid Results

More precisely, table 6.4 quantifies the model fit using two different measures with different strengths and weaknesses: R^2 and R_{abs} ⁵⁶.

The fit results are further distinguished by *Training Fit* and *Test Fit*, which indicates whether R^2 or R_{abs} was estimated using the sample used during model fitting (training data) or using the test dataset⁵⁷. Since the test data is not used for model fitting, the

⁵⁶ R_{abs} is a model fit measure very similar to R^2 , with the only exception that instead of using the sum of the squared errors it uses the sum of the absolute value of the errors – see also section 4.2.1 on page 119

⁵⁷As described in section 6.2.5 on page 187, the test data is the test data from bootstrapping and thus is different for each individual model in the bag.

R^2 and R_{abs} values based on the test data are expected to be a less biased estimate of predictive power than the corresponding estimates based on the training data (see section 6.2.7 on page 195).

Moreover, the fit results are shown separately for the overall *bagged model* and as an average fit of the *individual filtered models*. As discussed in section 6.2.6 on page 192, the final model, which is used for predictions and inference, is a single bagged model consisting of the ‘individual filtered models’, i.e., the top 25% of all individual bootstrapping models with the same final model structure. For the individual filtered models, the model fit estimates listed in table 6.4 on the preceding page are calculated for each model individually and then averaged over all individual models. In contrast, the bagged model is only a single model and thus only the simple R^2 and R_{abs} estimates based on the total data can be calculated (see section 6.2.7 on page 195).

The first impression of the model fit results in table 6.4 on the preceding page with values above 25% is good⁵⁸ – especially when compared to the results from the other tested algorithms (section 6.1.3 on page 175). One detail, however, may be surprising: the R^2 estimate for the individual models and the test data is higher than the corresponding R^2 estimate for the training data. As discussed in section 6.2.7 on page 195 and empirically supported in appendix section A.6.5 on page 314, R^2 is less robust to noise in the data than R_{abs} . This is why R_{abs} instead of R^2 was used for the model selection with BOGER and why R_{abs} will be used as a basis of the following assessment of the quality of the statistical model.

Considering primarily the estimate for R_{abs} in table 6.4 on the preceding page, it can be concluded that:

- The BOGER model only very slightly overfits: the training fit is only $\sim 2\%$ higher than the test data fit. This is a substantial improvement compared to any of the existing alternative algorithms, which all strongly overfit – see section 6.1.3 on page 175.
- The low level of overfit further implies that the simple R_{abs} model fit estimate for the bagged model fit will have a low bias due to overfitting. It can thus be interpreted as a robust and fairly accurate estimate of the true predictive power of the final bagged model.
- Therefore the predictive power of the final bagged model with approximately $R_{\text{abs}} = 28\%$ (or in terms of $R^2 = 34\%$) is fairly high – given that learning is a complex process with many latent and possibly overlooked variables, which makes it similar to typical social science applications (see section 4.2.2 on page 123). Judging from

⁵⁸Considering the level of noise in the data.

the variance of the training model fit distributions in figure A.19 on page 316, these bagged model fit estimates should be accurate within a tolerance of at most $\pm 5\%$.

- For the survey data, the model prediction performance is more than twice as good as the performance of the best existing alternative algorithm (cForest) – see section 6.1.3 on page 175.
- The preliminary pre-selection model from the screening stage has a training model fit of $R^2 = 32.4\%$ and a test fit of $R^2 = 25.9\%$. Thus the refined final bagged model has a substantially better test model fit than the pre-selection model and overfits less. Overall the predictive power of the final model is better than the fully automatically generated pre-selection model.

An empirical analysis of the model fit measures – by comparing the distributions of individual bootstrapping model fits in appendix section A.6.5 on page 314 – empirically supports the robustness of the above used assessment methods and results.

In summary, given the complex and stochastic nature of the learning process, the model fit is fully satisfactory. The detailed results confirm the robustness of this assessment.

6.3.3. Design Requirement Validation Summary

With the empirical results from the two following sections, it can be concluded that the BOGER algorithm meets all design requirements posed in section 6.2.1 on page 179:

- The mathematical model at the core of BOGER (eq. 6.1 on page 182) is a **multi-variate model**, which can deal with the **collinearities** in the survey data and **restricts the model flexibility** for increased robustness to overfitting.
- BOGER supports **systematic full model selection** in order to reduce the number of variables in the final model, with a fully automatic screening stage followed by an interactive full model selection stage.
- The combination of training and test model fits allows a **robust assessment** of the **predictive power** and robustness of the model.
- The BOGER results for the survey data show only a **small overfitting bias** (section 6.1.3 on page 175).
- The sample data is used with **high data efficiency** by full use of the information in the metric scales (by the use of a native regression algorithm instead of a classification algorithm in the background) and by the use of a data-efficient bootstrapping algorithm.

- With approximately $R_{\text{abs}} = 28\%$ and $R^2 = 34\%$, a high explained variance, i.e., **a high level of predictive power**, is attained for the survey data (section [6.1.3 on page 175](#)). The predictive power is more than **twice as high than the best alternative** algorithm (cForest) – compare sections [6.3.2 on page 199](#) and [6.1.3 on page 175](#).

7. Statistical Results and their Interpretation

Chapter Contents

7.1. Result Interpretation Procedures	206
7.1.1. Interpretation Criterion – Variable Importance	206
7.1.2. Interpreting Interactions	212
7.1.3. Model Shape Graphics by Variable	213
7.2. Overall Statistical Results	217
7.2.1. Survey Questions of the Model Variables	217
7.2.2. Variable Rankings and fitted Model Parameters	219
7.2.3. BOGER Model Parameter Results	224
7.3. By Variable Results and Interpretation	226
7.3.1. Learning Strategy Profile	226
7.3.2. Leadership Effect	231
7.3.3. Personal Interest	238
7.3.4. Personal Working History Variable Group	240
7.3.5. Learning Barriers Variable Group	243
7.3.6. Epistemological Beliefs about Learning	247
7.3.7. Task Type Variable Group	250
7.3.8. Description Detail-Level of Procedures	252
7.3.9. Number of Seminars	254
7.3.10. Job Closure	255
7.3.11. Openness to New Experiences (Big Five)	257
7.3.12. Task Difficulty	257
7.3.13. Fault Culture	258
7.3.14. Surprisingly insignificant Factors (Non-Factors)	260

7.1. Result Interpretation Procedures

As already discussed in section 6.2.6 on page 192, the final bagged BOGER model has one major disadvantage: In contrast to an ordinary regression model, for example, the parameter estimates cannot be used directly for inference.

Hence to draw inferences from an opaque bagged model that defies direct inspection, the following sub-sections will present alternative inspection and model structure analysis techniques in the form of a variable-importance measure and uni-variate model-shape graphs.

These inspection techniques are relevant for a detailed interpretation of the results. Readers looking for a summary of the statistical results in the context of existing theory may skip to the implications section 8.1.2 on page 267.

7.1.1. Interpretation Criterion – Variable Importance

This section presents the *permutation variable-importance measure* as proposed by Breiman (2001a), which allows the user to quantify the importance of a particular variable without the need to inspect the model's structure. Thus the variable-importance measure also works for models that appear as a black box and have an arbitrary complex structure. Moreover, with this measure, the model's independent variables can be ranked according to the strength of their effect on the outcome variable (in this case, the learning index) – independently from how the variable occurs in the model (even if the variable occurs in different configurations and multiple times in the model).

The Need for a Variable-Importance Measure and Existing Alternatives In ordinary multivariate linear regression, a comparison of the regression coefficients can assess the strength of each individual variable's effect on the model, as long as the data has been standardized (z-transformed) to equal mean and variance (Backhaus et al., 2006, p. 54) and the variables are not collinear¹. A few additional rules apply when interactions are used – as will be discussed in section 7.1.2 on page 212.

¹Collinearity, i.e., a substantial correlation of two variables A and B, makes a simple interpretation of regression coefficients b_A and b_B for A and B impossible. The additive structure of ordinary linear regression models suggests that increasing A alone will increase the dependent variable y at a slope of b_A . However, if B is highly correlated with A, B will also increase as A increases. This coupled increase of B multiplied by b_B needs to be added to the direct effect of A on y . Using a multiplicative interaction AB may alleviate problem. See also sections 7.1.2 on page 212 and 4.1.8 on page 117.

It is not as easy to directly assess the relative importance of the variables and their interaction using the bagged BOGER model, with its multiplicative terms and non-additive structure.

Existing Variable-Importance Measures Since many model building algorithms in machine learning are similarly opaque to direct inspection, various variable importance measures have been proposed in the machine learning literature – e.g.:

- Breiman’s Permutation Variable Importance (Breiman, 2001a; Strobl et al., 2007)
- Laan’s Absolute Prediction Loss Variable Importance (van der Laan, 2006)
- The Gini Coefficient from Information Theory (Breiman, 2001a; Strobl et al., 2007)
- Pratt’s Variable Importance Measure for Linear Regression (Thomas et al., 2007)

As Strobl et al. (2007) report, the Gini coefficient shows a very strong bias – at least for Breiman’s method of estimation. It has thus not been considered for this study. Pratt’s variable-importance measure is restricted to additive linear models. That leaves Breiman’s Permutation Measure and Laan’s Absolute Prediction Loss Variable Importance for use in this application.

Breiman’s Permutation Variable Importance is suitable It turns out that Breiman’s permutation variable importance is effectively a measure for the relative prediction loss, and thus it is very similar to Laan’s absolute prediction loss measure (van der Laan, 2006, p. 4) – except for the estimation methods. It is noteworthy that Strobl et al. (2007) demonstrates biased results from Breiman’s permutation variable-importance measure using simulated data. However, this bias stems from a bias in tree node split selection, i.e., from a bias in automatic variable selection during model building and fitting. Thus no biases have been reported for the permutation measure itself but rather for the random forest algorithm in certain circumstances.

The results of this study will be interpreted with Breiman’s permutation variable importance. This method was chosen instead of attempting to interpret individual parameters for the following reasons:

- **Complex interactions of the mathematical terms** in the model (e.g., the multiplicative and linear terms) are covered.
- **Collinearity** can be dealt with – see below.
- **Extensive coverage and testing** of Breiman’s permutation measure (Archer and Kimes, 2008; Strobl et al., 2007) has taken place. Other researchers have begun to appreciate and use this very flexible variable-importance measure in practical

applications (often in conjunction with the use of random forest), as these examples show: Archer and Kimes (2008); Li et al. (2006); Thuiller et al. (2006).

- The **robustness** or **statistical significance** of the variable-importance result can be assessed by **inspection of the distribution of variable-importance estimates** over different bootstrapping model fits. Using Breiman's permutation technique (see below), a variable-importance sample is estimated for each bootstrapping model contained in the final bagged model (see section 6.2.6 on page 192). These samples are then plotted in a frequency graph to visualize the distribution of the variable importance – as shown in fig. 7.4 on page 222. This distribution is effectively the variation of the variable-importance due to the sampling effect (section 4.2.1 on page 119) – the same statistical significance effect that is analyzed with Student's t-test. However, the analysis of the variable-importance distribution can be **more informative than the null-hypothesis testing** approach inherent to Student's t-test².
- **Any data** – even with skewed distributions³ – is accepted, and no standardization is necessary.

Following Breiman, the permutation variable-importance measure for a variable x_j is estimated as follows:

1. Calculate the R_{abs} ⁴ model fit based on a dataset (e.g., the collected sample or a particular bootstrapping sample)⁵.
2. For the statistical process p , generate samples from its multivariate distribution $\mathfrak{D}_P(\mathbf{X}, \mathbf{y})$ for the independent variables x_i (with $i = 1 \dots m$ incl. j) and replace the data of the focus variable x_j with samples from an independent univariate distribution of x_j . Drawing x_j from its univariate distribution makes it independent of the

²Ordinary null-hypothesis testing, such as Student's t-test, assesses the probability of a term being insignificant – i.e., the probability that a non-zero parameter estimate in ordinary regression is in fact zero and that the estimate is only non-zero due to sampling effects, not because of a real effect in the data. Thus passing the t-test means that the true value of a parameter is not zero and lies somewhere between zero and the parameter estimate – which is not very informative for most practical investigations. Certainly t-tests yielding very high significance levels suggest that the effect strength is likely closer to the estimator than to zero – yet this is not what Student's test measures. In addition, the test requires strong assumptions, such as normal distributions. Therefore null-hypothesis testing has received much criticism in recent statistical literature (Anderson et al., 2000; Lee, 1997; Lukacs et al., 2007).

³Skewness of variable distributions cannot be removed from variables simply by using result-neutral linear transformations such as the z-transformation.

⁴Breiman uses R^2 instead of R_{abs} . But since, in section 6.2.7 on page 195, R_{abs} has been found to behave more robustly on noisy data, R_{abs} was chosen here instead. Conceptually there is no difference, since both estimators are a measure for model fit.

⁵Given the instability of R^2 with small sample sizes, as described in section 6.3.2 on page 199, the variable-importance measure used for BOGER is based on the more robust R_{abs} estimate – described in section 4.2.1 on page 119

state of the other variables x_i (with $i \neq j$) and thus the information in x_j regarding the statistical process is replaced with noise, which is identically distributed as x_j .

In practice, these sampling steps are approximated (estimated) by simply using the data for x_i from the chosen dataset as is and replacing x_j simply by a permutation⁶ of the vector x_j . Thus all variables x_i are distributed properly, and the univariate distribution of x_j remains the same, but the information in x_j – i.e., its correlation with y (and the other x_i 's) is completely removed. Similar to the sampling issues of the mean estimator, this approximation is only as good as the used sample is representative of the larger population.

3. Calculate the model fit with the permuted x_j : $R_{\text{abs}}(\text{permuted fit})$.
4. Calculate the prediction loss PL due to the permutation of x_j , i.e., the removal of the information in x_j , as:

$$\text{PL} = R_{\text{abs}}(\text{model fit}) - R_{\text{abs}}(\text{permuted fit})$$

The higher the prediction loss, the more valuable the information x_j was for the predictive power of the model (with respect to the used dataset).

5. Repeat steps 2 through 4 a large number of times (>100) in order to ensure that the sampling by permutation is truly random. Then average the prediction losses. This average is the *permutation variable importance*.

Note that the resulting estimate of the predictive loss (i.e., the variable importance) contains no information regarding the direction of the effect of the investigated variable. To obtain the effect direction, a graphical analysis method is used as a complement to the variable importance – see section 7.1.3 on page 213.

Features of the Variable-Importance Implementation of this Study

Since the \mathbb{R} -implementations of the permutation variable importance by Liaw and Wiener (2002) in the `randomForest` package and by Strobl et al. (2007) in the `party` package (Hothorn et al., 2006) are tightly interwoven with their respective random forest implementations, the above algorithm was re-implemented in \mathbb{R} for the BOGER algorithm with some additional features:

Similar to the challenges with bias in estimating the model fit R^2 (see also section 4.2.2 on page 123), the bias in the variable importance measurement calculated from an R^2 estimate strongly depends on the dataset used in its estimation process described above. Commonly all available data (the complete collected sample) is used. However, if this is the data that the model is fitted on, then the R^2 , used internally, may have a strong overfitting bias (see also section 4.2.2 on page 123). Thus the variable-importance measure

⁶The vector x_j is reordered randomly.

would not distinguish between a variable importance that truly adds predictive power to the model and a variable importance that only increases overfitting.

For BOGER, the variable-importance can be estimated based on two different datasets:

- all data with respect to the bagged model; and
- only the internal test data⁷ – for a more independent estimate.

Unfortunately, the variable importance for the internal test data cannot be calculated for the bagged model, since the internal test data is only unknown test data for a single model in the bag. Thus the variable importances for the individual models in the bag are averaged instead of using the bagged model for calculations of the R^2 .

The variable-importance estimates based on the two different data sources both have weaknesses. Yet the two estimates are biased in different ways. In the following, two different variable-importance estimates will be calculated and provided for each variable⁸:

- a **complete data variable-importance estimate**, which is the variable importance based on the final bagged model and the complete data. This estimate is preferable to estimates based on the training variable importance, since it is based on the bagged model and slightly more data and both types of estimates may be subject to overfitting biases; and
- a **test variable-importance estimate**, which is the average of all variable-importance estimates based on the test data and the predictions of individual bootstrapping models⁹.

Analogous to the discussion in sections 6.2.7 on page 195 and 6.3.2 on page 199, both measures have different strengths and weaknesses. But the weaknesses can be mitigated when the two different estimates are used in combination: the complete data importance for the bagged model may be subject to a bias related to overfitting – see section 6.3.2 on page 199. The test variable importance estimate, in contrast, is completely independent of the model fit process and thus is not subject to an overfitting bias. However, the test data estimate is based on a smaller sample and is therefore more subject to random biases from the sampling process¹⁰. Hence both estimates should be used in conjunction in order to assess the magnitude of the different biases and thus also the accuracy of the two estimates.

⁷Since there is no true (untouched) test data for the bagged model, the variable importance based on the test data is an average of the variable importances in the individual models.

⁸Future users of this measure may consider combining the model fit results from the different datasets in order to obtain a combined predictive loss estimate, allowing for a combined and robust measure of variable importance.

⁹The selection of individual bootstrapping models is restricted to those “good” individual models that are contained in the final bagged model.

¹⁰For details on the sampling bias, see section 4.2.1 on page 119.

There is no estimate based on the training data, since it would have both weaknesses of the other two measures: it is subject to overfitting biases, and it is more subject to sampling biases, since it is based on a smaller sample than the complete data estimate. Hence a training data estimate would be in all respects inferior to a complete data estimate.

To gain further insights on the magnitudes of these biases – i.e. to get an estimate of the accuracy and robustness of the variable importance measure – the BOGER implementation in addition estimates distributions of the variable importance over the different bootstrapping models in the bag. Hence the permutation variable importance is estimated with Breiman’s method for each individual bootstrapping model with many permuted datasets that are either based on the complete data or on the test data for the respective bootstrapping model. The result is many “samples” of the variable importance estimate (one for each permutation and bootstrapping model) that can be used to generate frequency plots in order to illustrate the distribution of the estimate for a particular variable over different bootstrapping models. See also the result graphics for this study in figures 7.4 on page 222 and 7.5 on page 223.

A large variance of this distribution indicates a low robustness of the variable (and the associated parameter estimates) over different bootstrapping models and thus suggests that the variable importance is likely to be affected by overfitting or sampling biases. In the language of linear regression, one would describe the parameter estimate to have a low statistical significance.

Finally, a grouping feature has also been implemented, which circumvents problems with highly correlated (collinear) variables by assessing the model prediction gain of groups of variables jointly rather than individual variables separately. This is done by permutating groups of variables in such a way that the variable value combination for a individual participant always remains intact. Consider a variable `grp.ABC` consisting of the independent variables A, B and C represented in a matrix, with a column for each variable, and the answers for each survey participant in the rows. Only the sequence of rows is randomized, but each row remains unchanged. The group of variables is also treated as a group during the permutation step described before, and therefore their collinearities are maintained throughout the process. The result is a variable-importance measure for the entire group instead of for the individual variables¹¹.

In addition, this grouping feature can also be used for grouping by concepts, e.g., all task/job-related variables in a task properties group.

In summary, the extended version of Breiman’s permutation variable-importance measure presented here is a suitable and robust estimation method to quantify the effect strength (but not the direction of the effect) of individual variables or groups of variables

¹¹Similar to collinear parameters in ordinary linear multi-variate regression, variable-importances for individual but strongly collinear variables would not be meaningful, since their effect depends on the other variables in the group.

on the outcome variable (i.e. learning) – as defined by the fitted BOGER model. Thus it is a suitable tool to draw inferences from the fitted BOGER model without the need for direct inspection of the opaque BOGER model.

7.1.2. Interpreting Interactions

When using multiplicative interaction variables, as in this study, the interaction variable AB will be highly collinear with its primary effects A and B. This correlation makes any effect strength measure for the individual variables A, B or AB become meaningless for interpretation: any change in A would also lead to a change in AB, with an additional effect on the outcome of the model. Thus an individual effect strength measure only covering variable A would not cover the correlated (i.e., actually always synchronously occurring) effect of AB.

Therefore, in ordinary linear regression, interactions require special interpretation of the parameters – see figure 7.1 by Brambor et al. (2006). The figure shows an example of a dichotomous conditional variable Z, which is either 0 or 1.

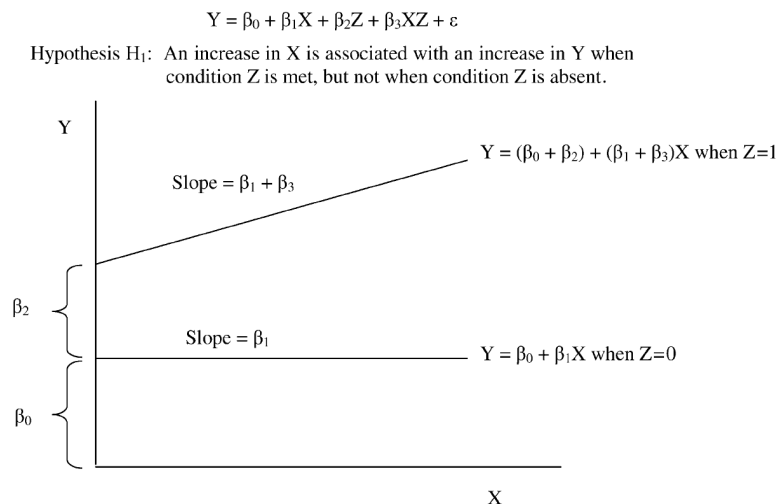


Figure 7.1.: Interpretation of Interactions (Source: (Brambor et al., 2006))

If variable Z is 1, then Y has a much stronger functional dependence on X. Thus Z can also be called a *moderator*. Neither β_1 nor β_3 is relevant for interpretation but instead $(\beta_1 + \beta_3)$. This also explains the frequently observed counteracting results of interactions and main effects: β_1 can be negative, even if the overall effect of X on Y (without conditioning on Z) should theoretically be positive. A negative β_1 in combination with a strongly positive β_3 will lead to a positive effect of X on Y for Z = 1. If the majority of the sample has Z = 1, then the overall unconditional effect of X on Y will be positive, even though the negative parameter β_1 suggests otherwise.

In a more general case, where Z is a continuous variable as well, and thus the functional relationship between Y and X lies between the two extreme functions of the dichotomous case, $(\beta_1 + Z \beta_3)$ is the relevant parameter for interpretation (Cleary and Kessler, 1982).

In the case of the mathematical model of the BOGER algorithm (see section 6.2.2 on page 180), the interpretation of the parameters is even more complicated: the multiplicative terms, for example, will interact with the linear terms. In addition, the data used by BOGER is not z-transformed to zero mean and standard deviation equal to 1, so the parameter values cannot be compared directly.

In summary, the interpretation approach by (Brambor et al., 2006), described above, is helpful to understand the signs of the BOGER parameter estimates and why interactions and main effects can have counter-acting directions. For interpretation of the BOGER parameters, however, the extended permutation variable-importance measure should be used instead, since it allows for a suitable treatment of the interactions with its grouping feature (section 7.1.1 on page 206).

7.1.3. Model Shape Graphics by Variable

Graphics can frequently visualize details and patterns more effectively than lists of single numbers, and thus computer-generated plots are a useful tool in modern statistics. Graphics showing the real data in suitable ways are valuable in detecting the unexpected (West, 2006):

*“The greatest value of a picture is when it forces us to notice **what we never expected to see.**”*, Tukey (1977, p. vi), bold in the original

Nevertheless, graphics have one severe disadvantage that needs to be compensated for: they can have at most three dimensions – or in animation, four dimensions (with the time as the fourth). Even though 3D visualization is possible¹², a good overview is frequently only achieved in a 2D plot.

Since the learning process is stochastic (i.e., noisy) and high-dimensional, these multi-variate properties of the survey dataset require a multi-variate model (as discussed in section 5.12.1 on page 163) – which was created with the BOGER model. Hence a low-dimensional analysis technique, such as 2D plots, cannot be used as a primary analysis tool for a high-dimensional problem.

However, 2D-graphics can be a useful complement to the variable-importance analysis described before, e.g., for the interpretation of the effect direction. The high-dimensional model shape will be plotted in many 2D graphs – one for each independent variable. These graphs show the dependent variable vs. a particular independent variable, graphically

¹²In \mathbb{R} , the user may create and script both 3D graphics as well as 3D animations.

illustrating the overall (unconditional) effect of a variable, including its direction (positive or negative).

The simplest approach for producing such graphs would be to use the mean of all independent variables except for the focus variable in the fitted model and vary the focus independent variable x_j over its typical range. With this artificial input data matrix \tilde{X} use the fitted model to make predictions for the dependent variable \hat{y} and plot the line of these predictions in a \hat{y} vs. x_j graph. This simplistic approach does not account for any collinearities. Thus the graphing method proposed here simply reuses the original samples. It effectively draws samples of the independent variables from the original distribution of the stochastic process $\mathfrak{D}_P(y, \mathbf{x})$ – and uses the resulting ‘artificial’¹³ input matrix \tilde{X} for predicting \hat{y} . The resulting \hat{y} vs. x_j graphs then do not show the model as a line but rather as a point cloud, since not only x_j is varied but all other independent variables as well. For better visualization, the point cloud is visualized in the form of a frequency graph showing the shape of the distribution of \hat{y} vs. x_j . For an example, see the right column of the graphs in figure 7.2 on the next page, which illustrates the dependence of learning intensity on personal interest in the learning topic (see section 7.3.3 on page 238).

Since the other independent variables are drawn from the multivariate distribution derived from the sample, collinearities are preserved automatically. Using the shape of the displayed bi-variate distributions or the median in the \hat{y} vs. x_j graphs (one for each independent variable), one can directly see the direction of the overall effect of the respective x_j .

Hence the practical implementation of the above procedure can be summarized as follows: Drawing samples from a stochastic process with the same distribution that generated the sample is approximated by using the original sample as \tilde{X} . Then, using the fitted model, predictions \hat{y} for the original input data X are calculated and plotted in many graphs – one for each independent variable.

In the following sections, figures (such as figure 7.8 on page 229) will include the distribution of the real complete data in the left column. The complete real data is shown without mean imputations and related filtering¹⁴ in the right column. Thus these graphs allow the researcher to compare the model shape (or in this case, a distribution) with the original data sample, labeled ‘*Training Data*’¹⁵ here. They will also aid in visualizing the

¹³The input matrix \tilde{X} is practically equal to the actually sampled input data X . Yet, to keep the argument as general as possible, this method would also work for an artificially generated input matrix as long as the matrix is identically distributed as the original sample, complete with all collinearities.

¹⁴Mean imputation was used to fill the many small gaps in the multi-variate dataset. Further filtering was applied for some variables to reduce the need for imputation – see also appendix section A.5.2 on page 303. This fourth and last filtering stage was not applied to the data shown in the left model shape graph as ‘training data’.

¹⁵Since the model shape is based on the bagged BOGER model, the ‘training data’ sample also refers to the complete and non-imputed data sample fed to the BOGER algorithm and should not be confused with the bootstrapping training data samples inside of the BOGER algorithm for individual bootstrapping models.

individual directions of the overall variable effects.

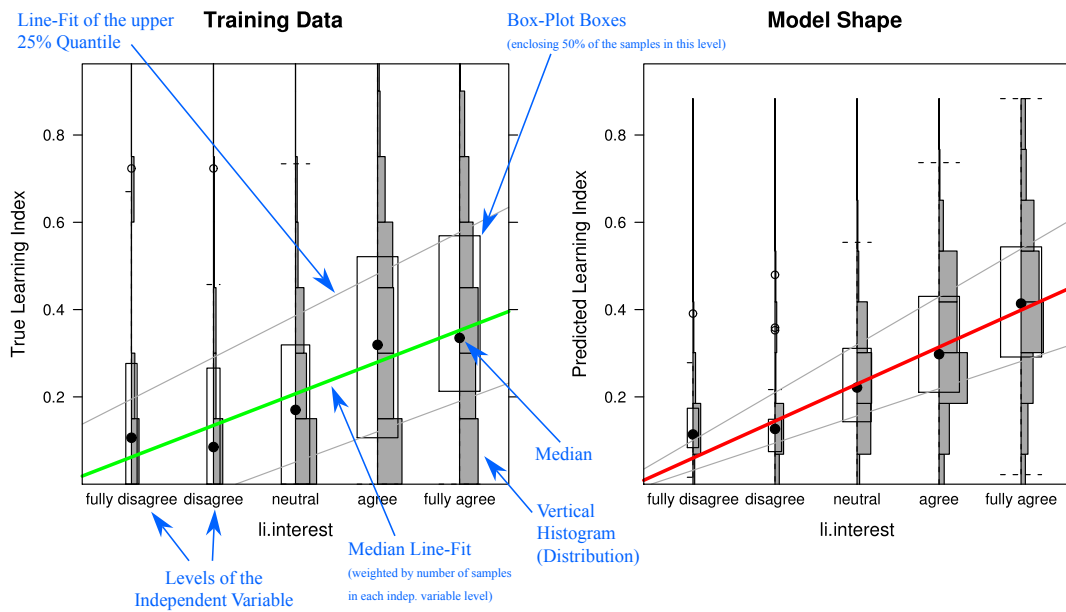


Figure 7.2.: **Annotated Model Shape Graphics:** The left graph shows the original (noisy) data for comparison with the expected value model prediction in the right graph. The distribution of the real and predicted data for the outcome variable ‘learning index’ is shown in histograms for different levels of a dependent variable ‘learning interest’. Medians and their linear fit line (in red and green) are added for a further abstracted aggregation of the relationship between the two variables’ box-plots.

To visualize as much information as possible and useful, each sub-graph in figure 7.2 shows three representations of the same data in an overlaid manner:

- The grey filled boxes show the data distribution in the form of vertically stacked histograms¹⁶ (i.e., vertical frequency plots).
- The rectangles with dark blue lines and no fill are boxes from a box-plot of the data. The blue boxes contain 50% of all samples and thus also provide information about the distribution of the data.
- The big black dots indicate the median of the learning index for a given level of the respective focus variable. The median was chosen here instead of the mean, since the median represents the typical learning index value for a given level of the focus

¹⁶A wider box means more frequent occurrence for a given range of the learning index and a given range of the independent focus variable. Thus this discrete representation is an approximation for the bi-variate distribution in the form of a probability density function (PDF).

7.1. Result Interpretation Procedures

variable, while the mean provides the average learning index. The typical learning index, i.e. the median, is more meaningful for interpretation¹⁷.

- The fitted line in green (for the real data) and in red (for the model shape) is a weighted line-fit of the medians¹⁸.

The shown graphics were created by the author using custom scripts in \mathbb{R} based on Trellis Graphics (Sarkar, 2008). Each graphic can be automatically generated with a single call of the respective \mathbb{R} function.

The vertically stacked histograms in figure 7.2 on the preceding page require an ordinal independent variable with a finite number of ordered categories, in order to be able to plot a distribution for each independent variable category. If a metric independent variable is used instead, categories for the metric independent variable (i.e., discrete levels) are automatically generated¹⁹.

These graphs also have some minor disadvantages:

- The number of levels for the independent variable may influence the slope of the line fit. In the case of an artificially categorized metric independent variable, the user may adjust the number of levels. Hence the user may influence the medians through the choice of categorization levels – and thus also the line-fits. Although more levels increase the line-fit accuracy, more levels also decrease the accuracy of the median estimator. Thus the user needs to find a good intermediate number of levels.

Using a mean estimator instead of the median will yet again lead to a slightly different line-fit.

- The BOGER model is an expected value model²⁰ of the learning index as a function of the independent variables. Hence the expected value model does not model the variance (the noise) of the process. However, the real data includes not only this expected value but also the variance of the process. Thus the real data and the model shape are not perfectly comparable, but this comparison is still informative.

In summary, despite these minor disadvantages, the advantages – such as accounting for collinearities – make this type of graph a useful analysis tool, suitable and sufficiently accurate for comparing model shape with real data and for assessing the direction of a variable effect, including collinearity frequently occurring with interactions.

¹⁷See also the nice example at http://en.wikipedia.org/wiki/Median#Popular_explanation.

¹⁸In more detail, the line-fit is a simple bi-variate ordinary least-square regression of the medians weighted by the number of samples underlying each median, i.e., the number of samples for each level of the independent focus variable.

¹⁹Alternatively, the user may provide custom intervals for the categorization of metric variables into custom levels.

²⁰see section 4.1.4 on page 108

Particularly noteworthy is the usefulness of the information inherent in distributions:

“[...] Thus, instead of characterizing statistical findings by stating percentages such as ‘70 percent of adult men have brown hair,’ researchers state, test, and do not reject the hypothesis: ‘Men have brown hair.’ Then they describe such findings by saying ‘Men have brown hair’ as if the description describes everyone or every situation. The distribution of hair colors becomes a generalization. Much of the time, such generalizations have no bases beyond computed averages, that is, ‘An average man had brown hair.’ Since social phenomena often have overlapping frequency distributions, comparisons between averages may say nothing about specific instances.”

Starbuck (2004, p. 1245)

Thus for stochastic (non-deterministic) processes, distributions are much more insightful than mean (or median) statistics. Furthermore, distributions are a direct view²¹ on the raw data without any (possibly biased and flawed) preprocessing²².

7.2. Overall Statistical Results

The following subsections show the statistical analysis results for the BOGER model as a whole. A by-variable discussion follows in section ([7.3 on page 226](#)).

7.2.1. Survey Questions of the Model Variables

In line with the research aims, the learning index is the outcome variable – quantifying the learning intensity that an individual participant experienced in a particular project or time frame at work (see section [5.4.2 on page 150](#)).

As described in section [6.3.1 on page 198](#), the application of the BOGER algorithm led to a reduced set of variables and parameters that were included as the strongest effects in the final bagged BOGER model. The 22 parameters systematically retained in the final model correspond to the variables listed in table [7.1 on page 219](#).

This final set of variables is listed with a brief description and the original variable short names, since the short names are short as well as unique and therefore reappear in a number of graphs. For the analysis of the variable importance (section [7.1.1 on page 206](#)), some of the variables have been grouped according to collinearities or on conceptual grounds. These variable groups are also listed in table [7.1 on page 219](#). Further details on the questions and scales are provided in the sub-sections referenced below.

²¹Except for choosing this type of visualization, the researcher does not make any (possibly already biased and subjective) choice regarding a filtering method – see section [2.3.3 on page 35](#). Hence the researcher effectively uses only a very light filter.

²²Unless the levels of a metric variable are artificially generated – as discussed above.

Group	Variable's Short Name	Description / Question
★ grp.LearnStrategy		Learning Strategy Group (section 7.3.1 on page 226)
	lst.reading	<i>"During your task, how have you learned ...?"</i> <i>"... by searching and reading"</i>
	lst.discussion	<i>"During your task, how have you learned ...?"</i> <i>"... by discussion with others"</i>
	learnSup.Strategy	Learning-supportive strategy scale – the inclination to support learning by reading and/or discussions.
★ grp.LeaderEffect		Leadership Effect Group (section 7.3.2 on page 231)
	dnvSc.	Knowledge-conductive task design scale, consisting of work inspiration, feedback and formal training.
	KnowlConduceTaskDesign	
	learnSup.Leader	Learning-supportive leadership style scale, consisting of feedback, initiative, climate, trust.
	inter.dnvSc.	Interaction of knowledge-conductive task design and learning-supportive leadership style.
	KnowlConduceTaskDesign_	
	learnSup.Leader	
• li.interest		Personal interest in the topic (section 7.3.3 on page 238).
★ grp.History		Professional History Group (section 7.3.4 on page 240)
	years.in.dept	Years in the department.
	age	Age of the participant.
	inter. years.in.dept_age	Interaction between <Years in the department> and <Age of the participant>.
★ grp.ApproachClear		Learning Barriers Variable Group (section 7.3.5 on page 243)
	inter.approachClear_	The three-way interaction of the learning barriers <approach not clear>, <no suitable contact found> and <expert not available> (the only variable remaining in this group – see section 7.3.5)
	foundContact_	
	expertAvail	
• EU_lmeth.IntSoc		Epistemological beliefs (EÜ) regarding social learning methods (section 7.3.6 on page 247).

• EU_lrnProc	Epistemological beliefs (EÜ) regarding the learning process (section 7.3.6 on page 247).
★ grp.TaskType	Task Type Group (section 7.3.7 on page 250)
isInnovation	Dummy variable, which is 1 if the learning index corresponds to the learning experiences from an innovation project.
isNormalWork	Dummy variable, which is 1 if the learning index corresponds to the normal work of the previous four weeks.
• proc.TaskDetail	Description Detail-Level of Procedures (section 7.3.8 on page 252)
• n.seminar	Number of seminars (section 7.3.9 on page 254)
• XDS_AG_jobClosure	Job Closure – the degree to which the participant works on a task from beginning to end (section 7.3.10 on page 255)
• bfi.open	Openness to new experiences – from the Big-5 scale (section 7.3.11 on page 257)
• taskDifficulty	Participant’s assessment of the task difficulty (section 7.3.12 on page 257)
• learnEncourage.Faults	Fear of mistakes (section 7.3.13 on page 258)
• interceptLin	Linear intercept β_0 in the BOGER model (equation 6.1 on page 182)

Table 7.1.: Variable Shortname and Question Overview

7.2.2. Variable Rankings and fitted Model Parameters

In this section, the structure of the fitted BOGER model will be analyzed by using the variable importance measure as variable ranking criterion – see also section 7.1.1 on page 206. In addition the accuracy and robustness of the variable importances will be analyzed with plots of their distributions.

As detailed in section 6.3.2 on page 199, the final BOGER model, which was fitted to the survey data, has a predictive power of $R_{\text{abs}} \approx 26\%$ ($R^2 \approx 32\%$) and shows little overfitting. Thus the model sufficiently accurately fits the data in order to draw inferences about the true stochastic process from the properties of the statistical model.

Section 7.1.1 on page 206 suggested the use of two different measures for variable importance: one based on the bagged model and the complete data, and one based on the

individual bootstrapping models and the test data. Since both estimates have different biases, they are presented side-by-side in figure 7.3 on the next page. Behind each variable name, the parameter value within the BOGER math model is shown.

Each graph is sorted by variable importance, leading to two different lists of variable rankings. It is a good sign that the variable-importance estimates decrease from a value around 0.15 to values close to 0 – even though the majority of estimates fall below 0.08. For these variables with an importance just below 0.08, the differences are small, which leads to clear differences in the ranking order between the complete and test data estimates.

To gain further insights into these differences, the accuracy and robustness of the variable-importance estimates were analyzed using frequency (i.e., distribution) plots of individual estimates in figures 7.4 on page 222 and 7.5 on page 223 – with one estimate sample for each permutation and bootstrapping model (see section 7.1.1 on page 206). While both figures show rather flat distributions, the majority of variable-importance estimates cluster closely around the median (indicated by the big round dots). This is supported by the narrow blue boxes, which contain 50% of all estimates (following the philosophy of box-plots). It is noteworthy that the variances of the different variable-importance estimate distributions represent an upper bound for the variance of the true variable-importance distribution for the bagged model. Since the frequency plots in the figures 7.4 on page 222 and 7.5 on page 223 are based on the individual bootstrapping models rather than the bagged model (see section 7.1.1 on page 206), the displayed variance is inflated and above the true variance corresponding to the bagged model. Hence the variable importances have been estimated with a limited but sufficient accuracy.

Given the sizable tolerances on the estimates, even small differences in the estimates cause significant changes in the ranking. However, it becomes clear from the distributions that the estimated differences (e.g., between the second and the eighth variable) are within the estimation tolerances. Therefore we only obtain enough information from the estimation to claim that the estimates are almost equal, and thus multiple rankings could be the result.

Yet the accuracy of the variable-importance estimates is high enough to make slightly softer statements regarding variable ranking, such as “*personal interest* is within the top five variables”. Such statements are well sufficient for interpretation and practical purposes. In the following, this slightly ambiguous ranking will be called *soft ranking*.

A comparison of figure 7.4 on page 222 and figure 7.5 on page 223 further illustrates that the differences in the ranking sequences between the estimates based on the complete data and the test data are also within the estimation tolerance and thus are marginal.

From the distribution plots, it can be concluded that the variable-importance estimates are not accurate enough to provide a single true ranking with a high confidence level. However, the distributions also show that the estimates are sufficiently accurate to provide a “soft” (i.e., approximate) ranking of the variables by effect strength – which is

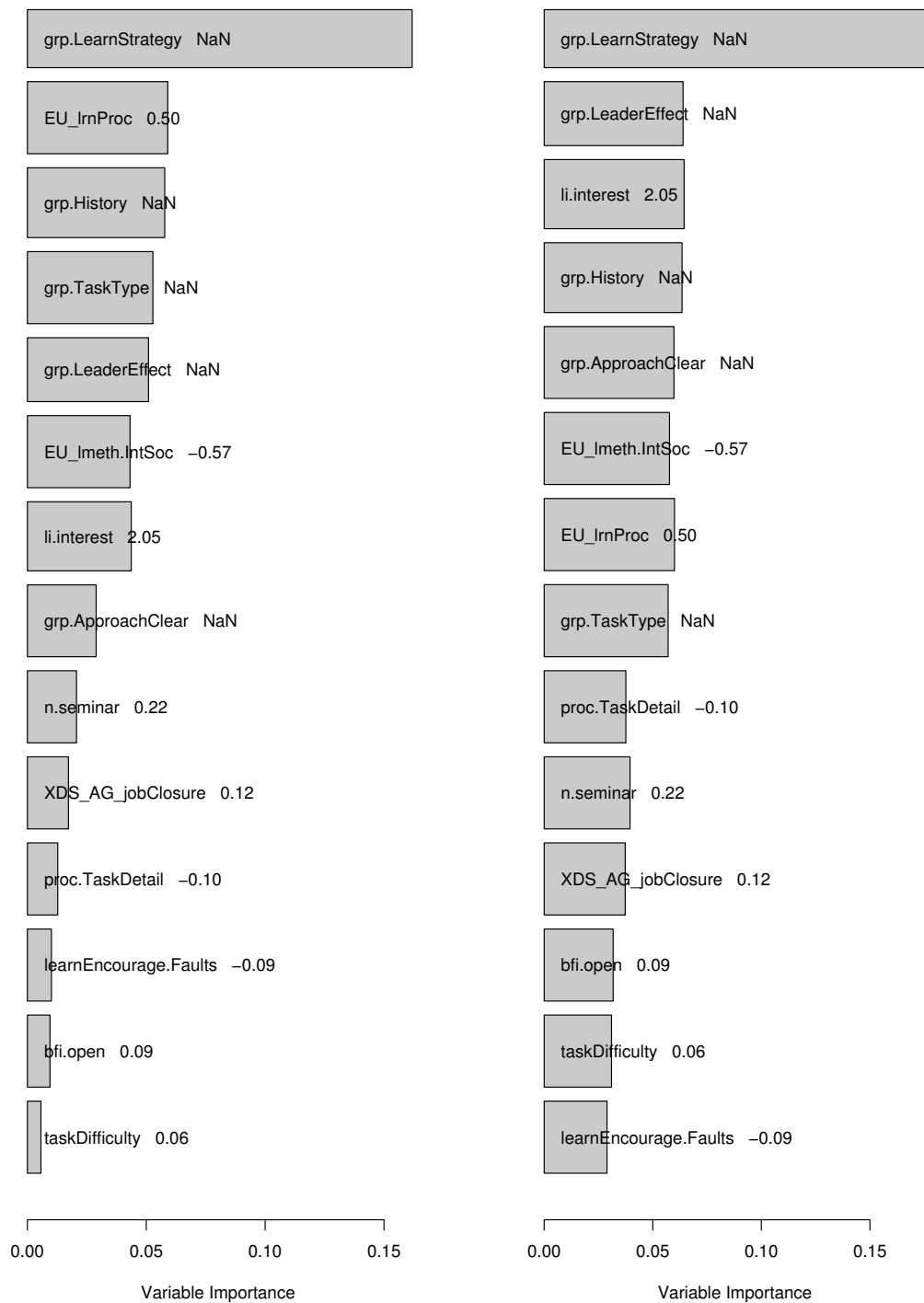


Figure 7.3.: The **Variable Ranking Results** based on the Variable Importance Measure (section 7.1.1) estimated with the **Complete Data** (left graph) and **Test Data** (right graph).

7.2. Overall Statistical Results

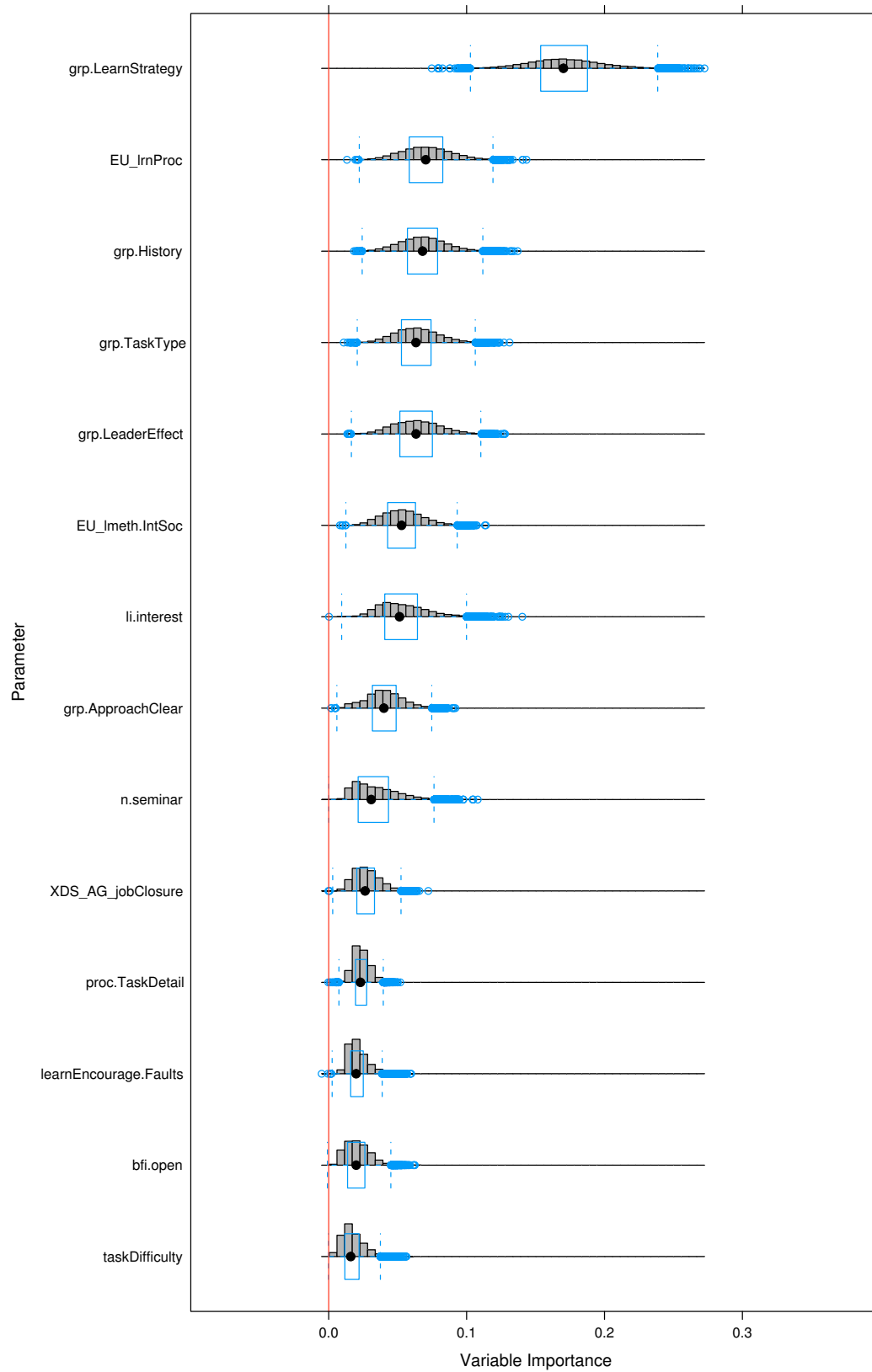


Figure 7.4.: **Variable Importance Robustness and Accuracy** Assessed with the Distribution of the **Complete Data** Variable Importance Estimate for Individual Bootstrapping Models.

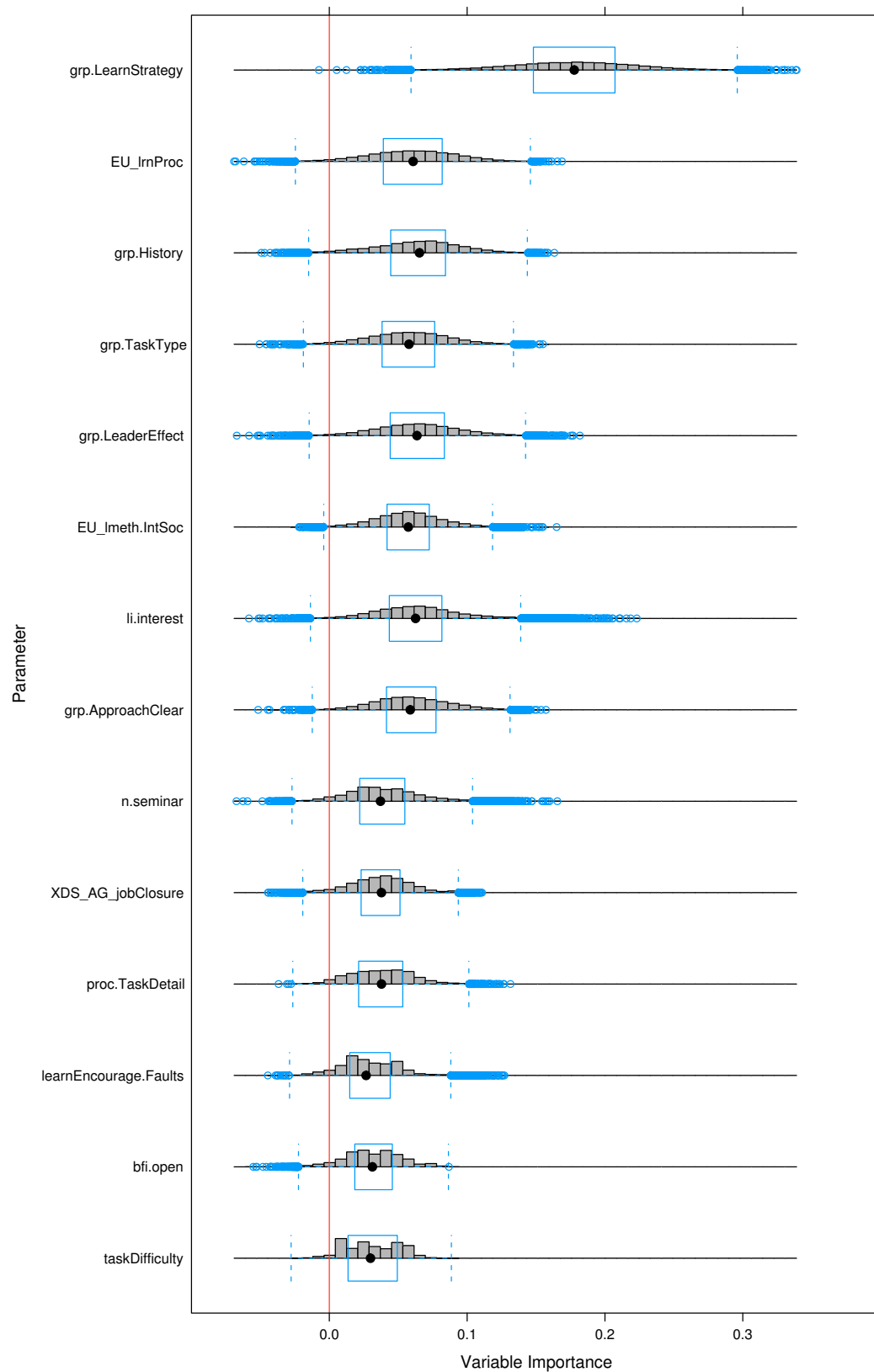


Figure 7.5.: **Variable Importance Robustness and Accuracy** Assessed with the Distribution of the **Test Data** Variable Importance Estimate for Individual Bootstrapping Models.

sufficient for the aims of this study. The distributions further allow the claim that the differences between test and complete data variable importances are marginal (i.e., within the estimation tolerances). Thus both estimation methods allow for a *soft ranking* of variables – with different but small biases. Therefore the ranking based on the test data has been arbitrarily chosen for use in the following sections on interpreting the results. This choice is equivalent – within the accuracy tolerances – to using the ranking based on the complete data.

7.2.3. BOGER Model Parameter Results

As discussed in sections [6.2.6 on page 192](#) and [7.1.1 on page 206](#), the parameter estimates for the mathematical model within BOGER (section [6.2.2 on page 180](#)) cannot be used for inference, since they stand in interaction with each other, and thus the parameter values are meaningless in isolation. The permutation variable importance measure is used instead. However, for reference the parameter estimates are shown in figure [7.6 on the next page](#).

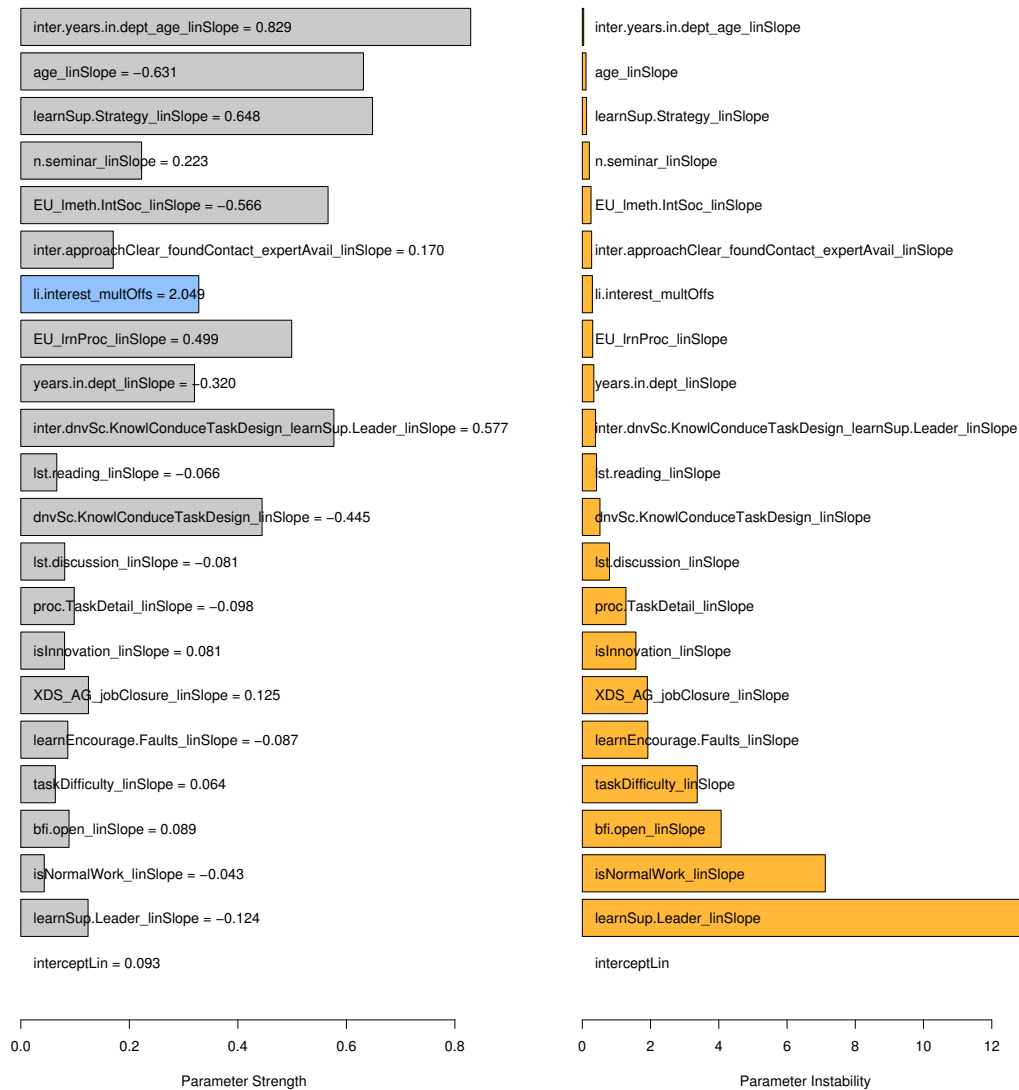


Figure 7.6.: BOGER Internals Inspected: **BOGER Parameter Values and Parameter Instabilities**

7.3. By Variable Results and Interpretation

The following subsections describe for each variable group:

- The individual variables (question items) in the variable group.
- Preliminary analysis steps and results – if any.
- The statistical results using the BOGER algorithm.
- The interpretation of the results, including the use of results from relevant literature.
- A summary.

A summary of the interpreted results can be found in section 8.1.2 on page 267 with the implications in section 8.1.2 on page 267.

7.3.1. Learning Strategy Profile

The Statistical Results – Learning Strategy The model contains a group of questions that all begin with the phrase “*During your task, how have you learned ...?*”, followed by:

- “... *by searching and reading*” (`1st.reading`).
- “... *by discussion with others*” (`1st.discussion`).
- “... *by trial and error as well as experimenting*” (`1st.experiment`).
- “... *by demonstration from other people*” (`1st.demo`).
- “... *by investigation and analysis of event in the past*” (`1st.analysis`).
- “... *by other methods*” (`1st.other`).

The questions are arranged on a single survey screen page, so that the five-level scale from “*not at all*” to “*very much*” is used to comparatively rank the different learning strategies. It is difficult to define scale anchors to connect the answering category “*very much*” with an absolute intensity of strategy use – especially when they need to be easily and quickly understandable by the participants. Moreover, the comparison of the different learning strategies is most important in the context of this study. Hence no attempt has been made to define hard scale anchors, such as “*daily*”. Since this scale design is only suitable for comparisons, individual answering biases²³ are removed by subtracting the mean of all answers for each participant from each scale variable. The result is an exclusively relative scale (rather than an abstract rating) – i.e., isolated comparisons of individual learning strategies are not meaningful.

²³ ‘*Answering bias*’ refers here to individual tendency to use the scale. For example, one participant might reserve “*very much*” for exceptionally intensive uses of a strategy, while another person might very commonly use “*very much*”, while both participants may mean the same intensity.

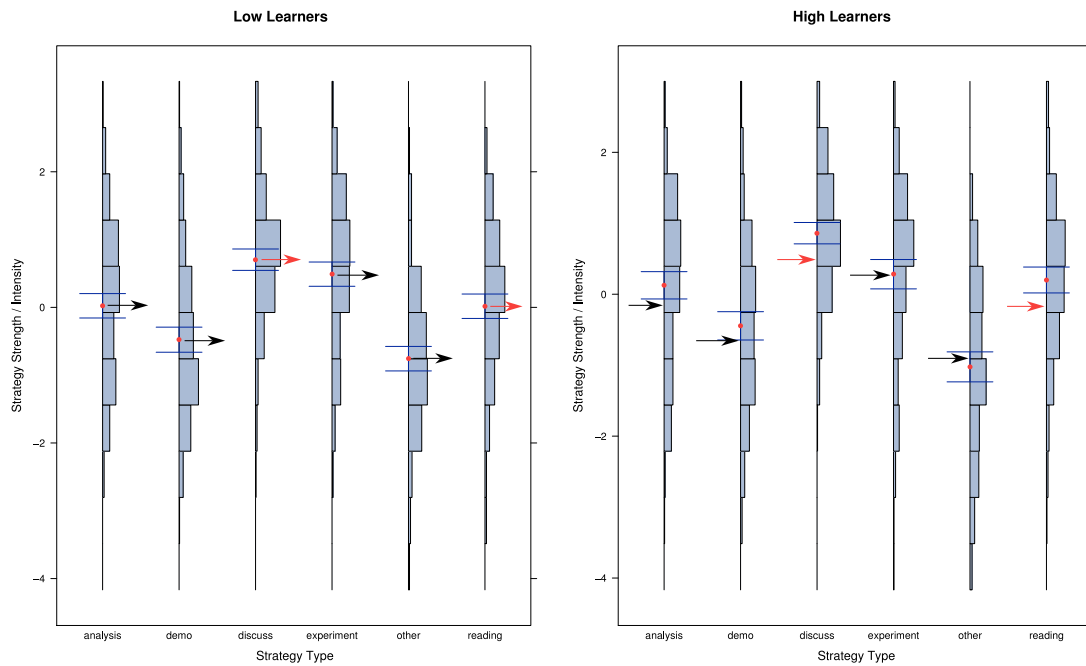


Figure 7.7.: **Learning Strategy Profile Dimensions Compared** for Participants with a High and a Low Learning Index. For comparison, the arrows in the right graph, for the ‘high learner’ group, indicate the values from the left graph, for the ‘low learner’ group. ‘High learners’ employ the **reading** and **discussion** learning strategies relatively more frequently than the ‘low learners’ (see red arrows).

For a preliminary graphical analysis²⁴, the profile of employed learning strategies is compared between a “*low learner*” group of participants, with a learning index in the lower 50% quantile, and a “*high learner*” group, with a learning index in the upper 50% quantile. The result is shown in figure 7.7. For each of the above learning strategies, the figure shows the distribution complete with mean and 95% confidence interval for the mean. The red and black arrows in the right graph show the values of the “low learners” from the left graph for direct comparison with the “high learners”.

Using the arrows as a visual aid, different profiles of learning strategies become visible between the *low learners* and *high learners*. While the use of different strategies varies strongly within the two groups, only the more frequent use of the ‘searching and reading’

²⁴This method is only a preliminary analysis because the choice to split participants into two or more groups may strongly influence the graphical result. Possible alternatives to the method described here would be to use more groups (i.e., multiple levels of the learning index), or to use the top 30% and bottom 30% of all samples as the cut-off point for the high and low groups in the learning index – as used in microelectronics. Continuous models, such as the one fitted by the BOGER algorithm, avoid the problem with selecting suitable levels completely.

as well as the ‘discussion’ strategies separates the high learners from the low learners – as the red arrows pointing to the differences in the means between the two groups suggest²⁵.

This preliminary finding and the interactive search²⁶ for a BOGER model with a high predictive power has led to the creation of the variable group `grp.LearnStrategy`, which consists of the following variables:

<code>1st.reading</code>	Question: <i>“How have you learned? ... by searching and reading.”</i>
<code>1st.discussion</code>	Question: <i>“How have you learned? ... by discussion with others”</i>
<code>learnSup.Strategy</code>	A special scale for this learning strategy profile calculated as the normalized sum of <code>1st.reading</code> and <code>1st.discussion</code> . This artificial variable acts in a similar way as an interaction of the two variables ²⁷ .

In the final BOGER model, the three abovementioned variables have a positive effect – as shown in figure 7.8 on the facing page. (Details on reading figure 7.8 on the next page are provided in section 7.1.3 on page 213.) Only the variables in the variable group `grp.LearnStrategy` – and not any of the other strategies – added predictive power. Hence the findings of the preliminary analysis are confirmed.

Going through different model fit iterations with the BOGER algorithm using either the combined scale (`learnSup.Strategy`) or the variables for the question items on reading and discussion individually, the results show that the scale has the strongest and most stable effect on learning. Compared to any of the other variables, the combined scale also has by far the strongest effect – see figure 7.3 on page 221. Both the training and internal test variable-importance measures agree. Interactions with other variables (e.g., leadership) could not be confirmed.

The increased stability of the BOGER model parameter associated with the scale is not surprising, since the scale enters the model with only one parameter compared to the two parameters of the two individual item variables.

Using a scale instead of a single item broadens the scope of the measurement. The scale, calculated from the normalized sum of the two individual item variables, measures the overall intensity with which the participant used the reading strategy or the discussion strategy in *any* combination. Since all other questions on learning strategies are excluded, the scale can be understood as a particular learning strategy profile. A high variable importance of the scale variable compared to the individual item variables, in conjunction with the fact that the reading or discussion strategies are interchangeable²⁸, strongly

²⁵Testing for statistically significant mean differences is not the intention here, given the preliminary nature of the analysis. The analysis of the BOGER model will provide more robust findings.

²⁶For the final full model search, see section 6.2.6 on page 192.

²⁸Possibly there is a non-linear combination in which the reading or discussion strategies act, but the data from this study is not sufficient to detect it.

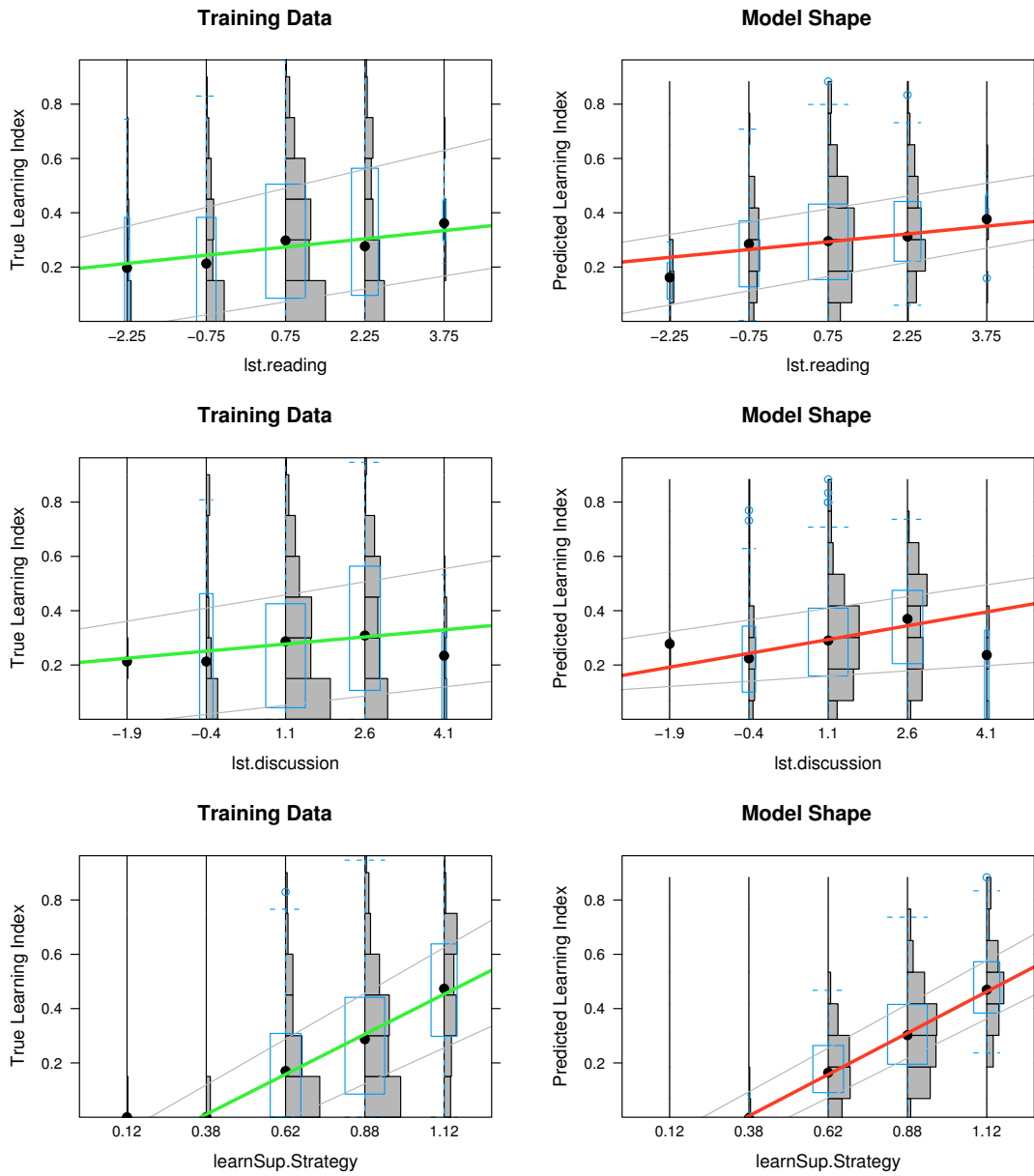


Figure 7.8.: **Positive Associations** of the Reading and Discussion **Learning Strategies** as well as the sum of both in a special scale (`learnSup.Strategy`) with **Learning** is shown with a Model Shape Graphic (see p. 215).

suggests that it is a common property of the reading or discussion strategy that drives learning.

Interpretation of the Results – Learning Strategy Reading & Discussion Hence the question arises: which common property of reading and discussion – not shared by any of the other strategies – is exerting such a strong positive effect on learning? In the survey data, no question item highly correlates with either the learning or the reading learning strategy. Thus the common property must be a hidden latent variable (Pearl, 2003) (a variable not covered by the survey), which drives learning indirectly – mediated by the reading and discussion learning strategy.

Compared to any of the other learning strategies (such as experimentation or analysis of past events), the strategies ‘reading’ and ‘discussion’ involve dealing with other people’s ideas and opinions regarding the subject, mediated by language – i.e., with a variety of perspectives on the problem.

The learning-supportive effect of dealing with a variety of perspectives and refining one’s own perspective on the problem has been observed by many scholars – as described in section 2.3 on page 28. As discussed in section 2.3.4 on page 41, language centrally shapes perspective setting, which in turn drives learning by integration of information suitably filtered from all available data (section 2.3.3 on page 34). Furthermore, in their case study on a civil construction company, Salter and Gann (2003) found that interaction with others supports learning and innovation most strongly.

Thus the strong effect of the “reading” and “discussion” learning strategies, combined with the results of these studies, strongly suggests that the exposure to many different perspectives mediated by language is the primary driver of on-the-job learning.

While the “reading” and “discussion” learning strategies linked to integrating facts yield the strongest effect on learning by far, this finding does not imply that other strategies focusing on the targeted collection of information, for example “experimentation” or “analysis”, are not important. In Orr (1996), a discussion that considers different perspectives is the central feature of copy machine fault diagnosis. Yet the ability to quickly check²⁹ and validate hypotheses by investigating the machine or exchanging a spare part is also an essential feature of diagnosis – helping the technicians to effectively steer and focus their discussion. Hence targeted experimentation and analysis, based on the current hypotheses, may still be an essential tool – even if not the central one – for reducing the number of plausible perspectives and creating a more solid link between the current perspectives and observations of reality³⁰. Without this grounding or validation effect, many

²⁹I.e., perform a *reality check*.

³⁰Personal Note by Author: My personal experiences during my doctoral research confirms this argument: most of what I personally learned is not from the survey data but from other studies. Yet, had I not run into problems with the conventional statistical methods, for example, I would never have looked for and learned about more modern methods.

more equally plausible perspectives would remain, and Orr's copy technicians would not converge towards a single perspective on the problem and a single solution that solves their copy machine problem in reality.

That this effect did not appear in the survey data during the statistical analysis, e.g., by an interaction, is not surprising, since this secondary effect is likely too weak to stand out from the statistical noise in the data.

A rival yet related explanation is that discussions have the additional benefit that they force the learner to articulate his/her reasoning, which is similar to the self-explanation strategies found by Siegler (2005). In Siegler and Chen (2008), he found that asking children to explain their reasoning for both correct and false answers strongly supported the children in building appropriate mental models to solve the example problem of Siegler's experiment.

In summary, the results of the survey suggest that exposure to other perspectives and with it the refinement of the learners perspective on the problem is one of the most important activities driving learning. The pure data/information gathering aspect of all learning strategies can not be equally important, since the strategies analysis, experimentation and demonstration show a much weaker effect – despite the benefit of generating truly new data. Thus the results confirm the literature findings behind the PIA-model (figure 2.1 on page 31), which describes perspective taking³¹ as a central preparatory step to learning.

7.3.2. Leadership Effect

The Statistical Results – Leadership The leadership style of the participant's superior (as perceived by the participant) was measured by a standard scale, which has been introduced by van de Ven et al. (2000).

The effect of leadership was analyzed in a similar manner as the learning strategies. A preliminary study, shown in figure 7.9 on the next page, indicates that the “high learners”³² are exposed to a different leadership style or profile than the “low learners”. The red and black arrows in the right graph indicate the mean value of the “low learners” in the left graph for direct comparison with the “high learners”.

The red arrows in figure 7.9 on the following page indicate the leadership dimensions, which are more pronounced for the leaders of the “high learners” (in comparison with the leaders of the “low learner” group):

- **ldr.OwnInitiative** – Leaders of this innovation encourage individuals to take initiative.

³¹See section 2.3.2 on page 30.

³²The participants with a learning index in the upper 50% quantile.

7.3. By Variable Results and Interpretation

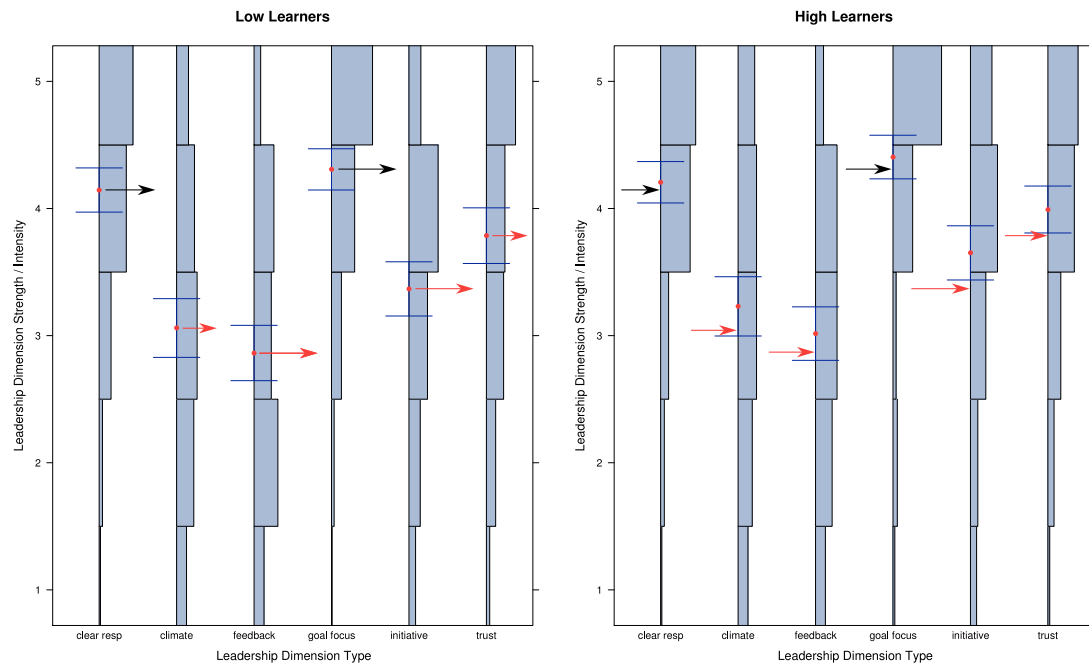


Figure 7.9.: **Leadership Profile Dimensions Compared** for Participants with a High and a Low Learning Index. For comparison, in the right graph, for the ‘high learner’ group, the arrows indicate the values from the left graph, for the ‘low learner’ group. Leaders of the ‘high learners’ on average more strongly encourage initiative, provide feedback and value group climate and trust (see red arrows).

- **ldr.SupFeedback** – Frequency that individuals involved in the innovation receive constructive feedback from the leader on how to improve their work.
- **ldr.ClimateFocus** – Leaders of this innovation place a strong emphasis on maintaining group relationships.
- **ldr.Trust** – Leaders place a high level of trust in individuals connected with this innovation.

In contrast, the following leadership dimensions do not differ much between high and low learners:

- **ldr.GoalFocus** – Leaders of this innovation place a strong emphasis on getting the work done.
- **ldr.ClearResponsability** – Individuals connected with the innovation are clear about their individual responsibilities.

This is striking, since these latter two dimensions can be seen as the principal or bare minimum leadership skills of a conventional leadership understanding. The above two dimensions still describe positive aspects of leadership; however, focusing on goals³³ and assigning clear responsibilities do not particularly support (or hinder) learning when it comes to getting departmental tasks done.

Similar to the learning strategies, a leadership profile scale was constructed from the sum of the four question items showing a raised effect for the high learners. Since some important dimensions of management are missing, this scale – strictly speaking – bundles only a subset of important qualities that may contribute to a leader’s effectiveness. Even though the scale is termed ‘*leadership profile*’, it is not a general and all-encompassing leadership measure but rather a measure of certain aspects of leadership that are relevant to learning.

This new profile variable `learnSup.Leader` showed higher predictive power for learning than the original full leadership scale by [van de Ven et al. \(2000\)](#) – see also variable `learnSup.Leader` in the model shape graphic [7.10 on the next page](#).

The Dimensions of Leadership in Literature Many authors stress the importance of leadership for long-term business success ([Collins, 2001b](#); [Deming, 1985](#)). A number of authors further highlight that certain management styles support learning ([Uhl-Bien et al., 2007](#); [Vera and Crossan, 2004](#)).

Despite this existing literature, there has been no investigation of how Van de Ven’s notion of leadership (as quantified by his leadership scale) particularly affects learning. However, the impact of the components of the scale have been covered as individual research subjects:

- **Individual initiative**, i.e., an active approach by the learner enhances learning – as discussed in section [2.3.1 on page 28](#). Leaders who trust their team members and encourage individual initiative thus allow for effective learning experiences if the team members make use of this autonomy in order to solve problems. Furthermore, by allowing for individual initiative, leaders show appreciation for employees’ skills and thus strengthen employees’ identity as specialists, which in turn is a strong motivator ([O’Donnell et al., 2003](#); [Orr, 1996](#)).
- **Good relationships** between the team members support communication within the team. [Cross et al. \(2001\)](#) refer to workers’ feeling of ‘safety’ in their relationship with colleagues when engaging in learning-effective discussions – especially when the questions reveal knowledge deficits. [Orr \(1996\)](#) and [Nonaka \(1991\)](#) also underline

³³[Collins \(2001b\)](#) argues that an unwavering commitment to corporate goals is one of the primary success factors of longer-term outstanding leaders.

7.3. By Variable Results and Interpretation

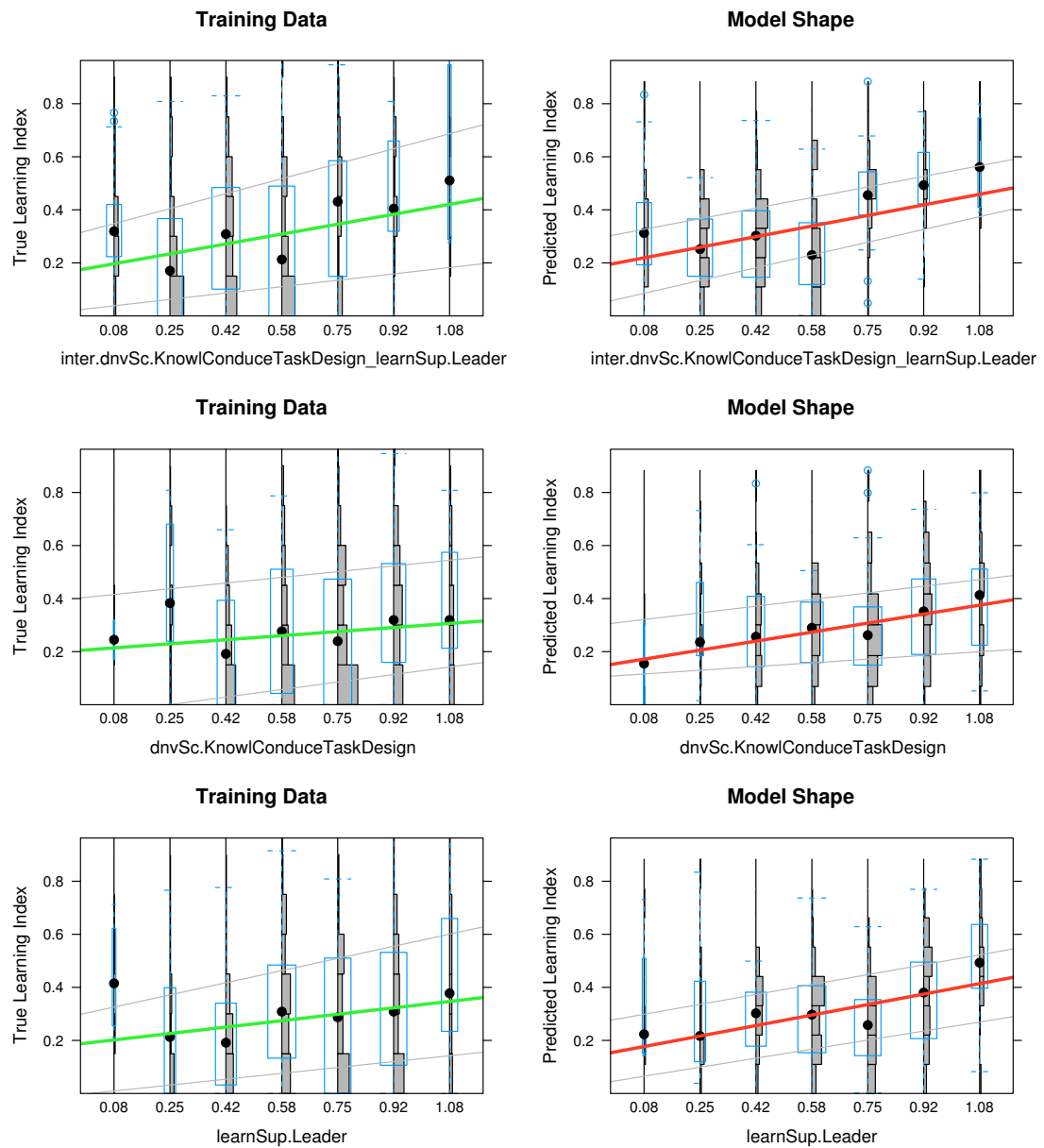


Figure 7.10.: A **Positive Association** of the Learning Supportive **Leadership Profile** with **Learning** shown with a Model Shape Graphic (see p. 215).

the importance of good personal relationships as a factor supporting learning and knowledge exchange.

- **Feedback:** The leadership profile comprises general feedback as well as feedback from the superior. However, the question items are not specific enough to know whether the feedback is on the task results or on the methods the individual employee used to attain the results.

The literature points to the importance of general feedback as well as feedback regarding the methods:

Sengupta et al. (2008) stress the importance of suitably visualizing³⁴ the interconnected and delayed effects of project management interventions – i.e., generating proper feedback about the project management task, allowing managers to learn and improve their project management skills. Similarly, feedback from measurements and experiments in the form of active hypothesis testing (by exchanging spare parts) is an essential part of the learning and problem-solving episodes described in Orr (1996) – second only to integrating different facts from prior knowledge and the feedback.

In educational psychology, Butler and Winne (1995) describe a learning process that involves not only external feedback, e.g., in the form of test grades, but also internal feedback about the learning process, automatically while learning³⁵. Similar models are used in Roßnagel (2008) and (Zusho et al., 2003).

Finally, feedback also shapes or focuses the perspective on a problem – as illustrated by the PIA-Model (fig. 2.1 on page 31) and the descriptions of its outer feedback loop in section 2.3.6 on page 47.

Section 2.4 on page 57, on the three industry practices EFQM, TPS and project management, discussed the creation of a shared perspective on the organization's processes and problems, a perspective strongly guided and systematized by management, which is a very effective measure to drive organizational change. Hence positive management influence on learning could also be mediated by jointly developing a shared perspective as a basis for learning. This aspect has not been covered in this survey and thus would be an interesting topic for future research.

Thus leaders may affect learning by cultivating diverse and alternative perspectives, which leads to the creation of more robust shared perspectives – as discussed in section 2.4.5 on page 66.

³⁴See also section 2.3.3 on page 35 on the effect of visualizations on learning.

³⁵While their research focuses on school or university settings with fixed and clear learning goals as well as pre-selected learning materials, the model allows for very different learning goals: truly acquiring a skill or only pleasing the teacher. Hence it is likely that the model generalizes also to learning situations in work settings with a task goal rather than a skill goal.

- **Trust:** Trust within the team has been widely accepted as an important basis for effective and intensive collaboration (Becker, 2006; Liker, 2004; Salter and Gann, 2003; Sandow and Allen, 2005; Szulanski et al., 2004).

Statistical Results – Leadership & Task Design In addition to the profile of learning-supportive leadership, an interaction between `learnSup.Leadership` and the DNV scale “Knowledge-Conducive Task Design” added predictive power to the model.

The DNV scale³⁶ “Knowledge Conducive Task Design” is composed of the following question items:

- “*My work offers many inspirations for new ideas and innovations.*”
- “*Seminars that are important for me, are offered to me.*”
- “*During my work I get feedback that helps me improve all the time.*”

Thus the scale describes a (perceived) learning-supportive work environment.

In order to check for counteracting effect, both the interaction as well as the two main effects were added to the model. Since the interaction and the two main effects are collinear, they have been joined to a common variable group `grp.LeadershipEffect` for the variable-importance calculation in figure 7.3 on page 221. Hence figure 7.10 on page 234 also shows all three variables.

When the ranking of variables by strength is concerned, the group `grp.LeadershipEffect`, consisting of all three variables, is placed second by the internal test variable importance measure but fifth by the complete data variable importance, even though the absolute values for variable importance are not very different – see figure 7.3 on page 221. Given that both variable-importance measures are only estimators with limited precision, only a *soft claim*³⁷ can be made:

The variable group for the effect of leadership is amongst the five most important variables.

Note that the parameter values in figure 7.6 on page 225 for the two main effects are both negative, while the interaction is positive. In figure 7.10 on page 234, however, all variables show positive effects (including the collinearity effect) – i.e., the interaction is stronger than the individual direct effects of the leadership profile and the knowledge-conducive task design variable. Thus both variables occurring with high values concurrently (in high correlation) drive learning more strongly than do the individual effects.

³⁶With a Cronbach’s alpha of 0.62, the scale is fairly reliable, i.e., in a factor analysis the vectors of its items would point in similar directions. The answers of the items are related but not perfectly identical and thus only partly redundant. Removing an item reduces the alpha to 0.50 – 0.56. Hence the scale cannot be improved by removing an item.

³⁷See also the discussion on “soft ranking” with the two variable-importance estimators in section 7.2.2 on page 219.

Hence a learning-supportive leadership style, occurring concurrently with a learning-supportive work environment, drives on-the-job learning even more strongly than do the two individual scales in isolation.

Interpretation of the Combined Results – Leadership & Task Design Given the result that the modified leadership profile and the knowledge-conducive task scale correlate and yield a higher predictive power in the multi-variate BOGER model, there are three different possible interpretations³⁸:

1. Leadership causes a knowledge-conducive task design.
2. A knowledge-conducive task design for the employees causes managers to adopt a certain leadership profile.
3. Leadership and a knowledge-conducive task design frequently occur concurrently (both at the same time) but are caused by another latent variable³⁹ that was not collected and thus remains hidden in the background.

Interpretation 2 suggests a causal link between the employee work environment and the manager's leadership behavior, which is not a plausible interpretation and thus can be discarded. In order to support interpretation 1 with few doubts, other methods – e.g., a true experiment or evidence from other studies – are needed to establish the causal link (Hitchcock, 2007). When no such evidence is found, interpretation 3 is the most likely interpretation, leading also to the weakest claim:

A leader who puts trust in his team members, encourages employee initiative, cares about group relationships and gives useful feedback frequently leads to productive results and possibly shapes an environment in which the team members perceive their jobs as inspiring, get opportunities for continuing education and get useful feedback from the superior or others. These inspiring working conditions for the team members lead to a higher team member learning index.

In summary, the literature widely supports the result that the aspects bundled in the leadership scale and the knowledge-conducive task scale have a positive effect on learning. From the analysis, a particular learning-supportive leadership profile emerges, including the following aspects: supporting employee initiative, feedback, group climate and trust. Not included are the classical leadership dimensions: goal focus and clear division and assignment of responsibilities. However, no support could be found for a causal connection in which leadership shapes a learning-conducive work environment. Hence, it remains

³⁸see section 4.1.5 on page 109, on causality, and section 7.1.2 on page 212, on interactions

³⁹An example of such a hidden latent variable could be a manager's specific leadership style or personality traits.

unclear how leadership affects the work environment in detail. Therefore the survey data combined with the literature only allows the claim that the learning-supportive leadership profile and a learning-conducive work environment together (i.e., occurring at the same time) strongly support learning, and that this combination is one of the most learning-supportive factors. Future research should aim to shed more light on how leadership supports learning – e.g., by investigating how leaders shape the perspectives of their employees and what the effect of this perspective shaping is.

7.3.3. Personal Interest

The Statistical Results – Personal Interest Another important factor in the fitted BOGER model is personal interest in the topic of the task or project. The participant’s level of interest was gauged with the question:

“I was also interested in the topic personally – independent from my tasks.”

This effect is strongly positive, as can be seen in figure 7.11.

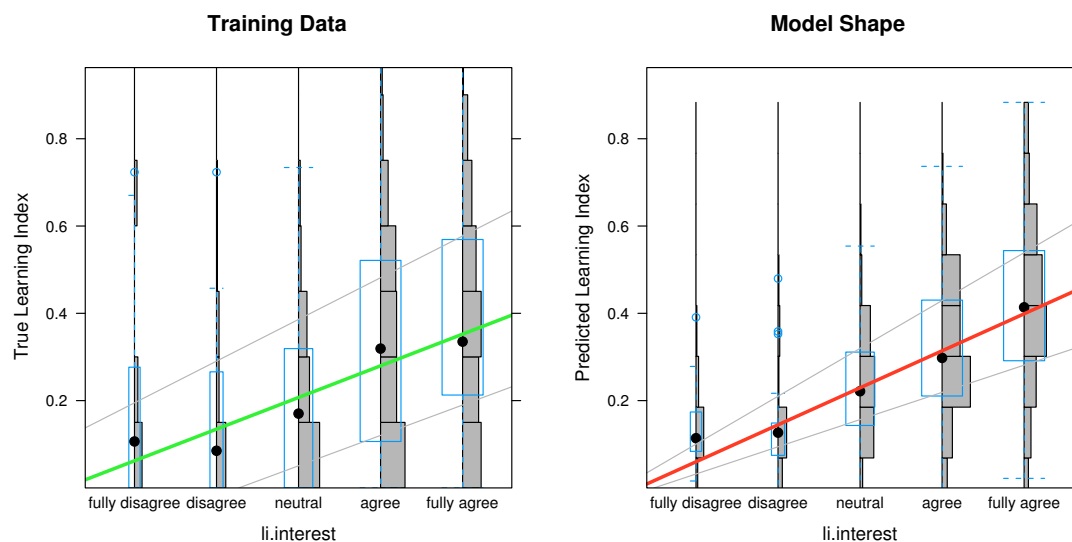


Figure 7.11.: A **Positive Association of Personal Interest in the Topic with Learning** shown with a Model Shape Graphic (see p. 215).

The internal test and complete data variable measure give slightly different rankings for the variable of personal interest: third and seventh place, respectively. Similar to the leadership effect, the absolute variable-importance measures do not differ much and are both close to 0.5. It is also noteworthy that this is the only variable that is stable with a (non-linear) multiplicative term in the final model (there is no linear term in addition).

A correlation of personal interest with other variables (e.g., leadership) was not found. Thus there is no indication in the survey data that personal interest is driven by any of the other survey variables.

Interpretation of the Results – Personal Interest In the theory section, it was argued that knowledge transfer critically depends on the active learning engagement of the learner (see section 2.3.1 on page 28). Not surprisingly, therefore, intrinsic motivation for learning is one the primary factors for driving the iterative learning process (section 2.3.6 on page 47). This is also in line with the fact that the multiplicative term is stable in the final model, since the multiplicative term suggests that personal interest acts on learning only in conjunction with (i.e., with an AND-relationship⁴⁰) other driving factors. As discussed in section 5.12.3 on page 168, personal interest acts like an amplitude function on the variance.

Colquitt et al. have found in a meta-analysis that motivation for learning is a strong predictor for the learning outcome, in some cases even stronger than cognitive skill (Colquitt et al., 2000, p. 681). Personal interest in the topic is very closely related to the learning motivation concept of Colquitt et al. (2000), and thus it can also be understood as an intrinsic motivator.

In another direction, Butler and Winne (1995) argue that (academic) learning activity is likely to halt when learning performance is much below the learner's goals and expectations, and as a consequence learning motivation is reduced during the course of a learning episode⁴¹ (details were discussed in section 2.3.6 on page 48).

Hence the following question arises: what is the direction of the causal link between intrinsic motivation and learning? Does personal interest in a topic drive learning, or does learning drive problem-solving performance and thus also positively affect personal interest?

Future studies should further investigate whether the causal effect is only in one direction or possibly even in both directions. A causal link in both directions would be in line with the theoretical insights on iterative learning (with feedback loops) presented in section 2.3.6 on page 47.

In addition, it should be investigated whether extrinsic motivators such as monetary incentives also lead to more learning in work contexts. For academic contexts, Butler and Winne (1995) claim that extrinsic motivators (e.g., good grades on an upcoming exam) lead to different and less effective learning behavior. Does this insight for academic learning also translate to problem-solving situations in organizations?

⁴⁰For a discussion of AND-relationships and multiplicative terms, see section 6.2.2 on page 180.

⁴¹In having personal interest as an independent effect on learning in the math model for this survey, it is implicitly assumed that personal interest in a topic is a fairly stable and independent characteristic of people and that the causal direction of effect is from personal interest to learning intensity.

Moreover, it should be investigated whether there are any other variables that support personal interest that have not been included in this survey (e.g., certain features of task design or a particular behavior of the superior).

In summary, from the empirical results together with the theoretical insights that intrinsic motivation causally acts on learning, it can be concluded that personal and intrinsic motivation for a topic strongly supports learning. In addition, it needs to be further investigated whether learning performance also feeds back into and thus affects personal motivation.

7.3.4. Personal Working History Variable Group

Statistical Results – Personal History A number of variables related to participants' age and personal working history have been tested for predictive power in the BOGER model, but only two variables and their linear interaction added predictive power:

- Age of the participant
- Number of years the participant worked in the current department

These variables and their interaction are part of the BOGER model with simple linear terms. The two main effects are highly correlated with $\rho = 0.95$ – which has led to the introduction of the interaction.

Drawing on the model shape graphics (figure 7.12 on the next page) and the results for the (average) parameter estimates within the BOGER model (fig. 7.6 on page 225), table 7.2 summarizes the statistical results. The 'Training Data Shape' column describes the results from the left shape graph, based on the complete training data, while the column '(Bagged) BOGER Model Shape' summarizes the results from the right model shape graph (fig. 7.12).

<i>Variable</i>	<i>Parameter Estimate</i>	<i>Training Data Shape</i>	<i>(Bagged) BOGER Model Shape</i>
<i>Age</i>	negative	neutral – no visible and strong effect (relative to the noise)	neutral but non-linear around 9 years
<i>Number of Years in current Department</i>	negative	neutral	neutral
<i>Interaction between Age and Years in Department</i>	positive	neutral	positive and slightly non-linear

Table 7.2.: Statistical Results – Personal History

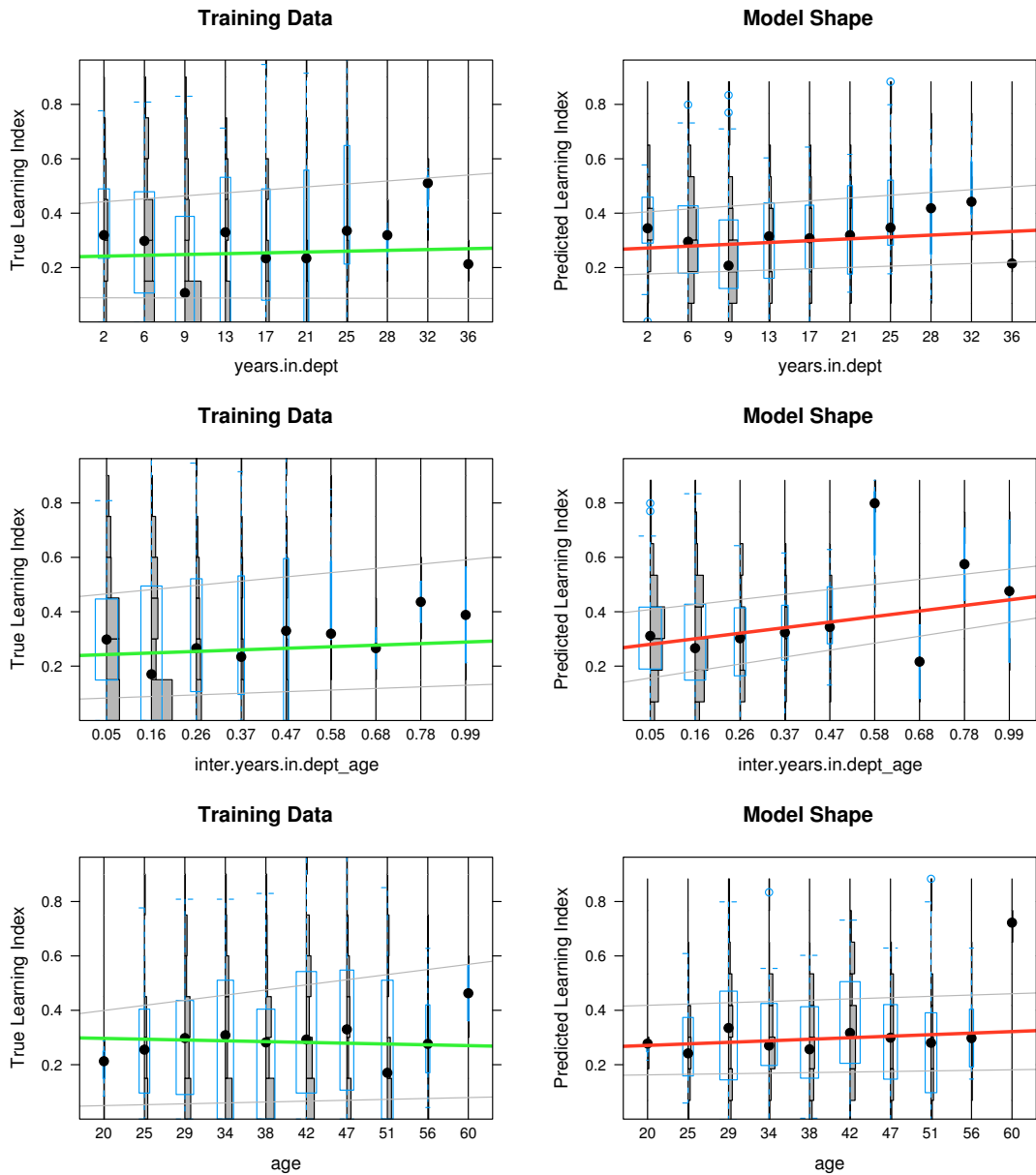


Figure 7.12.: Relationship of **Job History** with **Learning** shown with a Model Shape Graphic (see p. 215).

At first sight, it may be surprising that the parameter estimates for the main effects and the interaction in table 7.2 on page 240 have counteracting signs: years in department (negative), age (negative) and the interaction of both (positive). However, as detailed in section 7.1.2 on page 212, the parameter signs in isolation are meaningless, since main effects and the interaction always act in conjunction and thus the model shape graphics should instead be the source of insights on the direction of the effects.

In summary, the results from the model shape graphics in table 7.2 on page 240 reveal that neither age nor years in the department has a strong *direct* effect on learning. However, the interaction is strongly associated with a high learning intensity. Therefore in a variable group, together with the two main effects, the interaction was added to the final model – which has led to an increased predictive power.

Result Interpretation – Personal History The statistical results from the survey data imply that staying a long time in a single department at a young age reduces the learning effect, while staying longer as a senior member in a department at an older age supports learning.

While there is some evidence that it takes more effort to achieve the same learning results at an older age (Baltes and Staudinger, 1999, p. 476), other studies show that certain groups of older employees (e.g., managers) have found a way to maintain their everyday cognitive performance, allowing them to learn and adapt even in turbulent environments, which require such a learning ability for continued success (Colonia-Willner, 1999, p. 602). Roßnagel (2008) also reports that older employees can train and maintain their skills for formal learning. Hence, despite the popular “wisdom” that older employees learn less, evidence from the literature supports the statistical results that age is not necessarily a limiting factor for learning.

Regarding the effect of personal job history, the PIA-model (figure 2.1 on page 31) provides a lead: personal history as well as personal job history are incrementally created from many episodes of experiences – composing a person’s body of prior knowledge. This prior knowledge strongly affects perspective setting, i.e., the way one filters all available data, and thus it also affects learning and decision making (section 2.3.4 on page 41). A particular structure or kind of personal experience may lead to a learning-supportive filtering behavior and thus support the learning process as a whole.

In light of this theoretical background, a possible and plausible interpretation would be that at a young age it is still more important to orient oneself and explore different environments by switching departments (yet not too frequently), while at an old age – after sufficient orientation – becoming a specialist supports learning.

While the survey data does not entirely support this hypothesis, the principal insight from the data is nevertheless that personal job history in combination with age has a strong effect on learning. Yet the evidence is not sufficient to understand how the three

variables interact.

Consequently, future research would need to go one step further and gain a deeper understanding of the detailed mechanisms of how personal job history interacts with age and affects learning. Such a research step would require a suitable categorization and operationalization of personal histories, which will likely be a difficult task and is probably the reason why it has not yet been done.

In summary, even though the detailed mechanisms of personal history that affect learning require further investigation (possibly with different survey constructs), personal history has emerged from the the statistical results as a strong predictor for learning, which is in line with the theory behind the PIA-model.

7.3.5. Learning Barriers Variable Group

Statistical Results – Learning Barriers Other questions in the survey address barriers to learning specific to the participant’s task. All questions start with *“I could have learned more...”* and end with:

- *“... if I could have found more written information”* (`lbo.findInfo`).
- *“... if more written documentation had existed on the topic”* (`lbo.infoExistant`).
- *“... if I had found a competent contact to discuss this topic”* (`lbo.foundContact`).
- *“... if I had the chance to talk to a competent expert in time”* (`lbo.expertAvail`).
- *“... if I had known how to approach the topic”* (`lbo.approachClear`).
- *“... if I had the chance to experiment more”* (`lbo.moreExperiments`).
- *“... if the information had been reliable”* (`lbo.infoDependable`).
- *“... if I had had more time”* (`lbo.noTime`).
- *“... if I had more measurements from previous projects”* (`lbo.infoMeasured`).
- other barriers (`lbo.otherBarrier`).
- no learning opportunity existed (`lbo.noLearnOoport`).

The learning barriers are used in their raw form without any mean correction⁴².

⁴²Of the presented learning strategies, most participants will have used all of them to some extent, so the *relative* frequency of learning strategy use is of central interest, and therefore a mean correction for the answering bias of each participant is desirable. In contrast, the question items regarding learning barriers aim more at detecting whether a particular barrier has been a hindrance or not. Therefore the absolute (non-mean corrected) answers are of primary interest.

7.3. By Variable Results and Interpretation

Similar to the preliminary analysis of the learning strategies and leadership, the profile of learning barriers was plotted for a “high learner” and a “low learner” group in figure 7.13. Again, the red and black arrows in the right graph represent the mean value of the “low learners” for direct comparison with the “high learners” from the left graph.

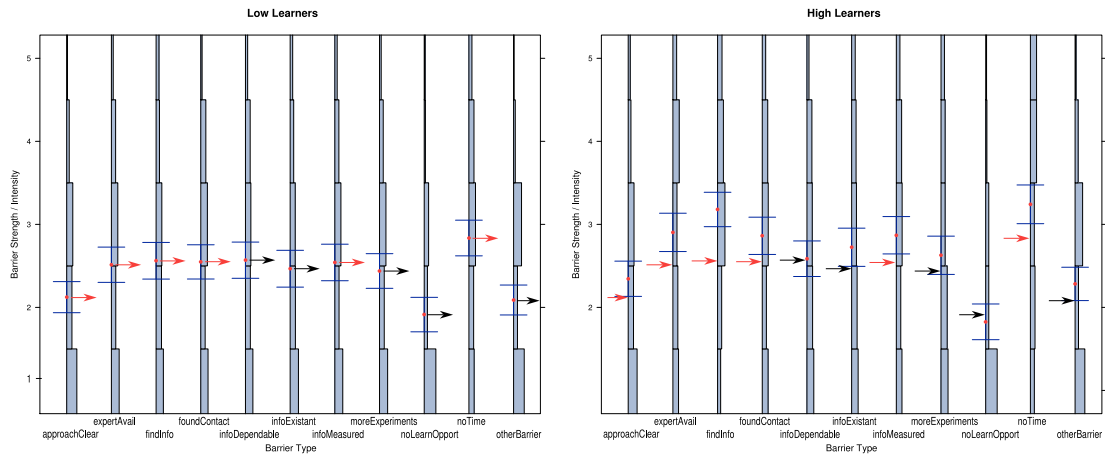


Figure 7.13.: **Learning Barrier Profiles Compared** for Participants with a High and a Low Learning Index. For comparison, the arrows in the graph of the ‘high learner’ group on the right indicate the values of the ‘low learner’ group from the graph on the left. Leaders of the ‘high learners’ on average more strongly encourage initiative, provide feedback and support group climate and trust (see red arrows).

The red arrows point to substantial differences between the low- and high-learner group for the following learning barrier dimensions (ordered by the magnitude of difference):

1. Information Not Found [lbo.findInfo]
2. Not Enough Time [lbo.noTime]
3. More Measurements Necessary [lbo.infoMeasured]
4. Expert Not Available [lbo.expertAvail]
5. Contact Not Found [lbo.foundContact]
6. Approach Not Clear [lbo.approachClear]

Since some of these barrier dimensions are related by topic (e.g., information not found, contact not found, expert not available), a further analysis was performed to determine whether a new aggregate variable, composed from a specialized combination – similar to the learning strategy or leadership profile approaches in the previous sections – would

yield a better predictive power. Hence the construction of a matching special profile scale, containing only the above dimensions, was attempted. However, this combination of dimensions did not add much predictive power to the model.

Instead it was found that the variables:

`lbo.approachClear` “[More learning] ... if I had known how to approach the topic”,
`lbo.foundContact` “[More learning] ... if I had found a competent contact to discuss this topic” and
`lbo.expertAvail` “[More learning] ... if I had the chance to talk to a competent expert in time”

correlate highly, and the interaction of these three variables added predictive power to the model. Fitting a model with this three-way interaction and the respective main effects further indicated that it was mostly the interaction that added predictive power and not any of the main effects, which were rather unstable (by parameter instability) and weak (parameter values close to zero). Therefore the main effects were omitted from later models, and the resultant variable “group” `grp.ApproachClear` in the final model contains only a single variable: the three-way interaction.

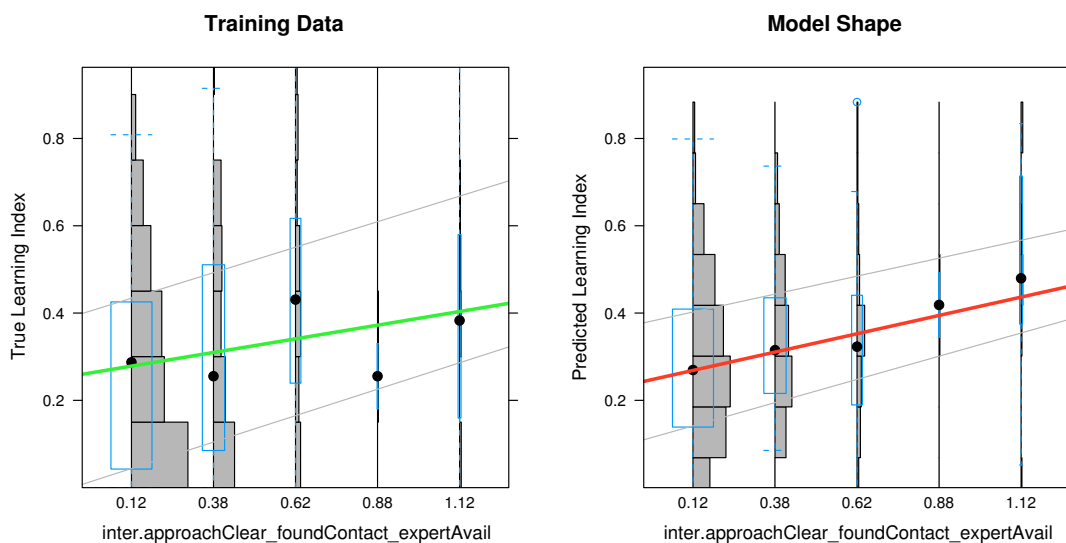


Figure 7.14.: A **Positive Association** of the interaction **Approach Not Clear & Expert Not Found or Available with Learning** shown with a Model Shape Graphic (see p. 215).

The model shape graphic for the three-way interaction in figure 7.14 shows a strongly positive effect on learning. Thus the joint occurrence of “approach not clear” and the lack

of discussion with a knowledgeable expert is on average *associated* with a higher learning effect. The direction of the *causal* effect needs to be determined with reference to theory.

Result Interpretation – Learning Barriers Rather surprisingly at first sight, participants who encountered many barriers to learning actually learned more. This observation can also be expressed in other words: the participant’s greater effort at learning led him/her towards more barriers to learning. The latter interpretation of the results is consistent with the learning feedback effect outlined in section 2.3.6 on page 47: The ‘high-learner’ group has gone through a number of successful learning iterations, leading to a high learning effect. During a learning episode, the momentum for iterating in the learning feedback cycle⁴³ was supported by the progressive learning effect, which has maintained or even improved the high-learner’s high learning motivation and high search effectiveness (section 2.3.7 on page 51) – leading to an even further improved learning effectiveness.

At some learning level, this virtuous spiral meets new barriers that were not significant at a lower learning level. A behavior that limits an upward-spiraling feedback process is very common for feedback systems in other fields: in many engineering applications, such limiting effects occur due to an increased resistive force, while biological systems frequently run into new bottlenecks. For example, a plant’s growth may be limited by the amount of sunlight it receives. But once the plant receives enough sunlight, it will not have an infinite increase in growth. Growth will instead be limited by a new bottleneck, such as a scarcity of nutrients in the ground.

In the case of learning, the statistical results suggest that once learning gains momentum, it is limited by new bottlenecks – listed here along with literature references supporting the respective claims:

1. limited access to information (section 2.3.2 on page 30),
2. lack of time for further investigation (Salter and Gann, 2003),
3. limited access to knowledge through a social network (lack of contacts and access to specialists) (D’Eredita and Barreto, 2006b; Dodgson et al., 2007; Haas and Hansen, 2005; Sandow and Allen, 2005),
4. and a lack of a suitable perspective on the problem (it was not clear how to approach the topic) (Badke-Schaub et al., 2007; Weick, 1993).

If these barriers are only reached after other factors (such as motivation or personal learning predisposition) have propelled the learning loop to a high level of learning, the direction of the causal effect is at first⁴⁴ *from* the high level of learning *to* the learning

⁴³‘Learning feedback cycle’ refers here to both the internal and external feedback loops of the PIA-model – as described in section 2.3.6 on page 47.

⁴⁴While approaching a high learning level.

barriers – i.e., encountering these barriers is caused by the high learning level. Once a high learning level has been reached, the causal direction will reverse, and the barriers will negatively act on learning intensity and, with it, limit the learning spiral.

Given these results, future research should focus on two issues:

- How can organizations be designed, and leaders act, in order to move more employees from the group of low-learners to the high-learner group⁴⁵? How can more employees gain enough learning momentum in order to reach a high learning level and begin to encounter the selected learning barriers presented above? (For factors driving the learning loop, see also other variables included in the survey – e.g., personal interest in the learning topic in section 7.3.3 on page 238.)
- Once the barriers are encountered, what are the barriers that limit the high-learner group in more detail? How do these barriers limit learning? How can organizations eliminate the barriers or mitigate their effect?

The fact that the interaction of ‘*Approach Not Clear*’, ‘*Expert Not Found*’ and ‘*Expert not Available*’ adds predictive power to the BOGER model implies that this particular combination of barriers is a particularly good predictor of learning. This combination further suggests that the learner is faced with a particularly difficult and open problem that requires a challenging refinement of the learner’s perspective on the problem – which was (unsuccessfully) attempted by seeking discussions with experts. Yet for now this statement has to remain a hypothesis that requires further detailed investigation in future research.

In summary, the learning barriers captured in this survey have been found to become relevant learning bottlenecks only after a high level of learning effectiveness is reached. Thus organizations should first concentrate on the factors that get the learning momentum going and only then focus more attention on learning barriers, such as access to information and experts as well as sufficient time.

7.3.6. Epistemological Beliefs about Learning

Statistical Results – Epistemological Beliefs The final BOGER model contains two variables regarding the epistemological beliefs (EÜ⁴⁶) about learning, collected by using the two following scales:

- **EÜ Learning Process (EU_1rnProc)** consists of the question items with five-level answering scales (disagree ... agree):

⁴⁵First, the learning level at which the other barriers become significant bottlenecks would need to be investigated. Possibly the barriers, as discussed here, only gradually become more dominant, like air resistance eventually dominates the rolling resistance of a car.

⁴⁶The German acronym *EÜ* stands for “Epistemologische Überzeugungen”, or epistemological beliefs.

7.3. By Variable Results and Interpretation

- “When I am learning, I am mostly passive.” (Neg.)
- “Learning means solving problems.” (Pos.)
- “While learning, one needs to concentrate on the most important issues.” (Neg.)
- “To be able to learn something, one needs a teacher.” (Neg.)
- “One can also learn without instruction.” (Pos.)

This scale gives high scores for more differentiated notions of the learning process that go beyond old-style school- or seminar-like settings and towards more active forms of learning – such as learning during problem solving.

- **EÜ Internal Social Learning** (EU_lmeth.IntSoc) consists of the question items with five-level answering scales (disagree ... agree):

- “At the workplace one can learn well for the job.”
- “From colleagues one can learn well for the job.”
- “From family members one can learn well for the job.”
- “From friends one can learn well for the job.”
- “From superiors one can learn well for the job.”
- “Alone one can learn well for the job.”

This scale gauges the notion of social learning, i.e., whether the participant counts discussions with others as an important part of learning.

Figure 7.15 on the facing page shows the overall effect of these two scales on learning:

- A differentiated notion of the **learning process** (EU_lrnProc) on average leads to a higher learning index.
- A more **social notion of learning** (EU_lmeth.IntSoc) leads to a lower learning index.

Result Interpretation – Epistemological Beliefs In two experiments, Schommer (1990) observed that epistemological beliefs affect the nature of knowledge and the comprehension of academic subject matter by college students. Hence this effect might also be present in work settings.

Epistemological beliefs are closely linked to an awareness about the learning process, e.g., that learning is an active endeavor (section 2.3.1 on page 28) and that non-traditional problem-centered forms of learning support learning more strongly. However, a lack of this awareness does not directly imply a reduced learning intensity. The participant without this awareness may simply be less aware of the learning effect while working on his or her task but may still learn useful lessons.

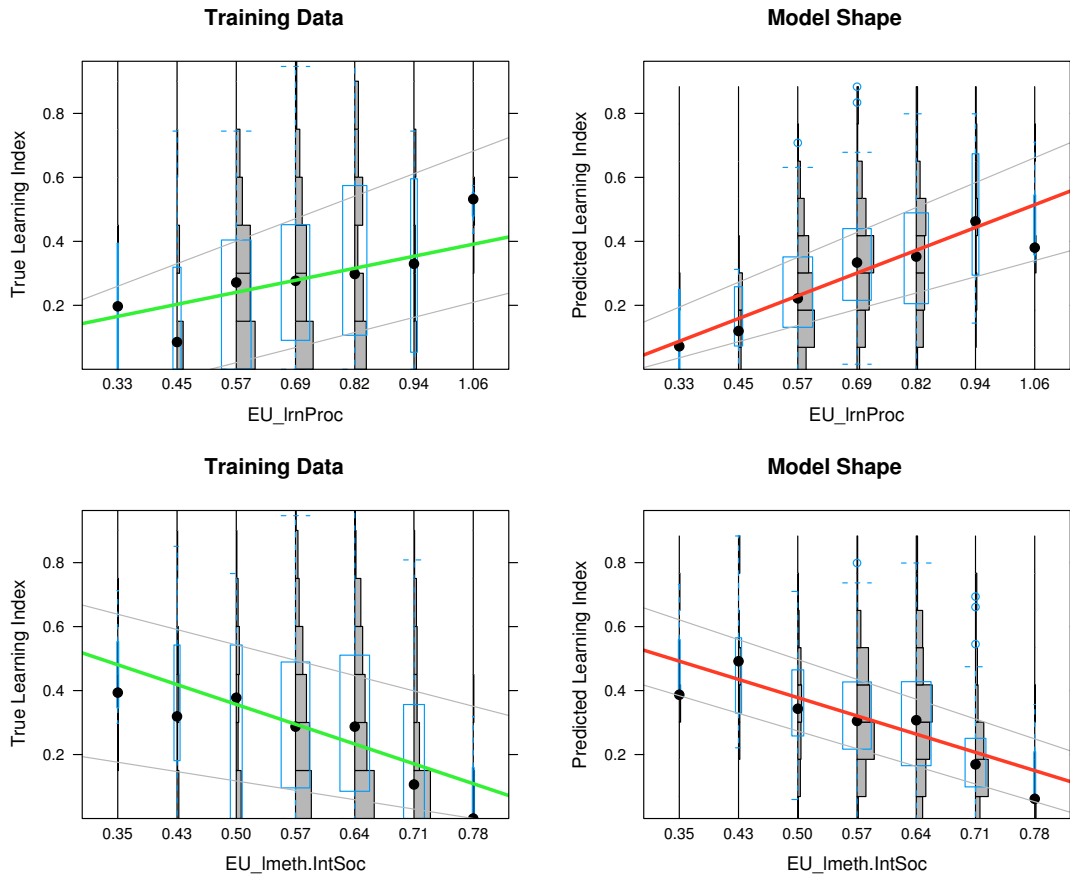


Figure 7.15.: A **Positive Association** of the **Learning Process-oriented Epistemological Beliefs with Learning** and a **Negative Association** of the **socially oriented Epistemological Beliefs with Learning** shown with a Model Shape Graphic (see p. 215).

Thus two reasons for the effect of a differentiated awareness of learning processes on the learning outcome are conceivable:

- A more differentiated awareness of the learning process may be associated with a more reflective approach to learning and allow systematic improvement (Dahl et al., 2005)⁴⁷. Only the conscious awareness of an ongoing learning process allows a conscious reflection on this process, involving an assessment of its efficacy and an iterative improvement (see section 2.3.6 on page 48).
- A wider notion of learning processes will cause the participant to classify a larger

⁴⁷ “As hypothesized, the more students believe that learning ability is fixed, the fewer the strategies they report using to connect their prior knowledge with new knowledge that is to be learned, or to think critically about the information that they are processing”, Dahl et al. (2005, p. 269).

number of events that occurred while working on the task as learning episodes. In other words, a wider understanding of the question used to calculate the learning index will also lead to a higher learning index, without necessarily actually increasing the learning effect.

Hence `EU_lrnProc` may cover the real effect of reflective learning but may also be related to a possible bias in the learning index.

Figure 7.15 on the previous page further suggests that a more social notion of learning (`EU_lmeth.IntSoc`) leads to a lower learning index. Inspecting the real data, it becomes evident that most of the collected samples are clustered around numerical values corresponding to the scale points “neutral” and “agree”, which suggests that this result is not as robust as the result for `EU_lrnProc` with a more uniform data density – see section 4.2.2 on page 123.

Yet combined with the results regarding learning strategies (section 7.3.1 on page 226), which suggest that *discussion* alone is not as helpful as *discussion* combined with other learning strategies, such as *reading*, a negative effect of an overly strong focus on social learning is plausible.

Given its low statistical robustness, the results from the social epistemological belief scale are not further considered.

In summary, both epistemological belief variables appear to have a strong effect on learning. However, given the open questions surrounding the links of the two variables with artifacts of the learning index and low robustness of the statistical results, further research for clarification is required, and thus this result was not included in the implications.

7.3.7. Task Type Variable Group

Statistical Results – Task Type The variable group `grp.TaskType` contains effectively three dummy variables indicating what type of project or task the participant chose as a concrete example during the survey.

- Innovations (`isInnovation`)
- Larger (and Longer) Projects (no dummy variable included, since when the two other variables are zero, then the task branch must be “larger projects”) (`isLargePrj`)
- the Normal Work of the Past Four Weeks (`isNormalWork`)

See also the details on the survey *task branch* in section 5.7 on page 153.

Figure 7.16 on the next page shows a positive effect on learning when an innovation project was chosen, hardly any effect for large projects and a negative effect for the normal work of the past four weeks.

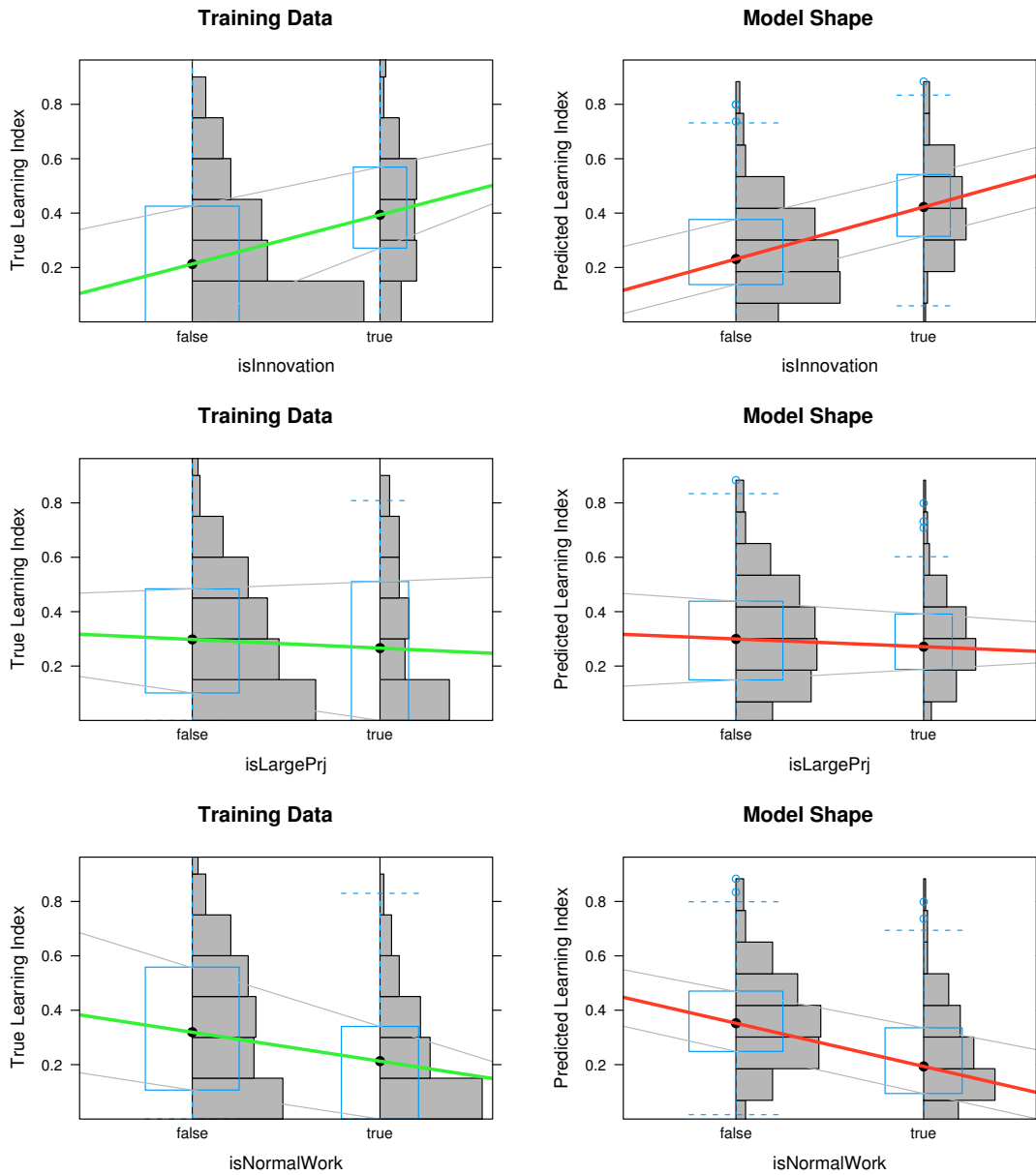


Figure 7.16.: A **Positive Association** of Innovation Tasks with Learning and a **Negative Association** of Normal (Non-Project) Work with Learning shown with a Model Shape Graphic (see p. 215).

Result Interpretation – Task Type These statistical results support the insights from literature: In their case study at the civil engineering company Arup, [Salter and Gann \(2003\)](#) found that more openly defined projects provide more opportunities for innovations rather than projects following a very detailed specification – with most of the solution already described or implied in the specification.

In addition, the statistical results could also be caused by correlations with other variables. As described in section [5.12.1 on page 163](#), employees with a higher education level are more likely to participate in larger projects, including innovation projects. This correlation may suggest that education level acts as a latent variable on learning and that the task branch derives its predictive power only from the correlation with education level. However, since the interaction of education and task branch was not found to add predictive power to the model, there is no evidence that supports the hypothesis that education level acts as a latent variable here.

In summary, the type of task affects learning. Of the three different task types, innovation projects by their very nature lead into uncharted territory and thus will pose many inspiring challenges that require finding innovative solutions and thus also lead to a strong learning effect.

7.3.8. Description Detail-Level of Procedures

The variable `proc.TaskDetail` gauges the level of detail of the instructions for the participant's task. The participant is asked the following question:

“How accurately do working instructions describe how the work for your task needs to be done?” Answering scale: (“very inaccurate” ... “very accurate”)

This question is the opposite to asking about the openness of a task.

From theory, there are two plausible ways this detail level of task description may act on learning:

1. Strictly described and thus standardized tasks, when combined with statistical process control measures⁴⁸, open up opportunities for learning from the process data ([European Foundation for Quality Management, 2003](#); [Liker, 2004](#)).
2. Open-ended tasks (or ‘*weak situations*’ ([Mischel, 1977](#))) feature more challenges and opportunities to learn and to innovate ([Salter and Gann, 2003](#)). Following this argument, tasks with a very detailed description (and instruction) are expected to lead to less learning.

⁴⁸A simple variant of statistical process control is tracking key performance indicators (measures to assess the process performance) over time and dependent on changes made in “run chart” to improve the process ([Devor et al., 1992](#)).

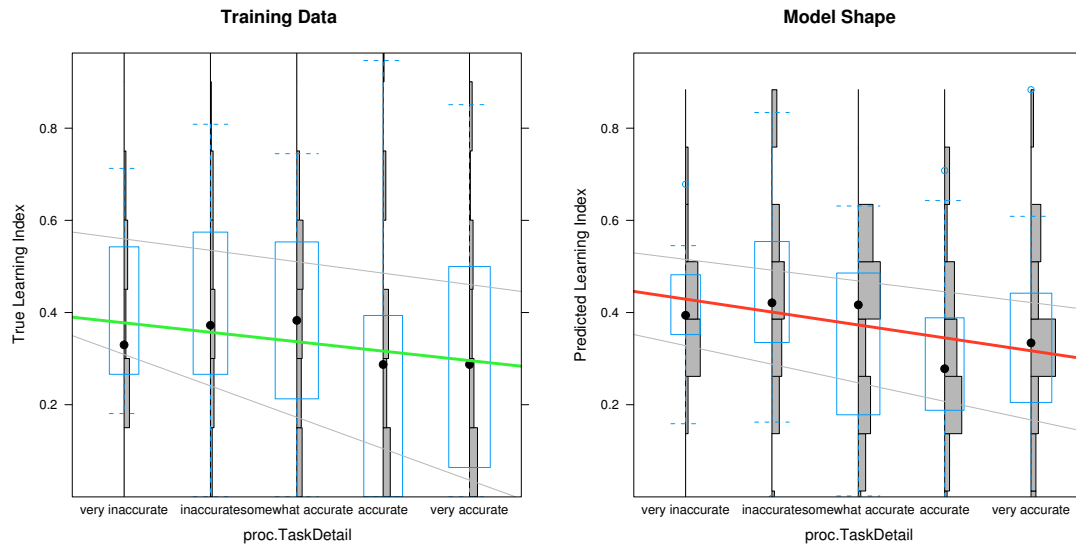


Figure 7.17.: A **Negative Association of Procedural Description Detail with Learning** shown with a Model Shape Graphic (see p. 215).

The statistical results in figure 7.17 suggest an (on average) negative dependency of the learning index on procedural description detail and thus support the second hypothesis.

Further support can be drawn from a question item on task routine level⁴⁹ that was posed for each workstep involved in the chosen example task:

“How frequently have you worked on tasks similar to <the workstep of the chosen example task> in the past?”

Figure 7.18 on the following page shows that a high routine level⁵⁰ decreases the learning intensity – i.e., there is weakly negative average dependency of the learning index, which also supports the second hypothesis.

In summary, the results from the BOGER model and the task detail variable as well as from the routine level variable support the second hypothesis: openly formulated tasks lead to more learning. However, the results do not contradict the first hypothesis: standardization of tasks may still support learning, since task standardization only becomes learning-effective when used in combination with key performance measures and statistical process control – which might not have been the case for the participant’s task⁵¹.

⁴⁹The task routine level could not be added as a variable to the BOGER model in a sensible way, since the routine level has been surveyed for each workstep (and person), and an average for each person across multiple worksteps is not very meaningful.

⁵⁰The figure shows the raw routine level samples for all worksteps and not for each person in an average.

⁵¹Standardization of process steps and close monitoring of the process within control limits has been proven to be very useful in many industrial applications (Devor et al., 1992). However, many of the

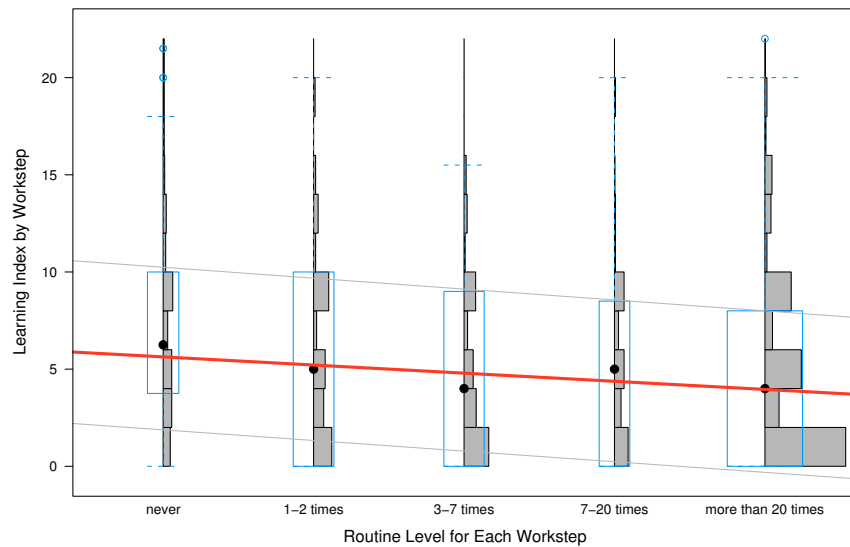


Figure 7.18.: Learning Index vs. Routine Level for each Person and Workstep – suggesting a **Negative Effect of Routine on Learning**.

7.3.9. Number of Seminars

The survey asked participants about the number of seminars they have participated in during the past year. This variable, `n.seminar`, shows a strong positive effect on learning – see fig. 7.19 on the next page.

A number of explanations for this association of number of seminars with learning are plausible:

- Seminars provide a theoretical background that also facilitates on-the-job learning when the new theoretical knowledge is transferred to practical application after participation in the seminar. This would be a direct effect.
- Learning skills and learning self-efficacy are maintained at a higher level when employees regularly participate in seminars (Roßnagel, 2008; Winne, 1995). This increased or at least maintained learning skill also has a direct and positive effect on on-the-job learning.
- Participation in seminars could be driven by a generally learning-supportive work

manufactured parts and steel assemblies at Meyer Werft change with every new product that needs to be manufactured. Thus standardization is frequently possible, but the comparison of the process data is difficult – yet not impossible with suitable statistical models. Given these increased challenges with applying statistical process control (SPC) in cruise ship construction, it is not used for all process steps at Meyer Werft. Nevertheless, even without SPC, task standardization may be useful, not primarily as a learning support but for increasing and controlling process stability and quality.

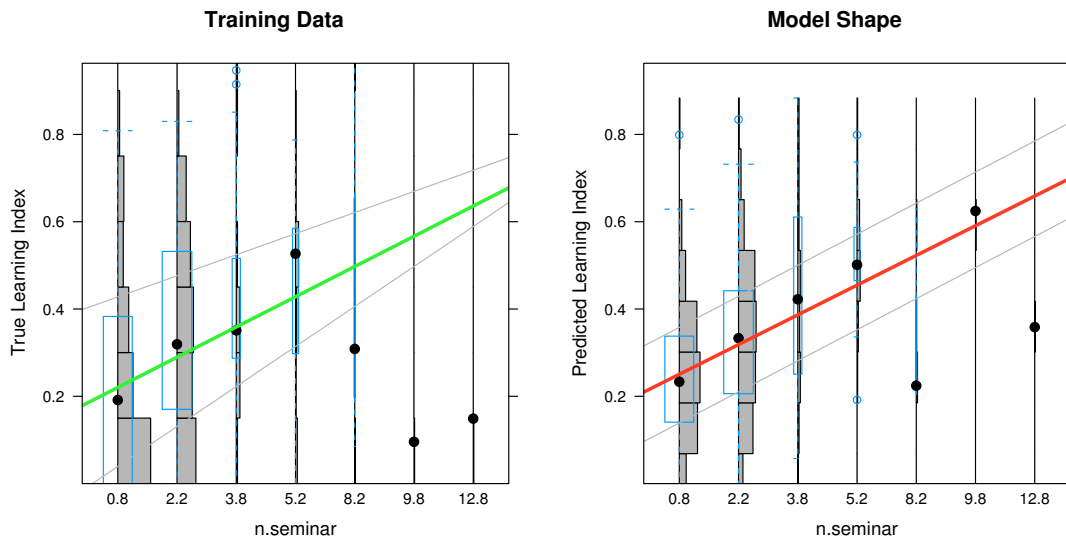


Figure 7.19.: A strongly **Positive Association** of **Number of Seminars** with **Learning** shown with a Model Shape Graphic (see p. 215).

environment or by the participant’s motivation to learn. In these cases, it is not the seminar participation that acts on learning but instead a latent variable (e.g., motivation or a learning-supportive environment) that drives learning. Thus the variable `n.seminars` is only by correlation a good predictor for learning. (Education level, which correlates mildly with `n.seminars`, can be excluded as a latent variable since it was tested as a separate variable directly in the BOGER model.)

In summary, the survey data has evidence for a spillover effect from formal learning to informal on-the-job learning. However, the survey data is not sufficient to gain further insights into the underlying mechanisms – which would require further research with other methods.

7.3.10. Job Closure

The standard job description scale (JDS) inventory Kulik et al. (1988) contains a scale that describes job closure, i.e., whether an employee, as part of his or her job, follows the entire process of a task or only sees a small fraction of it. The scale consists of the following question items⁵²:

- “To what extent does your work contain holistic, self-contained tasks?”

⁵²For participants in the innovation and large project task branches, these questions were slightly adapted to refer to the task instead of the participant’s entire job.

7.3. By Variable Results and Interpretation

- “My works gives me the opportunity to bring all started tasks to a finish.”
- “My task is designed in such a way that I have the opportunity to work on a task from start to finish.”
- “My task is designed in such a way that I do not have the opportunity to work on a task from start to finish.”

As shown in figure 7.20, job closure shows a weakly positive effect on learning.

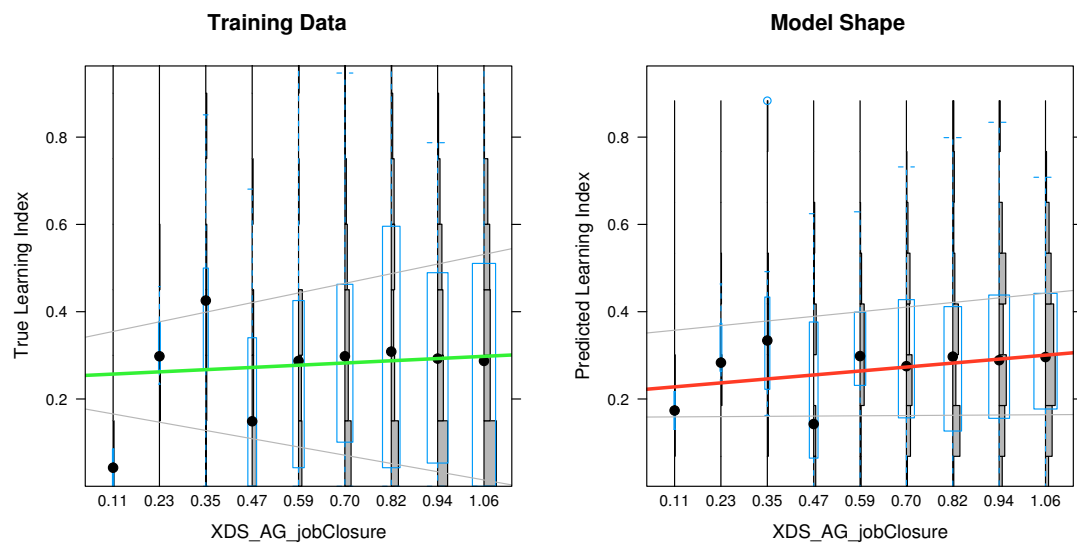


Figure 7.20.: A **Positive Association of Job Closure with Learning** shown with a Model Shape Graphic (see p. 215).

In line with the PIA-model and other sources (Cabrera et al., 2006; Deming, 1985; Liker, 2004), involving employees in holistic improvement efforts leads to a better understanding of the processes and thus offers more opportunities for learning, innovation and process improvement. Thus one of the best ways to support employees in finding new solutions is to allow them to understand the entire process by having them work on all steps of the process⁵³.

In summary, the statistical results as well as insights from literature support the claim that job closure (i.e., working on a bigger task from start to end) facilitates on-the-job learning.

⁵³This insight is also in line with Karl Marx’s claim that a Tayloristic disintegration of work into separated and isolated worksteps leads to an alienation of workers from their work.

7.3.11. Openness to New Experiences (Big Five)

A commonly used scale inventory for personality are the *Big Five* by John et al. (1991), consisting of the following personality dimensions:

- **Extraversion** Extraversion encompasses traits such as talkative, energetic and assertive.
- **Agreeableness** Includes traits such as sympathetic, kind and affectionate.
- **Conscientiousness** Includes traits such as organized, thorough and planned.
- **Neuroticism** Emotional instability. Includes traits such as tense, moody and anxious.
- **Openness to Experience** (Also called Intellect or Imagination.) Includes traits such as having wide interests, and being imaginative and insightful.

This survey used a shortened version, the NEO FFI (Rammstedt and John, 2007), in a German translation (Borkenau and Ostendorf, 1993).

Only the dimension “*openness to new experiences*” was found to add predictive power. The corresponding scale variable `bfi.open` consists of the following question items⁵⁴:

- “*I have only limited artistic interest.*” (Neg.)
- “*I have an active imagination and am imaginative.*” (Pos.)

The strong positive effect that openness to new experiences has on learning confirms the insights from theory well. As discussed in section 2.3.5 on page 45, the perspective-refining opportunities that derive from considering other people’s perspectives require a minimum level of (critical) openness towards others’ ideas in order to be realized. This perspective-setting effect is particular and unique to the openness dimension of the Big Five survey tool. Future research should investigate how much of the effect of openness on learning is mediated via perspective setting and how much directly acts on learning. Further support from literature can be found in Cabrera et al. (2006).

In summary, the only (Big Five) character trait that strongly supports learning is openness to new experiences.

7.3.12. Task Difficulty

Task difficulty can be expected to drive learning. In other words, if the task has no challenges, there is no need to solve problems and little need to learn on the job (see also Salter and Gann (2003)).

⁵⁴The NEO FFI question items were used in this unmodified and standard (German language) form from Borkenau and Ostendorf (1993).

7.3. By Variable Results and Interpretation

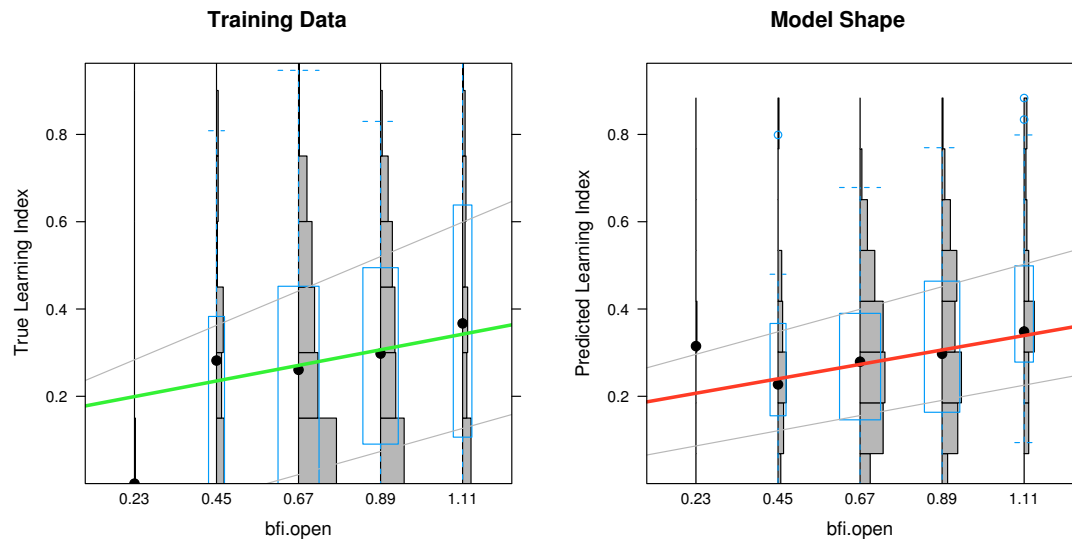


Figure 7.21.: A **Positive Association of Openness to New Experiences** (i.e. Personality) **with Learning** shown with a Model Shape Graphic (see p. 215).

Task difficulty was surveyed by the following question:

“How difficult was this task for you compared to other tasks of your work?”

As expected, figure 7.22 on the facing page shows a strong positive effect of task difficulty on learning.

7.3.13. Fault Culture

If employees fear making mistakes during their work, their willingness for experimentation and open exchange with colleagues will be limited. Therefore an overly strong fear of mistakes, i.e., a bad fault culture, is expected to reduce learning and continuous improvement (Deming, 1985).

The fear of mistakes was surveyed by the following question:

“If somebody tries something and makes a mistake, this can have very serious consequences for that colleague’s career.”

As expected, figure 7.23 on the next page shows a negative correlation between fear of mistakes and the learning index.

Hence less fear of mistakes, i.e., a suitable fault culture, supports learning.

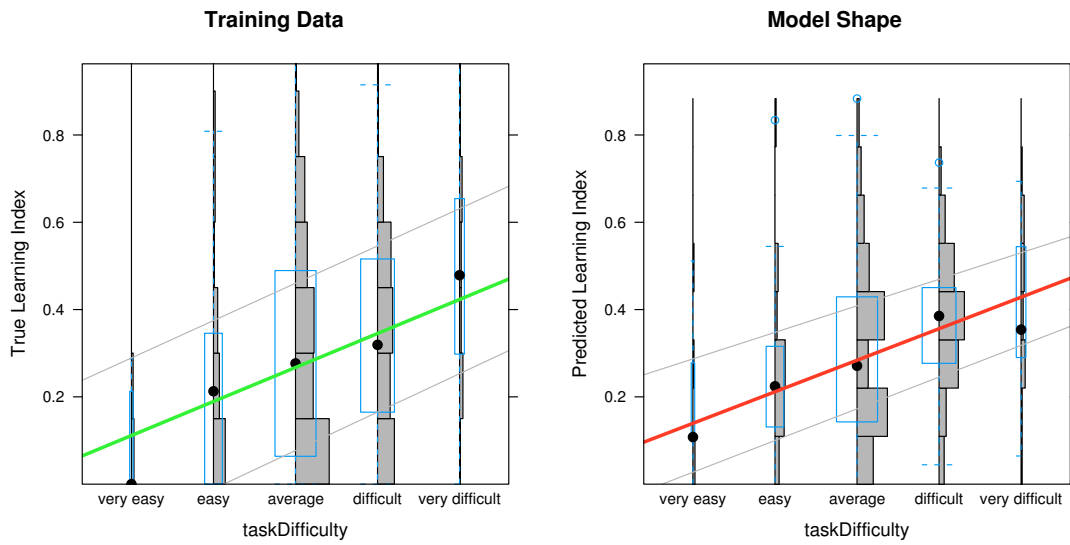


Figure 7.22.: A strongly **Positive Association** of Task Difficulty with Learning shown with a Model Shape Graphic (see p. 215).

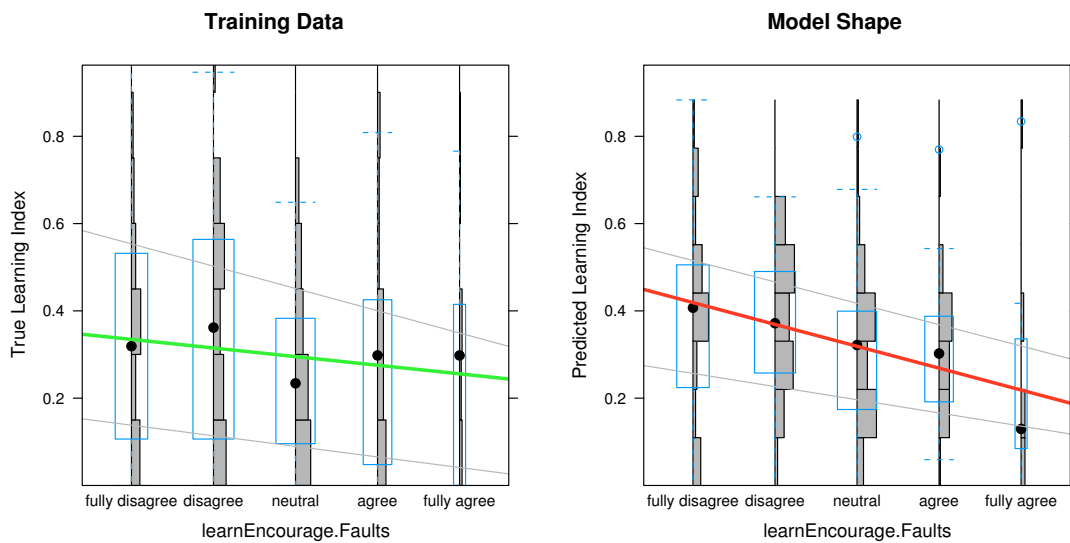


Figure 7.23.: A **Negative Association** of Fear of Mistakes (i.e., Fault Culture) with Learning shown with a Model Shape Graphic (see p. 215).

7.3.14. Surprisingly insignificant Factors (Non-Factors)

The previous sub-sections discussed only the positive statistical results. However, the non-results are also insightful⁵⁵ – i.e., factors that theoretically should support learning but did not add predictive power to the BOGER model and were therefore left out.

The following factors surprisingly did not add predictive power:

- Communication (i.e., discussions) greatly support learning (section 7.3.1 on page 226), and social networks are important for searching (section 2.3.7 on page 51) and diversity of perspective (section 2.3.5 on page 45). Not surprisingly, in the literature the total **number of contacts** has been observed to support learning, depending on the complexity of the subject (Hansen, 1999)⁵⁶. Hence the number of personal contacts as well as the intensity⁵⁷ of the relationship was surveyed for. However, none of these variables or their interactions added predictive power to the BOGER model.

As Hansen (1999) confirms, an assessment of a social network's value by the number of the learner's direct contacts is too simplistic and does not have much predictive power. Therefore future research would need to use more refined questions regarding the social network than the one used in this survey. Given the strong evidence in literature for the importance of social networks, networks of personal contacts should not be left out in future research efforts on on-the-job learning.

- **Age** was not a factor inhibiting learning (see section 7.3.4 on page 240) – in contrast to popular “wisdom” that older employees have reduced learning skill.
- The theory behind the PIA-model suggests that diversity in perspectives adds to learning (section 2.3.5 on page 45). Thus one might expect that employees who had other professional experiences before starting to work at the shipyard would have a broader perspective and thus learn more. Despite a careful investigation during the model building process, having additional **external professional experiences** in other companies did not contribute to the participant's learning.

It remains unclear, however, whether the diversity of perspectives of a participant with external professional experience supports learning among an entire group of employees (Kearney et al., 2009).

⁵⁵Looking for non-results, i.e., cases where a current problem does not occur, is also a common technique in quality management. See, e.g., Ford's Global 8D (eight-disciplines) problem-solving technique, which requires finding all cases where a defect does not occur in order to gain a deeper understanding of the current problem (see (Al-Mashari et al., 2005) or http://en.wikipedia.org/wiki/Eight_Disciplines_Problem_Solving).

⁵⁶The empirical results of Hansen (1999) are discussed in section 2.3.5 on page 44.

⁵⁷A question item surveyed for frequency of contact as a very primitive measure for the concept of *strong links* in Hansen (1999), which refers to a common understanding and shared meaning of language among the participants.

- As discussed in section [5.12.1 on page 163](#), **education level** was not found to be a strong driver of learning – which may also be connected to the fact that the learning index measures *self-assessed* learning intensity.
- Despite some indications from other studies ([Cabrera et al., 2006](#)), **personality traits**, as measured by the psychometric construct of the Big Five, did not strongly affect learning intensity – with the one exception of *openness to new experiences* (section [7.3.11 on page 257](#)).

8. Summary, Implications, Limitations and Future Research

Chapter Contents

8.1. Summary of Research Findings	263
8.1.1. Principal Insights from Literature	263
8.1.2. Results Overview	267
8.2. Practical Implications	270
8.3. Relevance of the Results to Literature	275
8.3.1. Relevance to Organizational Learning	275
8.3.2. Relevance to Industrial Practice Models	276
8.3.3. Relevance to Sense Making, Problem Solving and Knowing	277
8.3.4. Relevance to Statistics	277
8.3.5. Relevance to Knowledge Management	278
8.3.6. Selection of a Few from Many Plausible Explanations	279
8.4. Limitations	280
8.5. Areas for Future Research	282

8.1. Summary of Research Findings

8.1.1. Principal Insights from Literature

As described in section 2.2 on page 23, the literature search yielded a number of important insights from different research areas on which this study was based.

Learning, cognitive processes, knowing, individual and group decision making, and problem solving are all different and valuable perspectives on knowledge-intensive work and are addressed in a large number of different research fields. The different perspectives are shaped by the use of different research paradigms, by different basic assumptions about

8.1. Summary of Research Findings

science and by different terms. Given the differences in language and community, results and insights from one field do not automatically propagate to all other fields (which is also a knowledge management problem). This effect is further strengthened by the increasing importance of search terms in literature databases. While research in these fields comes to many similar conclusions, it also produces contradictory results – contradictions, which may even prevail over time.

As discussed in section 2.7.2 on page 80, this study's research focus is on informal individual on-the-job learning while solving job-related problems. As detailed in section 2.6 on page 77, individual learning is in this case a part of organizational learning.

Relevant insights from literature, presented in theory section 2.3 on page 28, were condensed in the PIA-model (figure 8.1). Further insights from literature were added in the detailed discussions of the results in section 7.3 on page 226.

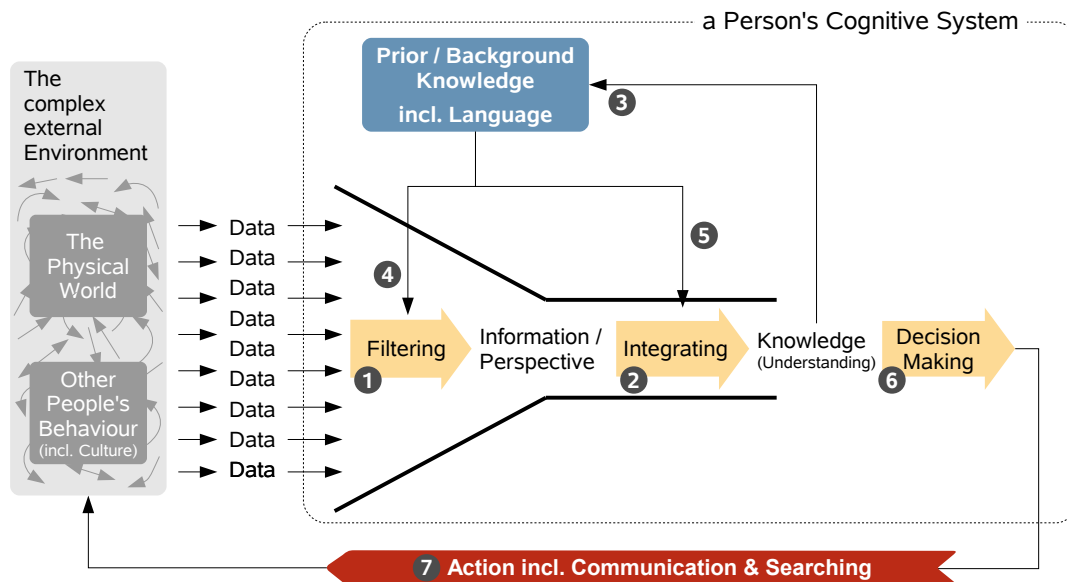


Figure 8.1.: Relevant **Literature Insights Condensed** in a Single Model: The Perspective Taking / Action / Integration (PIA) model [figure 2.1 on page 31 repeated for convenience] (Source: Author)

The PIA-model reflects the following general insights, which form the basis of this study:

- Before a workplace problem can be solved, a knowledge worker first has to make sense of the situation: humans are constantly faced with a very large stream of *data*, which includes any visual, auditory or other sensory input and which may or may not be relevant to the problem at hand.

Since we as humans have limited cognitive resources, the stream of data is in most cases too large to handle directly, and thus the data stream needs to be **filtered** down to *information* – a set of features that an individual person considers relevant for the application at hand. It is this skill for complexity reduction that allows us humans to behave effectively in a world that is too complex for our cognitive resources (section 2.3.2 on page 32).

In section 2.3.2 on page 30, this filtering step (step 1 in the PIA-model, figure 8.1 on the facing page) is described as ‘**perspective taking**’ – following Orr (1996)¹. Hence by filtering we assume a complexity-reducing perspective on a complex situation, which highlights certain aspects while hiding others².

- Next, the filtered information needs to be **integrated** to a sound judgment that sufficiently explains the situation or problem (step 2 in the PIA-model)³. Such explanations can be compared with scientific models: they explain complex phenomena in a simplified or approximate, and hence never perfect, manner.

Finally, with an understanding of the problem, new knowledge is created – or, in other words, a **learning** episode occurs. This new knowledge can then be the basis for decision making (step 6 in the PIA-model) or can become part of the stock of prior knowledge for later use (link 3 in the PIA-model). For details, see section 2.3.3 on page 34.

Knowledge is then the result of integrating new information into a person’s individual web of prior knowledge. This definition draws a fairly clear distinction between information and knowledge – as argued in section 2.5.5 on page 76. It also implies that only information can be captured and stored in documents⁴.

- The data filtering and integration process partially depends on a person’s **prior knowledge** (links 4 and 5 in figure 8.1 on the facing page), which has been built up over many episodes of experience (section 2.3.4 on page 41). Thus, on the one hand, newly created knowledge is the basis for decision making, yet on the other hand, it

¹Orr (1996) presents an ethnographic study showing how copy machine technicians constantly challenge and shape their perspective on a problem by telling narratives of past repair jobs until they find a perspective that fits with the current feedback they have received by inspection and exchange of spare parts, until they can integrate all the facts into a coherent explanation.

²Language can play an important role in this filtering and aspect-highlighting process as a tool for thought – see section 2.3.5 on page 43.

³Here the recognition of relevant features and the integration of features are modeled as two discrete steps for illustrative purposes, though technically these two stages are a continuum of a single cognitive process beginning with feature recognition and ending with integration of features.

⁴The reader may think of a university textbook on the shelf. Just because a person owns the textbook, does that imply that he or she has the knowledge? I argue here that most of the textbook’s value only becomes realized when the person actively attempts to understand the arguments presented in the text. According to my definition, knowledge is only created after the active and personal information integration process.

also sharpens one's view: the newly added knowledge has a **feedback**⁵ **effect** by refining the filtering process and thus further improving filtering and new knowledge creation (section 2.3.6 on page 47).

Since most people live in constant interaction with others, they share their views and even some experiences, which leads to a limited alignment of their prior knowledge and thus also their perspectives. Hence a part of their prior knowledge will be socially constructed – which includes the implicit rules and filtering schema that constitute societal as well as **organizational culture**⁶ (section 2.3.4 on page 42).

This dependency of the data filtering process on prior knowledge, and indirectly on personal history and organizational culture (link 3), makes it a personal and **subjective** process, since it is a complexity-reducing step, which leads to a simplified and possibly biased representation of the actual world. Hence the selection of relevant information from a large stream of data is a subjective process, which is this the basis for learning and decision making.

- Given these mechanisms of cognition, knowledge transfer involves learning and possibly also teaching. The PIA-model illustrates that humans cannot directly import knowledge from other people. Instead, they have to process any data they receive from other people by filtering it first and then actively integrating the extracted information into their individual and historically grown structure of prior knowledge (steps 1 – 3 in figure 8.1 on page 264). Hence this process cannot be forced and instead other people can only support the learner's individual knowledge creation effort in the filtering step by pointing him or her to relevant information⁷.

Hence **knowledge transfer** is an **active endeavor**, mostly on behalf of the learner. The teacher or the employee sharing knowledge can 'only' create a suitable environment and example situation to support the learner (see section 2.3.1 on page 28).

Since in most cases the biggest challenge is **learning**, the learning process is the focus of this study. This insight also has direct implications for organizations. For example, people and the organizational environment, rather than databases, should be at the center of any knowledge management initiative.

- Filtering and thus also decision making can be improved by **visualization** (section 2.3.3 on page 35). Visualization in this context can be the result of a manual

⁵The PIA-model shows an inner feedback loop (without external feedback) in steps 1, 2, 3, 4 and 5, in addition to an outer feedback loop (with external feedback) in steps 1, 2, 6 and 7 – see also section 2.3.6 on page 47.

⁶This includes aspects such as fault culture.

⁷This is the reason why the term *knowledge management* is misleading on the individual level as a unit of analysis. Knowledge in people's heads (as defined in section 2.6 on page 77) cannot be managed directly. That limits active management to shaping and optimizing the environment and organizational processes to support a learning activity that by itself cannot be forced.

process, such as a sketch, as well as the result of a computer automated process – such as the model shape graphics (figure 7.2 on page 215).

- **Diversity of perspectives** supports learning, since the diversity can be used to create a single common and shared perspective, which, as a product of intense debate and discussion, is more robust than isolated individual perspectives. Conversely, a variety of diverse yet completely unrelated perspectives is not helpful. A **common perspective** – a shared way to describe the situation with graphics or language – is important to understanding each other, especially across different domains of expertise, and can facilitate a deep and constructive discussion. See theory section 2.5.3 on page 74. For a practical example, see section 2.4.5 on page 66.
- **Tacit knowledge** and implicit learning poses a challenge for knowledge transfer within organizations as well as for research design. In particular, the unconscious nature of tacit knowledge poses challenges. For example, in teacher/learner interactions, the teacher can only give limited learning support to the student, since the teacher cannot verbalize his or her tacit knowledge that complements his explicit knowledge on a topic or skill. Learning becomes even more challenging when the ‘knowledge’ is distributed only in the form of information in documents (section 2.3.8 on page 52).

Nevertheless, implicit learning rarely occurs in complete isolation from explicit learning, and thus a part of the learning effect is consciously noticeable in most cases.

The challenges surrounding tacit knowledge have been considered for this study, but since explicit and tacit knowledge in most cases are created, transferred and used jointly, this study’s research design does not contain an explicit distinction of the two types of knowledge and learning.

It is noteworthy that a large part of knowledge management literature uses a different paradigm: the paradigm of an *object-like* tacit and explicit *knowledge*, which suggests that knowledge is a quasi-tangible form of capital that can be counted, stored, transferred and managed. Although this perspective may be useful when conceptualizing knowledge transfer across corporate divisions, it is not useful for the purpose of this study – as discussed in detail in section 2.5 on page 68

8.1.2. Results Overview

As presented in the theory chapter 2 on page 21, on-the-job learning in an organizational environment is a multifaceted and complex problem with many factors supporting or hindering learning. For an actual organization with limited time and resources, it is not practical to consider and address all of these factors to improve individual and thus

organizational learning. Therefore the interactive survey (chapter 5 on page 137), which is a principal part of this study, was designed to quantify learning intensity for different participants in different working environments with the aim of obtaining a ranking of the most important factors driving or inhibiting on-the-job learning.

The ranking of the strongest factors from the statistical analysis (chapter 7 on page 205) confirms the main elements of the PIA-model developed from theory in chapter 2 on page 21 (the (x)-numbers indicate the approximate importance ranking of the result with respect to its effect strength on learning):

- (1) **Perspective taking** is an important prerequisite step in learning (see theory section 2.3.2 on page 30). The results show that learners most effectively refine their perspectives by comparing their own with the **perspectives of others** (section 7.3.1 on page 226). The data further suggests that perspective refinement is a **stronger** driver of learning **than** successfully **obtaining information** – e.g., due to good searching skills or availability of good information sources. This finding is particularly important and surprising at first sight, since a large fraction of the knowledge management literature is based on the implicit assumption or simplification that there is only a single *true* perspective on a problem⁸.
- (3) **Learning** requires the **active** engagement of the learner in order refine the learner's prior knowledge in **iterative learning feedback** loops (as described in section 2.3.6 on page 47, with confirming results in section 7.3.5 on page 243. Hence also the strong and positive influence of topic-specific **intrinsic motivation** (i.e., personal interest) in learning (section 7.3.3 on page 238).
- (4) A person's **prior** or background **knowledge** is **built incrementally over** many episodes of **experience** (section 2.3.3 on page 34). Consequently, the statistical results show that personal history matters (section 7.3.4 on page 240).

Furthermore, the following additional results are drawn from the statistical analysis:

- (2) **Leadership** is a surprisingly **strong** driver of learning. The data can even isolate a particularly learning-supportive leadership **profile**, including the features of supporting employee initiative, giving feedback, and fostering a group climate and trust, but excluding the classical leadership features of focusing on goals and

⁸Many knowledge database systems are geared towards storing a single version of the true knowledge. However, there are also ways to overcome this limitation. The open web encyclopedia Wikipedia (<http://www.wikipedia.org/>), for example, allows for discussion threads on each topic. The result of these discussions may eventually be consolidated in the main article. Furthermore, the different perspectives presented in the discussion allow the reader to compare the different views and make his or her own judgment, which in most cases will be a better judgment rather than just superficially accepting the single version of the knowledge on the topic presented in the main article.

clear division of responsibilities. An effect of leadership on **work environment** also becomes visible, but the details remain unclear (section 7.3.2 on page 231).

- (5) Limited **access to information** and **expert advice** (e.g., by lack of IT tools for efficient searching or by lack of availability of experts) is a serious challenge but **only for** those who are already **the most active learners** (section 7.3.5 on page 243).
- (6) The **nature of tasks** affects learning. Innovation projects, difficult tasks and tasks with open challenges provide more opportunities for learning than other tasks do (see sections 7.3.7, 7.3.12, 7.3.8). Furthermore, jobs that allow a person to work on a task from start to end are the type that provide a more complete perspective on the processes and thus support learning (section 7.3.10 on page 255).
- (7) A person's **openness** to new experiences and perspectives supports perspective refinement by comparison with others' perspectives (section 7.3.11 on page 257).
- (8) **Fault Culture**: Excessive fear of mistakes hinders learning (section 7.3.13 on page 258).

It is also worth noting the following important “**Non-Results**” – i.e., factors that were expected to affect learning, based on other studies or popular belief, but that did not add predictive power to the statistical model of this study (details in section 7.3.14 on page 260):

- Communication (i.e., discussions) greatly support learning, but the learning effect does not increase with a higher total **number of contacts**. This supports the insight by Hansen (1999)⁹, who concluded that the depth of personal relationships supports knowledge transfer of complex knowledge.
- **Age** was not a factor inhibiting learning, thus confirming the idea of lifelong learning.
- **Personality traits**, as measured by psychometric constructs, did not affect learning intensity – with the exception of openness to new experiences, from the Big-Five psychometric survey tool (section 7.3.11 on page 257).
- The theory in section 2.3.5 on page 45 suggests that diversity of perspectives supports learning. Thus also **external professional experiences** should facilitate learning. Surprisingly, however, in this study's survey data, external experience is not a sufficiently strong factor to appear within the above lists – which requires further investigation.

⁹The empirical results of Hansen (1999) are discussed in section 2.3.5 on page 44.

8.2. Practical Implications

The general insights from literature and the results lead to the following implications, which provide levers for supporting on-the-job learning. Since these levers are intended for application by organizations, they are grouped by categories that describe different aspects of an organization (following the EFQM¹⁰ leading indicators):

- People (Human Resources)
- Processes
- Leadership
- Partnerships & Resources (including Technologies)
- Policy & Strategy (Overall Organizational Aims)

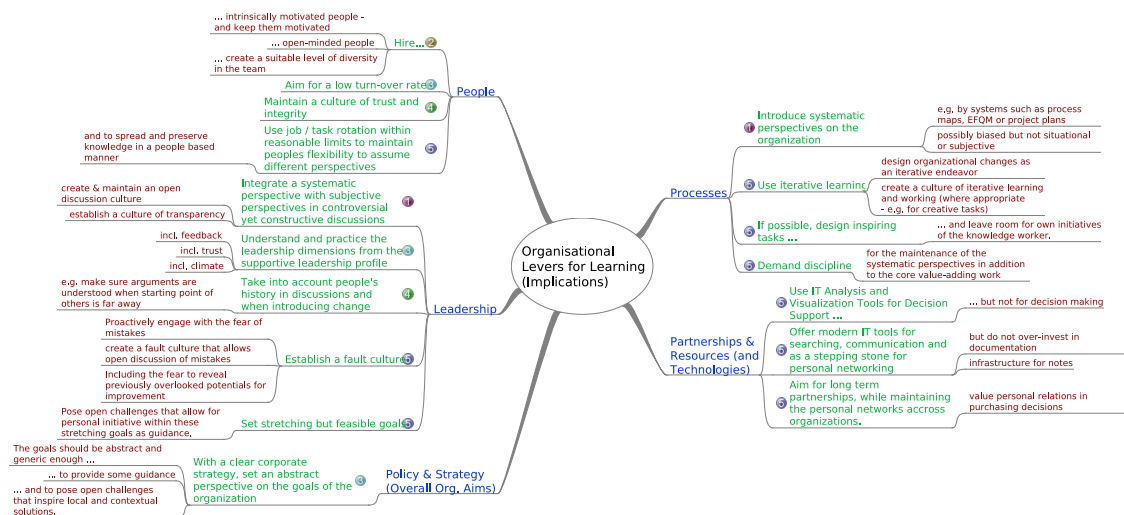


Figure 8.2.: Implications Summarized in a Mindmap

Figure 8.2 gives an overview of the implications. The numbers group the implications by their importance, based on the ranking from the statistical results (section 8.1.2 on page 267). The ranking is particularly useful for practical applications, since most organizations have to focus their limited time and resources on a few measures to support on-the-job learning. Hence the ranking facilitates the prioritization of organizational change according to the Pareto principle¹¹.

The following list discusses the points of the mindmap in further detail:

¹⁰see section 2.4.4 on page 62

¹¹80% of the success can be achieved with 20% of the conceivable effort.

- **People**

- (2) Hire **intrinsically motivated** people with a passion for the product or the work and support this intrinsic motivation. Only people who are personally interested in their work will seize learning opportunities and go through enough learning iterations.
- (New) Cultivate a **pull-approach to learning** in order to leverage employee learning motivation to solve a particular problem at work. Practical measures to support pull-learning could be communicating and discussing the pull-approach with the relevant employees, offering relevant seminars on an as-needed basis, offering relevant documentation for immediate self-education (e.g., with suitable online platforms) and offering coaching.
- (2) Hire **open-minded** people, who are open to engage with the perspectives of others and use these comparisons as an opportunity to refine their own perspective.
- (4) Maintain a **culture of trust** and integrity.
- (3) Aim for a **low turnover rate** for the workforce in order to support and maintain a common language and personal relations.
- (5) Use **job or task rotation**¹² within reasonable limits to maintain people's flexibility to assume different perspectives.
- (5) Since exposure to multiple perspectives on a problem supports learning, hiring a few experienced outsiders injects new and potentially inspiring views in order to create a sufficiently **diverse organization**.

- **Processes**

There are a number of working processes that in the short run are less efficient but because of a learning side-effect are very effective for the organization's performance in the long run. For example, in many business settings it is easier – i.e., more efficient in the short run – to find a quick fix or work-around for a concrete problem (symptom) in a process instead of performing a structured analysis with the aim of finding the root cause and thus learning something useful for all similar problems in the future¹³. This challenge of balancing short-term effort and efficiencies with long-term learning effect and long-term effectiveness becomes even more difficult for

¹²Rotate tasks in such a way that multiple people are 'experts' on a particular topic. This forces new people with a fresh perspective to reconsider the current solution and distributes the knowledge within the organization without the need for expensive databases, documentation and seminars.

¹³Management approaches that aim to support learning therefore frequently mandate problem-solving procedures that involve a structured analysis until the root cause is found. Examples are Toyota's practical problem-solving process (Liker, 2004, p. 256) and Ford's 8 disciplines (Global 8D) approach (Al-Mashari et al., 2005).

processes where problems are less obvious than in many manufacturing settings – e.g., for recruiting processes¹⁴. Therefore a learning effect should be designed into the process.

- (1) Introduce one or more **systematic perspectives** on the organization. Examples are a process map complete with key performance indicators (section 2.4.3 on page 60), an EFQM assessment of the organization (section 2.4.4 on page 62) or a project plan that is used as a visualization of the project status (section 2.4.2 on page 58).

The aim is to create a new level of **transparency** on the processes, which by their nature will have a systematic bias but are without bias due to special circumstances of the current situation or due to the current hopes and wishes of individuals (section 2.4.5 on page 66). The EFQM RADAR method¹⁵ is an example of an assessment method that combines a systematic and **neutral** assessment with an **intelligent but subjective** human judgment (section 2.3.3 on page 35). Similar approaches can be used for status-reporting systems.

Aside from its neutrality, the importance of the systematic perspective lies in the basis it provides to **visualize the processes** in order to create a deeper shared understanding and facilitate **deeper discussions** (sections 2.3.5 on page 44).

- (5) Use iterative processes for **iterative learning** in all challenging and creatively demanding activities. For example, use iterative product design processes, use project review sessions, set up continuous improvement processes (which are iterative by definition), or iteratively design and deploy IT projects¹⁶.
- (5) If possible, **design inspiring tasks** and leave room for the knowledge worker's own **initiatives**. Empowering employees to take responsibility for processes in the organization is a central concept in both the Toyota Production System (TPS) as well as the EFQM excellence model (see [European Foundation](#)

¹⁴Poor recruiting practices lead to poor business performance ([Collins, 2001b](#)), but this link is not obvious in an operational setting, since it acts in a delayed fashion and requires systematic tracking of the results and analysis.

¹⁵See page 64.

¹⁶Corporate IT projects are traditionally executed without a feedback loop in a linear process involving the steps of creation of the IT system's specification, implementation and training. This approach works well for small IT projects, where most aspects of the challenge can be overlooked and it is feasible to create a specification covering all important aspects. In larger IT projects involving many human actors in the design as well as the operation, it is far from trivial to create an all-encompassing specification. Therefore iterative IT project approaches have been advocated, which involve multiple iterations of specification, deployment and feedback [Orlikowski and Hofman \(1997\)](#), which allow the involved actors to iteratively develop and refine their view on the project and make sensible design decisions along the way.

for Quality Management (2003); Liker (2004) and section 2.4 on page 57).

- (5) Demand **discipline** for the maintenance of the systematic perspectives in addition to the core value-adding work (Senge, 2004). An example is the discipline to give regular and accurate status reports.

- **Leadership**

- (1) **Integrate** a **systematic** perspective **with subjective perspectives** in controversial yet constructive discussions into a robust and shared perspective on the situation or problem. Create and support a culture of controversial discourse across all levels of the organization.
- (3) Understand and practice the leadership dimensions from the supportive leadership profile, including **personal feedback** and **group climate**¹⁷ in order to support collaboration and learning.
- (4) **Take into account people's history** in discussions and when introducing change.
- (5) Establish a **fault culture**: Proactively engage with and reduce the **fear of** working openly on finding solutions and preventive measures for mistakes.
- (5) Set **stretching**¹⁸ **but feasible goals**. Pose **open challenges** that allow for personal initiative within these stretching goals as guidance.

- **Resources**

- (5) Use **IT analysis and visualization tools** to help the participating actors refine their perspectives on complex problems in order to support them in sound human judgment and intelligent decision making (section 2.3.3 on page 35).
- (5) Since the learner's active engagement is required for learning, **IT systems for storing, searching and sharing** information only become useful when motivated employees start to use them on their own initiative. Hence IT systems can be a focus second to the more human aspects of learning (e.g., motivation or a culture of open and controversial discourse).

Furthermore IT systems should not only provide access to information by storage and searching facilities but also connect people, i.e., suggest knowledgeable contacts on a topic, and allow for multiple opinions¹⁹.

¹⁷Note that the aspects of leadership discussed here are those that support learning. The aim is not to describe 'good' leadership in general.

¹⁸'Stretching' as in 'challenging' goals.

¹⁹For example, many wiki systems, such as the web encyclopedia Wikipedia (<http://www.wikipedia.org/>), feature a discussion feature that allows the discussion of the contents of a wiki page.

- (5) Storing information may be useful for the future, but the danger of **over-investment in documentation** or the danger of not being able to keep the documentation up to date should be carefully hedged. Sparse documentation with information suggesting knowledgeable contacts may be an effective alternative.
- (5) Maintain **long-term partnerships** with external partners – especially customers and **suppliers**. Personal relationships and a common language also support learning across organizational boundaries.

- **Policy & Strategy (Overall Organizational Aims)**

- With a clear corporate strategy, set an abstract perspective on the **goals** of the organization. The goals should be abstract and generic enough to provide some guidance but also pose open challenges that inspire local and contextual solutions.

So far the organizational levers have been presented. Organizations will, however, have to design contextually adapted organizational changes to support learning – for which a few design criteria should be considered:

The organizational change to support learning should be ...

- ... **self-sustainable in the long run** Once organizational changes to support learning have passed an introductory period, the utility of the change should be substantial and widely recognized to make it independent in the long run from promotion by individuals.
- ... **aligned with the organizational goals and priorities** in order to target the resulting needs in decision making and knowledge-intensive problem solving (avoiding a decoupling of knowledge from its uses – see (Fahey and Prusak, 1998))
- ... have a true **focus on knowledge flows** (learning), not just information with a new label. The human side of knowledge management needs to be designed into the organizational change.
- ... **show early wins** to gain acceptance during the introductory phase (a common change management technique)

In summary, there are many ways to support learning, but one of the most important is to create and maintain a shared perspective of all involved actors – based on systematic as well as individual perspectives. The shared perspective serves as a common language for deep discussions, leading to widely accepted and better decisions. Rather than solely focusing on knowledge databases, IT systems can play an important role as support tool

for generating a visualization of the shared perspective. Yet all these measures can only be fruitful if the participating organizational actors are motivated to learn.

Not surprisingly, a number of effective industrial practice models already rely on many of the described effects – most notably EFQM and the Toyota Production System (section 2.4.1 on page 58). Yet a refined understanding of the drivers for the effectiveness of these practices allows organizations to improve and better adapt these practices to new contexts (i.e., industry- or organization-specific challenges).

8.3. Relevance of the Results to Literature

The results of this study are furthermore relevant to a number of discussion streams in literature – as will be presented in the following.

8.3.1. Relevance to Organizational Learning

A major challenge in the field of organizational learning is to measure the learning effect. In some cases the productive effect of learning can be used to indirectly measure individual learning – such as in the study by [Argote \(1999\)](#), who uses the increase of shipbuilding tonnage as an indirect gauge for the learning effect. Along the same lines, in their study on a consulting firm, [Haas and Hansen \(2005\)](#) compare the fraction of successful consulting pitches with and without the incentivized use of a knowledge database. Similarly, learning can be indirectly measured in all settings in which productivity trends over time are driven only (or mostly) by learning. Examples are error rates in classical production or yield learning in the semi-conductor industry.

Using performance measures as an indirect measure for learning has the advantage that it is an indirect yet external and not self-reported measurement. Not relying on self-reports eliminates self-reporting biases (see section 5.4.1 on page 146). Yet both of the abovementioned example studies measured performance, not learning. While performance is in many cases preceded by learning, learning alone is a necessary but insufficient condition for improving performance. In the study by [Haas and Hansen \(2005\)](#), experienced teams were more successful with their pitches when they relied on their personal contacts to gain access to knowledge rather than on the electronic knowledge database. Yet the cause of their superior performance may still have been their superior level of experience, which caused them to rely on personal contacts and succeed with the pitch without intensive use of the database²⁰.

In contrast to these indirect learning measures, the learning index used in this study is a direct measure of experienced learning episodes. It is based on the results from a

²⁰This is yet another example of challenges with causation in statistics – as discussed in section 4.1.5 on page 109.

novel interactive survey mechanism that was designed following state-of-the-art surveying principles – e.g., asking questions about concrete learning examples rather than asking for general statements that might be strongly biased by prior beliefs. Nevertheless, the learning index is still a self-reported measure of learning experience. Thus there is no need to infer the learning effect indirectly, and difficult questions regarding causation do not arise, but researchers instead face challenges stemming from self-reporting biases (see [section 5.4.1 on page 146](#)).

Both types of learning measures have their own *different* strengths and weaknesses. Future studies would benefit from using both types of measures simultaneously in a multiple-method approach. Using both measures would provide insights towards the magnitude of inaccuracies caused by the respective biases²¹.

8.3.2. Relevance to Industrial Practice Models

While there is already a lot of literature on the application and effectiveness of industrial practice models such as the EFQM, the Toyota Production System and the Balanced Score Card, there is hardly any literature aiming to explain why and how these models work using scientific evidence ([Kujala and Lillrank, 2004](#), p. 43).

As discussed in [section 2.4.1 on page 58](#), the three industrial practice models share the following features:

- Create a shared understanding first (e.g., creating transparency by visualization).
- Repeat and iteratively improve.
- Engage in deep discussions regarding to create a contextually adapted solution – rather than following “best practices” blindly.

The PIA-model, which is based on scientific evidence from scholarly literature and the survey data from this study, is one way to explain how these common features of the industrial practice models are effective (see also [section 2.4 on page 57](#)) in many circumstances.

Since organizations, including businesses, operate in many different environments and face many different challenges, successfully applying these industrial practice models in new contexts will require some adaptation of these “best practice” models to the particular organization (instead of blind imitation).

A deep understanding of how a particular industrial practice model works will thus increase the chances for successful adaptation to a new context. The insights from the PIA-model would for example direct the practitioner’s focus on the visualizations used

²¹In principle, a similar multiple-method approach was used for the variable-importance measure in [section 7.1.1 on page 206](#), where two different variants of variable importances with different weaknesses are used.

and draw the attention to the careful and suitable designs of these visualizations in a way that highlights the key points of the particular business context.

In summary, a scientifically based model, such as the PIA-model, can be a helpful guide to gaining a deeper understanding of a particular industrial practice model and can support the effective adaptation and application of the industrial practice model.

8.3.3. Relevance to Sense Making, Problem Solving and Knowing

In the literature, the field of knowledge-intensive work has been approached by a wide variety of methods. Insights have come from qualitative as well as quantitative (usually laboratory experiment-based) research efforts – e.g. Orr (1996) or Siegler (2005).

Using a fully structured survey to obtain empirical data for this study has a few challenges (see section 8.4 on page 280), but two important advantages can complement the existing body of knowledge: the survey can be economically applied to a large sample of participants in real organizational settings, and it allows a **quantitative ranking of many factors driving learning**.

Many of the ranking results are not very surprising and confirm existing literature. For example, intrinsic motivation to learn has received a high ranking from the survey data, which is confirmed by a number of scholars (see section 7.3.3 on page 238).

Nevertheless, some ranking results are surprising and therefore suggest further investigation. For example, judging from the survey data, exposure to different perspectives on a problem is an important factor driving learning in knowledge-intensive work (see sections 7.2.2 on page 219 and 7.3.1 on page 226). In the literature, some but not many studies (e.g., (Orr, 1996; Prusak, 2005; Schreyögg and Geiger, 2007; von Krogh and Grand, 2000)) mention this factor and highlight its importance. Thus the results from this study point towards a closer investigation of this frequently overlooked factor.

8.3.4. Relevance to Statistics

The large number of variables potentially driving learning and the properties of the data (e.g., much noise – see section 5.12 on page 163) created a number of challenges for the statistical analysis that could not be dealt with using conventional methods (see section 6.1 on page 172). Therefore a substantial effort had to be put into creating the **BOGER algorithm** (section 6.2 on page 179) combining state-of-the-art statistical methods in a novel way in order to meet the special requirements stemming from the properties of the dataset and the desired analysis output.

The resulting BOGER algorithm features:

- A **robust** mechanism to systematically build a statistical model (based on **full model selection**) in a **data-efficient manner for metric scales** (see section 6.3.3

on page 202). The **predictive power** of the model was superior (section 6.3.2 on page 199) to any other existing algorithm that was tested on this study's dataset (section 6.1.3 on page 175).

- Breiman's variable importance measure was used and improved (section 7.1.1 on page 206) to **provide a ranking of factors** (section 6 on page 171) that makes it possible to draw inferences from the robust but opaque BOGER model.
- Finally, a **special type of frequency plots** has been developed to compare the raw data with the model shape not by a single number but graphically (section 7.1.3 on page 213). Numerical (single scalar) indicators for effect strength and model fit are an extreme form of complexity reduction. They are commonly calculated to answer the following two underlying questions: 1.) What is the nature of a particular relationship? and 2.) How good is the model that this insight about the relationship is based on? Yet the conventional R^2 model-fit estimator, for example, occasionally gives biased results and thus provides a non-robust answer to question 2 (see figure 4.2 on page 127).

As illustrated by figure 7.2 on page 215, a suitable graphical representation can provide a better overview and more robust insights on model fit as well as effect strength. The graphical approach is therefore proposed as a standard tool for comparing (in this case metric) statistical data with a statistical model. Using suitable graphics in the early analysis steps is useful as a complementary approach to using numerical indicators, in a second step, to quantify results that have been found to be robust.

All these statistical tools are generic enough to be useful in other studies with a similar data structure (many variables, many conceivable models, metric scales, strong collinearities and a high level of noise).

8.3.5. Relevance to Knowledge Management

In section 2.5 on page 68 a number of concepts frequently used in the literature on *knowledge management* are contrasted against partly opposing concepts derived from other theories and the results of this study, which have been condensed in the PIA-model. Most noteworthy are:

Knowledge is framed in the PIA-model as a *personal skill*, which includes and highlights the notion that knowledge depends on an individual, partly shared and historically evolved perspective on facts. The term "knowledge", which triggers an immediate association of **knowledge as an object** that can be stored, transferred and managed (directly), is therefore a metaphor that is at least a strongly simplifying description of the processes on the individual level (see section 2.5.5 on page 76).

The effect is not surprising: higher-quality articles frequently address the human side of knowledge, while simpler treatments of this subject do not cover this important facet of the topic. The popularity of the knowledge metaphor in the knowledge management literature can in part be explained by the focus of much of this literature on knowledge processes between groups of employees or different parts of large organizations (see, for example, (Argote et al., 2003; Augier and Knudsen, 2004)).

Using the *knowledge as a personal skill* perspective has profound implications that are not considered in large parts of the knowledge management literature, such as:

- Since knowledge is the integration of information with a person's background knowledge, knowledge can hardly be identified as a set of small knowledge objects that are meaningful without the person's background knowledge. Thus **mapping of knowledge** can only be the mapping of information, not person-bound knowledge. Mapping of information can be useful but its value only comes to bear when motivated learners use the information to learn (see section 2.5.2 on page 73).
- Rather than a **single version of the truth**, true mastery of a subject derives from critical examination with multiple expert perspectives (see section 2.5.3 on page 74).
- **Valuation of knowledge** is very difficult, given that one can hardly identify individual knowledge objects (as mentioned above) and given that the usability of knowledge is difficult to predict (see section 2.5.4 on page 75). Therefore, intellectual capital balance reporting approaches (e.g., (Bornemann and Alwert, 2007)), with the intention of assessing the value of knowledge within firms, can only be very rough approximations of the value of the knowledge within the heads of the firm's employees.

The results of this study highlight some important aspects of knowledge management that have in many studies received little attention but are generally accepted principles in other literature streams.

8.3.6. Selection of a Few from Many Plausible Explanations

Many insights related to on-the-job learning can be found in the literature – especially when not only studies under the heading 'learning' are considered (see section 2.2 on page 23). Most of these insights are plausible and are supported more or less well by evidence. These studies frequently focus on a few factors and test their effect on learning.

This study's data can support a subset of these insights, and thus a number of the results are in themselves not completely new. Yet a number of insights popular in the literature did not find very strong support in the survey data (see section 7.3.14 on page 260).

Therefore part of the value of this study's results is in selecting insights from a larger set of plausible insights and underpinning them with empirical evidence.

The statistical ranking results guided an additional second wave of literature research with new and refined search terms (section 3.2 on page 97). Thus the ranking results were a useful tool for **iterative perspective refinement** on the challenge of supporting on-the-job learning (section 3.1.6 on page 93).

Thus **this study's results contribute to the literature by the collection and selection of partly existing insights scattered across many research fields as well as by the complexity-reducing integration of these findings in the PIA-model** (figure 2.1 on page 31) – i.e., the creation of a suitable inter-disciplinary perspective specialized for the challenge of on-the-job learning – all **supported by empirical evidence**.

8.4. Limitations

Research is rarely without compromises. In particular, requirements and constraints in the following dimensions are frequently contradictory in their nature and thus need to be traded off against each other: suitable perspective, validity and reliability constructs, suitable choice of participants and case studies, limited resources (including time) and limited overall duration of the study (section 3.1.8 on page 96).

The research design really aims at achieving the best compromise among all the design requirements and limitations mentioned above. Since research designs are principally concerned with minimizing the total risk of obtaining false results, the risks should be reduced in a balanced manner – i.e. the biggest risk should be the first target for risk mitigation²². In addition couplings among the different risks should be considered: e.g. estimation errors of effect strengths may be overshadowed by a weak statistical model (with low predictive power).

In particular, this study's research design is subject to the following risks:

- **Construct validity** – The participants (or sub-groups of them) may have misinterpreted some of the questions. In addition, some aspects of on-the-job learning are difficult to survey directly with a fully structured survey – an example is the professional background and experience of the participant. In addition, despite the pilot phase, some questions might have been misunderstood. Thus there is a risk that some components of the survey measure something other than what they are intended to measure.

²²The aim of research design should be the reduction of the total risk of false results. Thus, following general risk management techniques, the probability of a false results by a particular flaw or bias should be multiplied by its negative impact on the accuracy of the results. Then for all possible flaws or biases the result of the multiplication should be summed to obtain the total risk.

- **Surveying Detail Level of the Important Factors** – Some factors, such as the question item on “personal interest” for an intrinsic motivation (see section 7.3.3 on page 238) have come into an unexpectedly strong focus following the results of this study. In hindsight, a more detailed coverage of motivation with multiple facets by using multiple question items – that are suitable for this type of learning situation – would have been desirable. Hence with the results of this study, future studies on this topic would be able to refine the surveying technique, concentrating on the most important factors with more detailed question items.
- **Omission of Important Variables** – Important variables describing relevant aspects of the work environment (including the organization), which could have led to an improved predictive power of the model, could have been omitted²³.
- **Biases** – The variables may be biased due to various reasons, e.g., social desirability or self-selection effects. This risk also applies to the learning index as outcome variable but was mitigated as far as possible – see sections 5.4 on page 145 and 5.11 on page 162.
- **Robustness** – Despite the efforts to make the BOGER algorithm perform robustly on the data (section 6.2 on page 179) and the subsequent investigation of its actual performance (section 6.3.2 on page 199), there remain some smaller risks of getting spurious results.
- **Small but Pivotal Effects** – Some factors may have a small yet causally pivotal effect – which makes them hard to detect given the high level of noise in the data. An example is the importance of getting additional ‘hard’ information by experimentation for validation of a refined perspective on a problem – see section 7.3.1 on page 230.
- **Misinterpretation of Causal Mechanisms** – The statistical results only provide insights on associations, not causal links (section 4.1.5 on page 109). Therefore, in section 7.3 on page 226, the statistical results were fused with insights from literature to draw inferences on the causal mechanisms. Yet integrating literature findings with statistical results requires sound judgment, which is never completely free of subjective bias (section 3.1.3 on page 86). Hence there are few situations in which a claim for a plausible explanation with support from the literature can be made without the smallest doubt.

²³As discussed in section 5.12.2 on page 166, certain situation-specific aspects of the learning episode were not and could not be surveyed and thus become latent variables. However, there might also be variables that can be surveyed and have been overlooked and omitted.

- **Generalization Beyond the Surveyed Organization** The ranking results from section 7.2 on page 217 are based on a survey from a single organization: Meyer Werft in Papenburg, Germany. Thus the question arises whether and to what extent these results may be used to generalize about other organizations.

Nevertheless, the survey covered all departments and thus a wide variety of different tasks and working environments²⁴. Therefore, given the solid support for the results in the literature (section 7.3 on page 226), applying the same data collection and analysis method to other organizations would most likely lead to similar ranking results. Yet a few properties of the organization Meyer Werft apply in the same way to all employees. In another organization with other properties, other effects may be observed. At the same time, such effects might not be detectable with the dataset from a single organization. Conversely, some properties of the organization, such as the level of competitive pressure from the global shipbuilding market, may not be shared by other organizations, and thus effects due to such properties would not generalize to other, principally different organizations.

Given these risks, which overall could not be further mitigated by the research design²⁵, the best test for these results is inspection and validation by other researchers with new data in other studies – i.e., by an external research iteration 3.1.6 on page 93.

8.5. Areas for Future Research

The analysis and the interpretation of the results raised a number of questions for future research in two main directions:

- the mechanisms driving on-the-job learning
- the methods – including the statistical tools and the survey instrument

Future Research on the Mechanisms of On-The-Job Learning Section 7.3 on page 226, on the detailed interpretation of the results, discussed various questions that could not be clarified fully with the information about the association of the factors with learning and literature. The following list summarizes the most important open questions (details can be found in section 7.3 on page 226):

- **Learning strategies** that involve the views of others have been found to be most supportive of learning (section 7.3.1 on page 226). Yet there is more to investigate

²⁴Compare, e.g., the work of a welder with that of a technician from the IT department.

²⁵The risks could not be further mitigated without reaching a different design point, e.g., by increasing the research cost/effort substantially.

about how these strategies are effective. In particular, an investigation of the weaker strategies, such as experimentation, would be worth the effort.

- Personal interest has been identified as a strong **intrinsic motivator** for learning (section 7.3.3 on page 238). Yet questions arise about whether and how this personal interest depends on the nature of the task, and whether and how learning and problem-solving success conversely affects the personal interest. Personality traits also appear to interact with motivation (Judge and Ilies, 2002). In addition, it should be investigated whether **extrinsic motivators** such as monetary reward systems have a similar effect as intrinsic motivators, since some organizations use bonus systems to promote knowledge sharing and searching (Haas and Hansen, 2005).

Furthermore, as mentioned in the limitations section 8.4 on page 280, motivation is covered in a very simple manner – yet from theory, and as supported by the statistical results, it is an important factor that deserves more detailed coverage with more facets. Thus future research should focus on dominant theories on work motivation (Steers et al., 2004), such as the “*expectancy theory*” by (Vroom, 1964) – with suitably detailed surveying tools in order to be able to clearly separate the motivational elements mentioned in theory²⁶.

- A particular **leadership** profile has been found to have a strong effect on learning (section 7.3.2 on page 231). Yet a number of questions about the detailed mechanisms remain open, e.g.: How does leadership shape a knowledge-conducive task design? How do the properties of these knowledge-conducive tasks affect learning?
- As predicted in the literature, a person’s **individual professional history** strongly shapes learning (section 7.3.4 on page 240). However, more knowledge about which kinds of professional experiences support learning would provide useful guidance for human resource development. Further investigations regarding the detailed mechanisms that cause professional experience to have an effect on learning, as well as the development of improved survey tools to assess the professional biography, would be valuable.
- As discussed in section 7.3.5 on page 243, the effect of **learning barriers** in the context of the learning feedback loops is non-trivial. It would be insightful to investigate the barriers in detail, as well as their effect, e.g., with ethnographic research

²⁶Valence, Instrumentality and Expectancy are similar constructs that have somewhat redundant predictive power. Yet the product of all three factors (i.e., the AND connection of the factors – as predicted by Vroom’s theory) in the meta-analytic study by Van Eerde and Thierry (1996, p. 581) does not yield more predictive power than the individual components, which may be due to some challenges of the meta-analytic design of the study.

approaches²⁷.

- A significant association of **epistemological beliefs** with learning has been found in the survey data (7.3.6 on page 247). However, the causal direction remains unclear. Therefore future research should include an assessment of epistemological beliefs – preferably with alternative measures for learning in order to understand the effect of epistemological beliefs on possible biases in the learning index. If there is such an effect, this would be very relevant to organizations, since the form of secondary education, which employees have received during their upbringing, affects epistemological beliefs (Schommer et al., 1997).
- An effect of **education level** on the learning index, a relative and self-reported learning measure, was not found (section 7.3.14 on page 260). However, is the absolute (not self-reported but independently measured) learning effect truly independent of the type and level of education?
- The variables regarding the **personal network of contacts** have not shown an association with the learning index (section 7.3.14 on page 260). However, in the literature a number of authors cite a causal but non-trivial connection. The effect of the personal network should therefore be reassessed with different methods, in particular with different surveying tools.

Aside from laboratory studies and surveys, ethnographic research (Orr, 1996), natural experiments (Starbuck, 2004) or the actor’s approach (Arbner and Bjerke (1997) in section 3.1.3 on page 89), in addition to reflective practical applications in actual organizations, lend themselves to the investigation of these questions. To monitor the overall effect on the organization, a framework such as the EFQM model could be used (section 2.4.4 on page 62) to yield standardized assessments that would allow for comparisons across organizations²⁸.

Future Research on the Methods Even though substantial effort has already gone into validating the learning index as a survey tool for the learning effect (sections 5.11 on page 162 and A.4 on page 290), further validation, e.g. against other measures for learning, would certainly be helpful to gain a deeper understanding of the accuracy and biases in this surveying tool. For the survey in this study, a self-reported learning assessment was the only type of measure that was feasible within the time restrictions given the large number of other factors assessed (chapter 5 on page 137). Therefore, a comparison of the

²⁷Similar to the ethnographic research approach of Orr (1996), who followed Xerox copy technicians in their normal work for months.

²⁸Following the spirit of the RADAR approach in the EFQM model, such a comparison would not be a simple comparison of criteria scores but a good judgment with the help of various results from the RADAR assessment (EFQM, 2001).

learning index with an absolute (and independent) measure of the learning effect would be insightful. Possibly a research approach similar to the one introduced by [Siegler \(2005\)](#), which mixes qualitative with quantitative methods, should be considered.

The BOGER algorithm, with its screening stage and the final interactive model building stage, is a significant improvement compared to other algorithms (e.g., ordinary multivariate regression or step-wise regression) for similar applications with similar datasets, since it robustly and efficiently²⁹ allows the researcher to perform a full-model search (section [6.3.3 on page 202](#)). Yet the algorithm would benefit from the development of more efficient search strategies for the final full-model search stage (section [6.2.6 on page 192](#)).

²⁹BOGER is efficient regarding the required sample size, computational effort and required researcher interaction.

A. Appendix

Chapter Contents

A.1. Writing Style and Conventions	287
A.2. Searching in Literature	288
A.3. Company Profile – Meyer Werft	289
A.4. Validity Investigation of the Learning Index in Detail	290
A.4.1. Inspection of the Input Data to the Learning Index	291
A.4.2. Distribution of the Learning Index	295
A.4.3. Cross-Validation of the Learning Index with Related Questions	297
A.5. Data Pre-Processing Details	302
A.5.1. Filtering and Outlier Removal	302
A.5.2. Imputation and Missing Value Filtering	303
A.6. Details on BOGER	305
A.6.1. Generating Data Frequency Equalized Bootstrapping Samples	305
A.6.2. Implementation Details of BOGER in \mathbb{R}	307
A.6.3. Flexible Model Fitting - an Interesting Accident	310
A.6.4. Residuals	311
A.6.5. Empirical Robustness of Model Fit Measures	314
References	316
Index	339

A.1. Writing Style and Conventions

The writing style I used in this thesis follows the following recommendations of the publication manual of the American Psychological Association ([Association, 2002](#)):

A.2. Searching in Literature

- **reducing wordiness** – e.g. “several students completed...” instead of “there were several students who completed...”
- **avoid passive tense** – use ‘I’ or ‘we’ instead of ‘the researchers’ and passive tense for clarity
- use **pronouns** only in as clear and **unambiguous** reference
- **consistent** use of **tenses**
- **avoid colloquial** expressions
- **word choice** – adhere closely to official and strict meaning (unless redefined)
- mix **short** with **long sentences**
- **clear comparisons**

The APA recommendations were chosen for their emphasis on clarity and economy of expression.

A.2. Searching in Literature

Electronic Search Early knowledge management literature, aside from explicating knowledge and storing it (Wexler, 2001), frequently focused mostly on electronic searching – for documents or other bits of information in knowledge databases (DeMocker, 1998; Kaufman, 2002).

Yet many authors quickly recognized that electronic search is much less popular than a search via networks of personal contacts (social networks) (Jacobson and Prusak, 2006; Salter and Gann, 2003; Sandow and Allen, 2005; Voelpel et al., 2005).

Haas and Hansen (2005) even found that searching electronically rather than via social networks can actually hurt performance. In his case study on a consulting firm, the consultants were formally incentivized to use electronic search on a new knowledge database, rather than relying on their social networks. Comparing the average business success of the teams with these incentives and some teams without incentives (and thus a more social search behavior), especially highly experienced teams showed a reduced performance when using electronic search rather than relying on their rich social networks.

A part of this effect might be explained by the fact that electronic search itself is a skill that is trainable (Debowski et al., 2001; Weiss et al., 2004) – yet formal trainings on using electronic search tools are rare.

Searching – a Filtering Challenge Another challenge with electronic searching is to filter the results. Good search terms certainly go a long way toward reducing and refining the results, but in the end the user, not the machine, needs to perform an intelligent selection (i.e., filtering) task, which non-intelligent computers cannot perform.

Consequently it is in this filtering task that some authors see a challenge of information overload:

“[...] while knowledge search and management efforts to date have been valuable, future payoffs will depend less on enhancing systems that track down information than on devising strategies to help employees use what they’ve found.”, (Jacobson and Prusak, 2006, p. 34)

However, relying on personal contacts is different: other humans perform the intelligent filtering task, based on their associative knowledge of information sources, for those who ask. Salter and Gann (2003) observed this effect with technical design engineers, who dealt more effectively with information overload by relying on personal connections rather than ICT search engines.

The idea that humans are very effective at filtering will not come as a surprise to the reader after the arguments presented in section 2.3.2 on page 30.

Social Networks for Search Consequently, social network theory became relevant for analyzing knowledge flows within organizations (Cross et al., 2001) and researchers began to analyze the structure of these networks and the nature of the links:

Hansen (1999)’s study of an engineering firm found that very large yet shallow social networks were effective in obtaining simple information efficiently, but small networks of personal relationships, built over a long time, were more effective in transferring complex knowledge – which is in line with the theory of mental models and shared perspective from section 2.3.6 on page 47. The perspectives of people with a lot of social interaction are likely to develop in a similar (socially constructed) direction (Tsoukas, 2005b).

In summary, the literature on searching indicates that social searching is frequently more effective than electronic searching, which is not surprising given that asking other humans is like leveraging a very intelligent search engine, which applies an effective filter and strongly reduces the challenge of information overload for those who search. Furthermore, all of these insights regarding searching fit well into the PIA-model in figure 2.1 on page 31.

A.3. Company Profile – Meyer Werft

The Meyer Werft shipyard is located in the town of Papenburg in northwest Germany. Together with its sister shipyard Neptun Werft in Warnemünde (near Rostock in northeast

Germany), Meyer Werft specializes in high-value ships, with a focus on cruise ships, river cruise liners, gas carriers and ferries.

With over 213 years' history with smaller vessels, Meyer Werft entered the cruise ship market in 1985 with the cruise ship *Homeric* for Home Lines. The MV *Homeric* was about 200m long and had a size of 42,000 GT (gross tons¹). Since then the sizes have almost tripled to 122,000 GT and 315m with the delivery of the *Celebrity Solstice* in 2008.

In 2009, two larger cruise ships were delivered. A cruise ship is a very large project of around half a billion euros, of which Meyer Werft and Neptun Werft outsource roughly 70% to suppliers and contractors.

In contrast to cargo ships, cruise ships are commonly built in small series of two to six ships, since cruise ship operators compete by continually offering new cruising experiences to their passengers. Therefore the shipyard custom designs each ship for its owner, which involves the classical ship and steel design, system integration of various machinery, accommodation design (following the design of the ship owner's architects) and the purchasing and subcontracting of about 70% of the ship's value. During the design as well as the construction phase, the ship owner closely collaborates with the shipyard on any open detail design decisions and for further optimization of the ships.

The production process resembles a large flow line that begins with the automatized manufacturing of simple steel assemblies (e.g., a plate with stiffeners), continues with the assembly of blocks and sections, which are also outfitted with pipes, cable trays and A/C ducts. The ship is then assembled in a few months in one of two large covered dry docks. Each of the assembled pieces is unique, hence mass-production principles are applied in an adapted manner.

Thus it is an important competitive factor for the yard to be able to develop highly optimized and customized prototypes as well as to build cruise vessels successfully and efficiently in small series and additionally to accept design change requests from the ship owner during the process.

As of 2009 Meyer Werft has about 2,500 employees, including more than 300 engineers and administrative staff. Employee turnover is very low. The survey was targeted at all employees from the production departments as well as the technical design and administration departments.

A.4. Validity Investigation of the Learning Index in Detail

Given the novelty of the learning index, its reliability and validity needs to be verified. Thus in the following the consistency of the learning index – internally and with other question items and constructs – is verified:

¹Gross tons are metric tons of displacement with a rough correction factor for the value and complexity of the vessel.

A.4.1. Inspection of the Input Data to the Learning Index

Learning Situations in Numbers

The learning index is composed by data about learning situations as well as learning importance and usefulness ratings. Therefore the quality of this input data is inspected here before inspecting the learning index directly.

	Mean No. of Learning Situations per Workstep	% of <i>all</i> Learning Situations
Work Step 1	1.25	38 %
Work Step 2	1.16	34 %
Work Step 3	1.16	28 %

Table A.1.: Usage of Worksteps

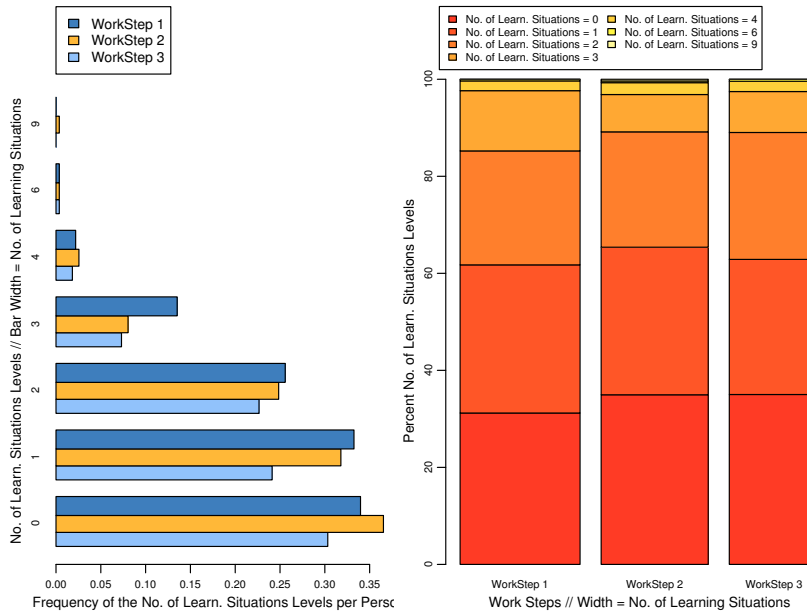


Figure A.1.: Frequencies of Learning Situations Levels

Table A.1 shows a fairly equal distribution of learning situations per work step. The fewer learning situations for work step 3 are explainable, since the participants were asked to split their example project into two or three tasks or to select two or three tasks from their normal work of the last four weeks.

In more detail, this is confirmed by figure A.1. The left sub-figure shows the absolute

A.4. Validity Investigation of the Learning Index in Detail

frequency of learning situation levels grouped by work step, while the right sub-figure shows the relative frequency of learning situation levels for each work step. The fall-off of learning situation frequencies over increasing learning situation levels (learning situation entry slots) is expected, since the participants will always use the first slot first, and only if there was another learning situation would they also fill in the next slot. Not having this fall-off here would have been an alarming sign.

One question during the design of the survey was whether participants actually get bored while going through the many iterations of the learning index survey tool. Given that participants listed a substantial amount of learning situations in work step 3, this concern thus cannot be confirmed.

Learning Situation Importance

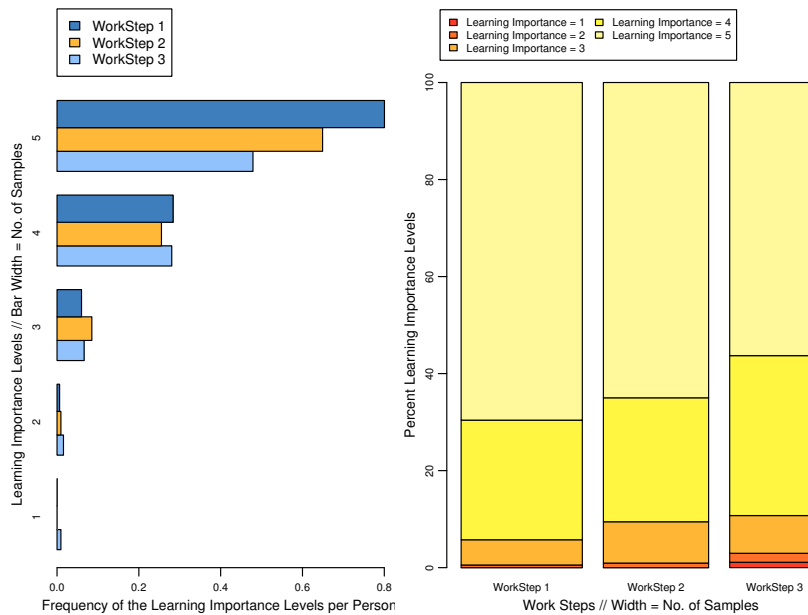


Figure A.2.: Frequencies of Learning Importance Levels by Workstep

Figures A.2 and A.3 on the next page illustrate the answering behavior for learning importance – distributions grouped by work step and by learning situation. Again the absolute frequencies are at left and the relative frequencies of the learning importance level are at right for each work step.

It is striking that at first the dominant overall rating is “*very important*” (the reader may look for *learning importance level 5* in the left sub-figure). This result implies that participants who experienced a learning situation mostly rated it as “*very important*”.

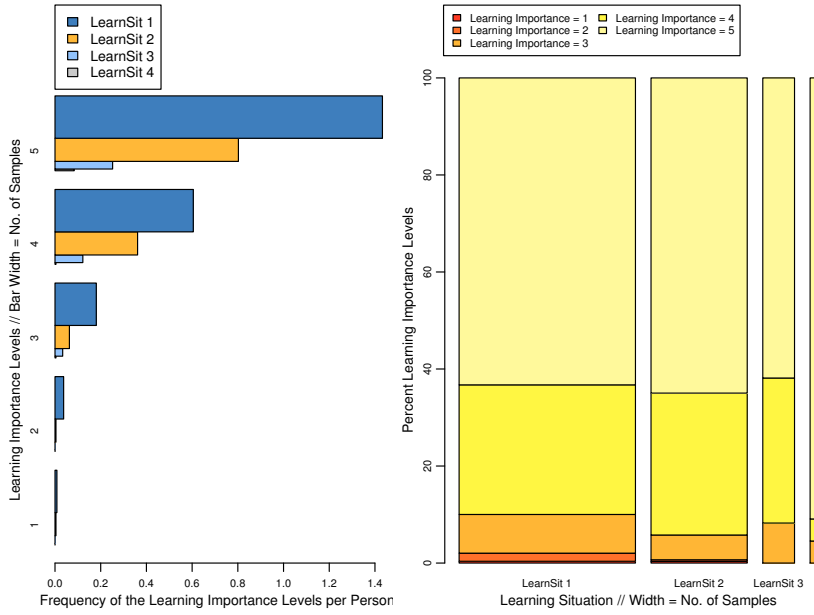


Figure A.3.: Frequencies of Learning Importance Levels by Learning Situation

This is not surprising, given that we remember events we consider important better than marginal events (Anderson, 1988). It is also likely that the participants prefer to mention important learning situations rather than unimportant lessons.

Thus the learning importance scale is used fairly unevenly with a strong bias to the “very important” end. Similar to the biases described in section 5.4.2 on page 150, this bias is systematic and applies independently of the independent variables and therefore does not degrade the usefulness of the learning index.

The right sub-figure of figure A.2 on the preceding page furthermore shows that the fraction of frequency of a particular learning importance level remains fairly constant of the work steps. This implies that there is only a very weak dependence of the work step on the learning importance. The left sub-figure A.2 on the facing page further reveals that this weak dependence exists only for learning situations rated as “very important”. If there had been a strong dependency, this would be a clear sign for an undesirable survey artifact.

The right sub-figure of figure A.3 presents a similar picture. There is only a weak dependence of the learning situation slot² on learning importance (the frequency fractions stay fairly constant).

²Learning situation slot 3 refers here to the actual entry field number 3 in the respective survey webpage.

Learning Situation Importance and Usefulness

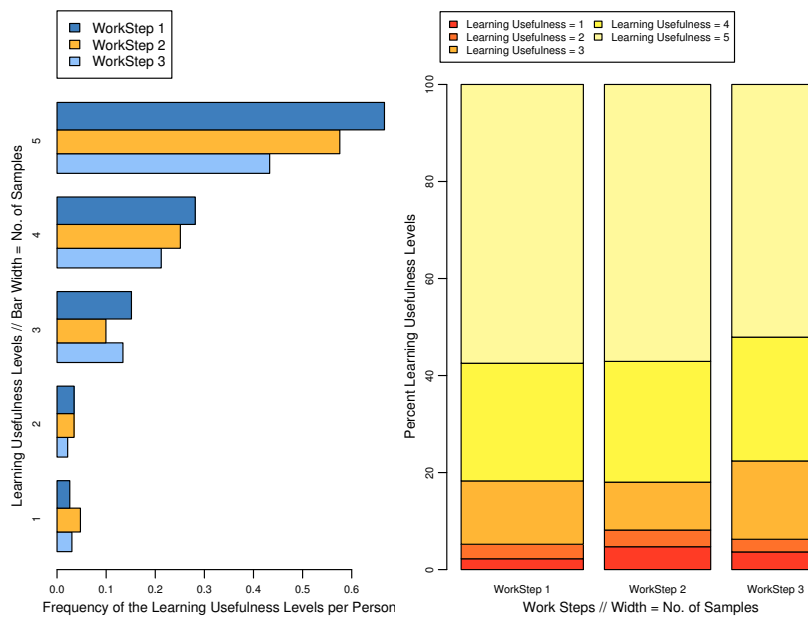


Figure A.4.: Frequencies of Learning Usefulness Levels by Workstep

A similar picture presents itself for the rating of the usefulness of a particular learning situation. Figures A.4 and A.5 on the next page show an even weaker dependence on work step or learning situation entry slot.

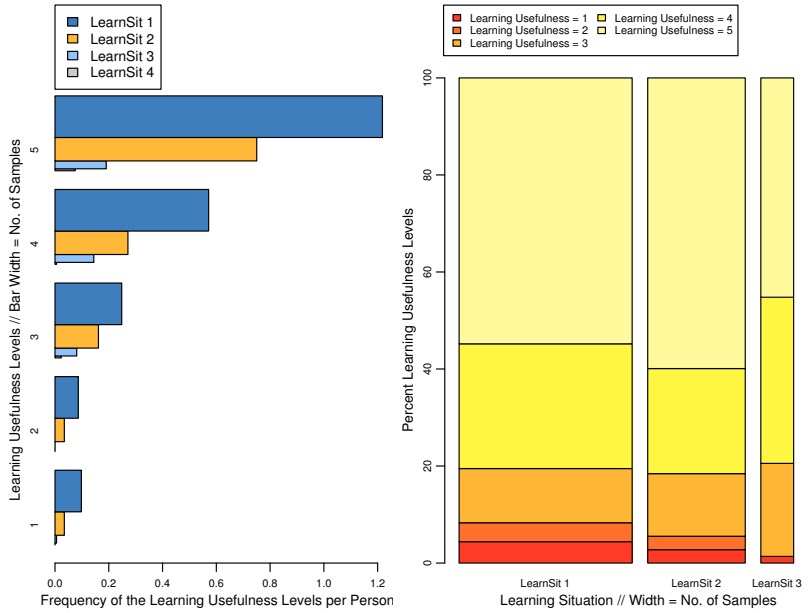


Figure A.5.: Frequencies of Learning Usefulness Levels by Learning Situation

Relationship between Learning Importance and Usefulness

With the reduced definition (equation 5.2 on page 151) of the learning index based only on learning importance and not also on the learning usefulness, as in the full definition (equation 5.1 on page 150), the learning importance is effectively used as an imputation³ value for missing values of learning usefulness (for those participants with a lower survey reduction level).

Thus figure A.6 on the next page is used to investigate whether learning importance and usefulness are related. The kernel density estimation plot with a scatterplot as overlay shows a mildly strong relationship (Pearson correlation = 0.358).

Thus both measures behave as expected from theory. Important learning experiences are important because they are likely to become useful in the future. In addition, the effective imputation strategy with the two definitions of the learning index is acceptable.

A.4.2. Distribution of the Learning Index

The distribution of the learning index in figure A.7(a) on the following page shows that many participants have not learned much in their example projects, which is not overly surprising – given that participants were asked to choose any example project or task, not

³For more on imputation, see appendix section A.5.2 on page 303.

A.4. Validity Investigation of the Learning Index in Detail

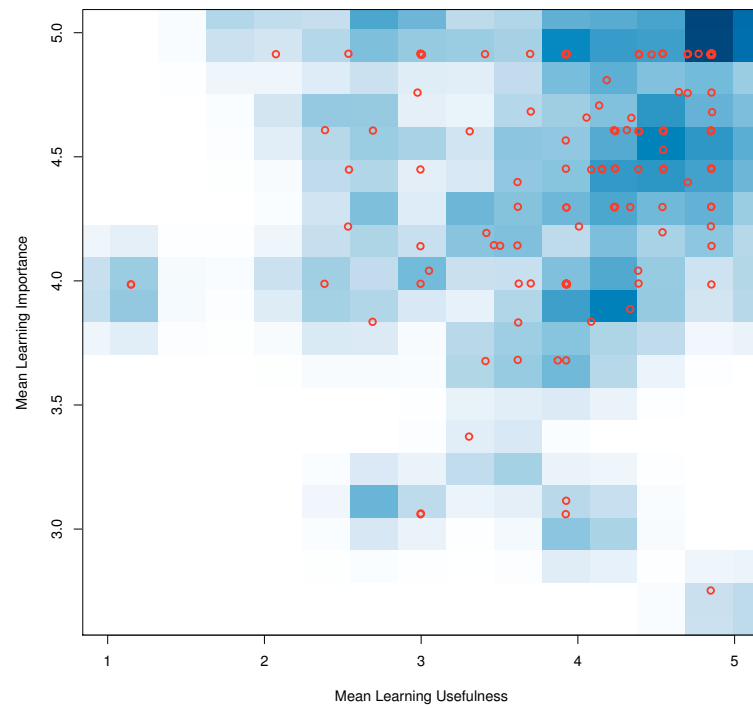


Figure A.6.: Relationship between Mean Learning Usefulness and Learning Importance

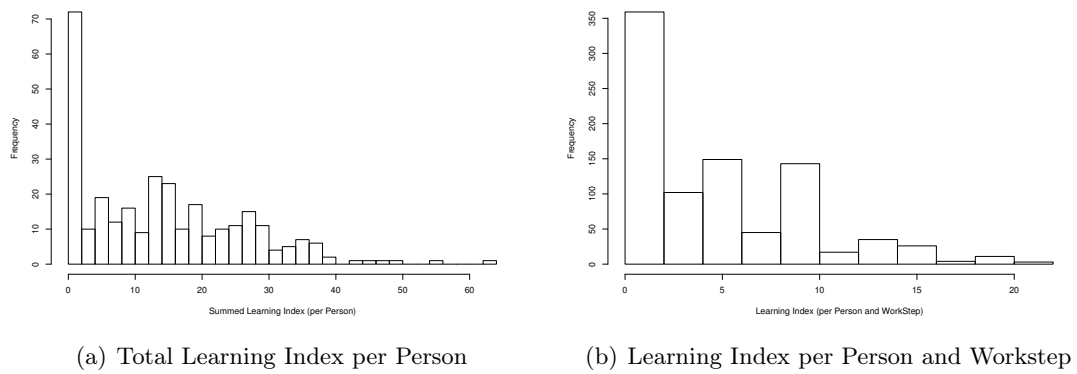


Figure A.7.: Distribution of the Learning Index

necessarily an extraordinarily learning-intensive example (see section 5.4.1 on page 146). Still, there is a substantial fraction of participants who had learning experiences.

The picture in figure A.7(b) on the facing page is more moderate, showing the (partial) learning index not summed over all work steps but just over the learning situations for each step. It shows that the learning index is dominated by many smaller learning episodes.

At first sight, it is striking that both distributions have a substantial amount of participants who did not cite any specific learning episodes for their example task and thus got a learning index equal to zero. Yet this is a good indication regarding the validity of the learning index, since it is not surprising that many tasks at the shipyard are performed without a consciously noticed learning effect. Therefore the large fraction of zeros is also an indication that the bias due to social desirability⁴ is small – especially when compared to the results from the much simpler question on the general learning effect without a link to a concrete learning episode (section A.4.3). Hence the context-specific link to learning episodes, which the participant can name, has achieved the desired ‘objectifying’ effect of bias reduction for the recollection of past events (section 5.4.1 on page 146).

Despite the many zero-learning cases, both distributions in figure A.4.2 on the facing page also have many small but non-zero learning cases, and thus the distributions appear to follow a continuous function similar to a log-normal distribution (with ρ around 1). Hence the distribution of the learning index has a shape that is common for many natural stochastic processes.

A.4.3. Cross-Validation of the Learning Index with Related Questions

General Learning Impression

Figure A.8 on the next page shows the dependence of the Learning Index vs. the General Learning Impression expressed in the following question (asked after the learning part):

Question for general learning impression (li.total):

“When you consider your <Task> as a whole in comparison with other tasks of work, did you learn much during <Task>?”

“Wenn Sie <Aufgabe> insgesamt betrachten und im Vergleich mit Ihren anderen Aufgaben bei der Arbeit setzen, haben Sie bei <Aufgabe> viel gelernt?”

[German Original]

The various features of figure A.8 on the following page are explained in text box 5.12.1 on page 166 and with more background and an annotated figure in section 7.1.3 on page 213.

The graph shows a weak dependence, while there should be a strong dependence. Noteworthy is the distribution of the general learning impression (shown in the lower part

⁴Social desirability is discussed in section 5.4.2 on page 150.

of fig. A.8), which is heavily centered on the scale midpoint and appears similar to a normal distribution.

The darker the blue in figure A.8, the denser the population of samples in the respective square. Hence this type of colorized plot visualizes the distribution of samples in this 2D-space better than a simple scatterplot (additionally provided by the grey circles).

The fact that the highest point density is located around 'average' for the general learning impression and a learning index below 10 suggests that a significant social desirability effect is acting on the simple general learning impression data. Still, there is a significant correlation of 0.421.

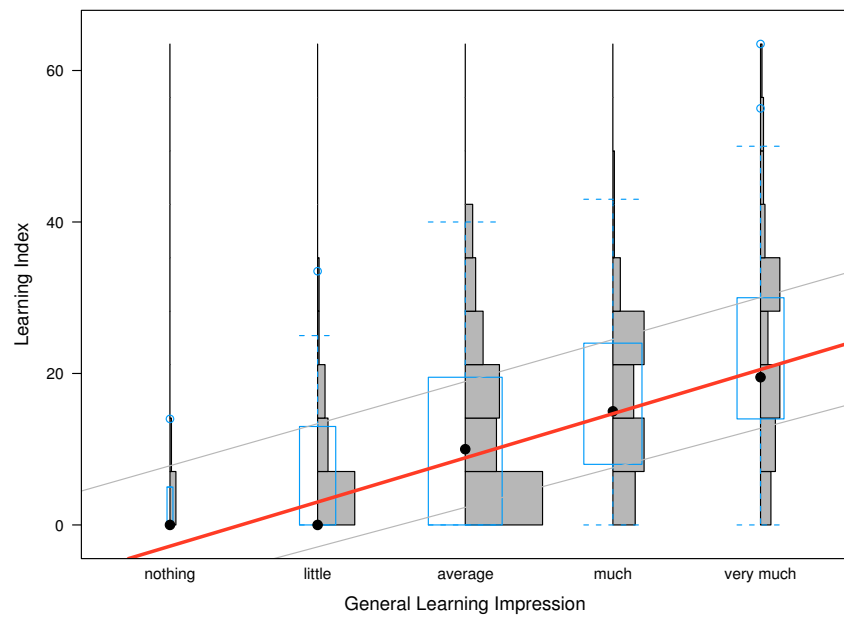


Figure A.8.: Distribution of Learning Index vs. General Learning Impression

Aside from the correlations:

- Correlation between the two real variables: 0.421
- Correlation between the two simulated variables: 0.043

figure A.8 suggests a linear but weak relationship. Aside from the social desirability effects, a strong linear relationship is expected from theory.

Average Post-Task Self-Efficacy

The question on post-task self-efficacy (Sicherheit) for similar tasks correlates highly with the learning index (see figure A.9 on the facing page) – a good sign.

Question for task assurance / self-efficacy:

“Did you become more self-assured while working on <Work Step> for similar jobs in the future?”

“Sind sie beim Bearbeiten des <Arbeitsschritt> für ähnliche Arbeiten in der Zukunft sicherer geworden?” [German Original]

This question is asked for each work step. For all valid work steps, the **mean** of these answers is calculated to obtain an overall self-efficacy increase for the task as a whole.

The correlation between the two real variables is 0.161.

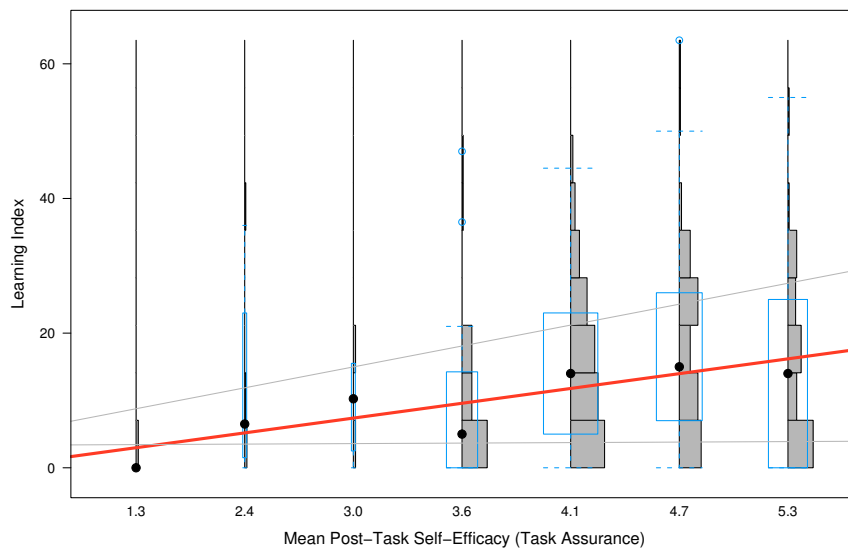


Figure A.9.: Distribution of Learning Index vs. Post-Task Self-Efficacy for each Person and Workstep

Summed Post-Task Self-Efficacy

In addition to the previous section, I investigate here whether the **summed** post-task self-efficacy is a better predictor for the learning index. As for the mean, the summed post-task self-efficacy is summed over all valid working steps and NA only if all working steps have this variable as NA.

The correlation between the two real variables is 0.342. As reasonable from theory, the learning index correlates highly, and the relationship in figure A.10 on the next page appears to be linear. This supports the learning index as a measure for learning.

Given that the summed Post-Task Self-Efficacy effectively (by the summing mechanism) includes the number of working steps (correlation = 0.735), it is not surprising

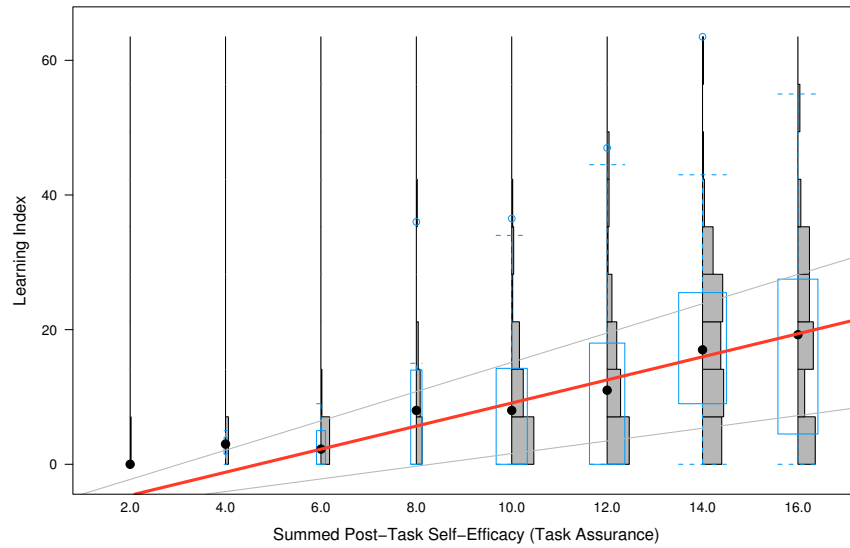


Figure A.10.: Distribution of Learning Index vs. Summed Post-Task Self-Efficacy

that – given the correlation between number of working steps and learning index (0.300) – the summed measure correlates more highly with the learning index (0.342) than the averaging measure (0.161).

Post-Task Self-Efficacy for each WorkStep

In the two previous sections, the post-task self-efficacy has been aggregated by either summing or averaging over the work steps in order to get a single measure for a person. To avoid this loss of data, in this section the data used is not aggregated by person, but instead data for each person and work step is used – thus $n = 3$ times the number of participants.

Since this measure does not include any side effects, such as the correlation with the number of working steps, it is a better test of the learning index than the summed or averaged measure. Correlation = 0.156.

See also figure [A.11 on the facing page](#).

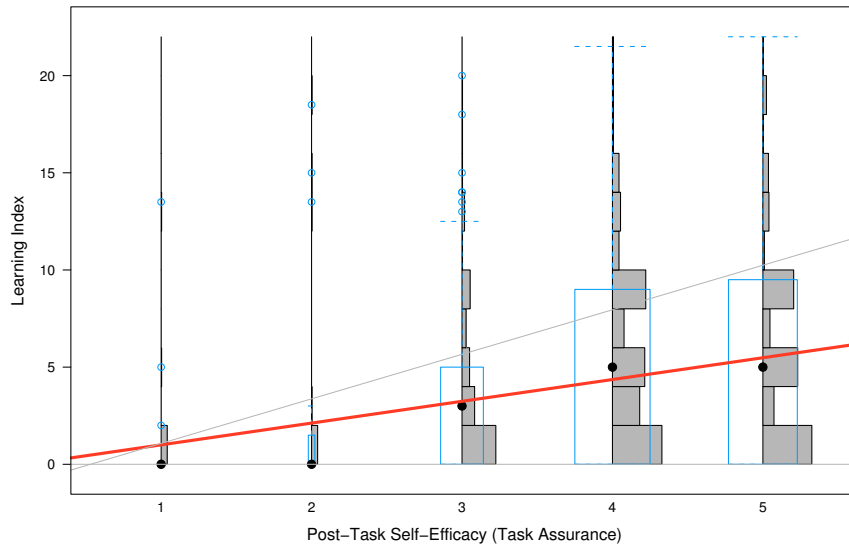


Figure A.11.: Distribution of Learning Index vs. Post-Task Self-Efficacy for each Person and Workstep

Mean Learning Strategy

For each task, after surveying for learning episodes, a question battery of six items starting with “*How much have you learned during <Task> by XYZ*” is posed to the participant, where XYZ is any of the following learning strategies: demonstration, discussion, analysis, reading, experimentation or other. The aim of these questions is to rank different methods of learning against each other. Hence the mean of the items for each person was subtracted from the items.

However, the mean of these question items is similar to asking six different facets of “*How much have you learned during <Task> in general?*” – thus it should highly correlate (like general learning) with the learning index. And as figure A.12 on the next page shows – it does correlate highly with 0.350.

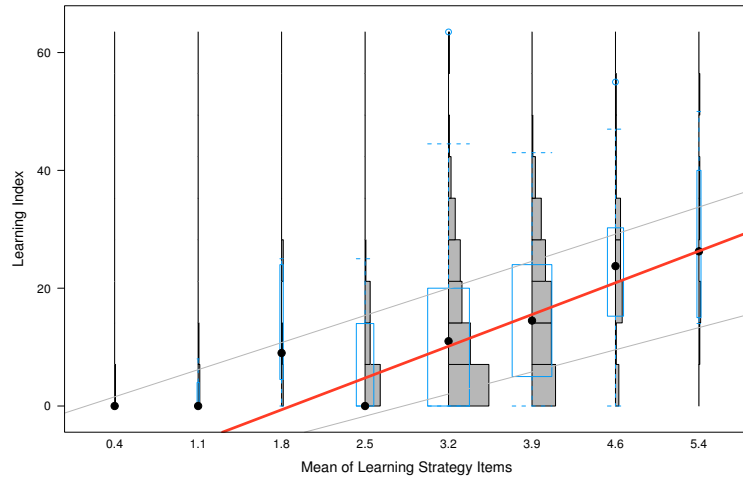


Figure A.12.: Dependence of Learning Index on Mean of Learning Strategy Items

A.5. Data Pre-Processing Details

A.5.1. Filtering and Outlier Removal

The raw dataset downloaded from the survey tool contains 1,583 rows of data – even though there were only 446 participants. On the other hand, improper outlier removal can falsify the results. Thus data pre-processing, including various stages of filtering, is documented in this section.

Filtering was performed in the following stages:

1. First, all data from the pilot phase and from **failed attempts** is removed by using date stamps and similar criteria – reducing the data down to 446 datasets.
2. As Chatfield (1995, p. 427) argues, the deletion of outliers leads to an overestimation of predictive power. Thus outliers should only be deleted on very strong subject matter grounds. Following this insight, there is no **outlier removal** in the classical sense – i.e., removing data points that appear to fall off very far from the fitted model in order to refit the model afterwards.

However, especially since the survey contains free text input fields in the learning frequency survey tool, the text entries were inspected manually and removed if it became clear that the participant did not understand or became uninterested in the survey⁵. Non-understandable answers (e.g., when participants used acronyms) were left in the data.

⁵For example, one participant entered ‘dog’, ‘cat’ and ‘mouse’ as learning opportunities.

In addition, and despite various precautions (by using additional validation questions – see [section 5.7 on page 153](#)), a number of participants clearly went into the wrong branch (innovation, large project, short repetitive task) judging from the text entries. Hence their data is not valid and needed to be removed.

All these removals, which all are solidly based on theoretical grounds rather than the model fit, lead to a further reduction of the data down to 329 datasets.

3. For simplicity of the analysis, the datasets of 27 **apprentices** were **removed**, since their exposure to the organization was not long enough yet and their working environment and embedding in the organization is different from an ordinary employee. Hence the learning experiences of apprentices are likely to be driven by other factors (or the same factors with different effect strengths).
4. As detailed further below, datasets with very high numbers of **missing values** were removed, leading to a further reduction down to 292 datasets. The criteria for this filtering step were: the dataset of a person must not have more than 40% missing values (NA) – see also the actual distribution of missing data per person in [figure A.13 on the next page](#). One case also had a missing value for the outcome variable (learning index), which was removed as well.

The number of variables were also pruned at this step. When a variable had more than 20% missing values, it was removed from the dataset – see actual distribution in [figure A.14 on page 305](#).

The result of all these filtering steps are 292 high-quality datasets – from participants who appeared to have understood the survey.

A.5.2. Imputation and Missing Value Filtering

Given that there are multiple paths through the survey, and some of the questions are mutually exclusive, not a single participant could provide a 100% complete dataset. Yet the amount of missing values for most participants lies well below 20% – see [figure A.13 on the next page](#). This is a good result, considering that the automatic survey-reduction mechanism may reduce the number of questions posed ([section 5.7 on page 153](#)) and thus introducing missing values for those questions that were not posed. Also when considering the missing values *by variable*, the results are good: most variables have 20% or fewer missing values – as shown in [figure A.14 on page 305](#).

Even though the fraction of missing values is acceptably small, most algorithms require a 100% complete dataset. Thus imputation of the missing data with some neutral replacement value is necessary in this case without any 100% complete datasets.

Current statistical literature and software offers a number of advanced imputation algorithms, e.g., the Expectation-Maximization Algorithm (EM) ([Rueda et al., 2005](#)),

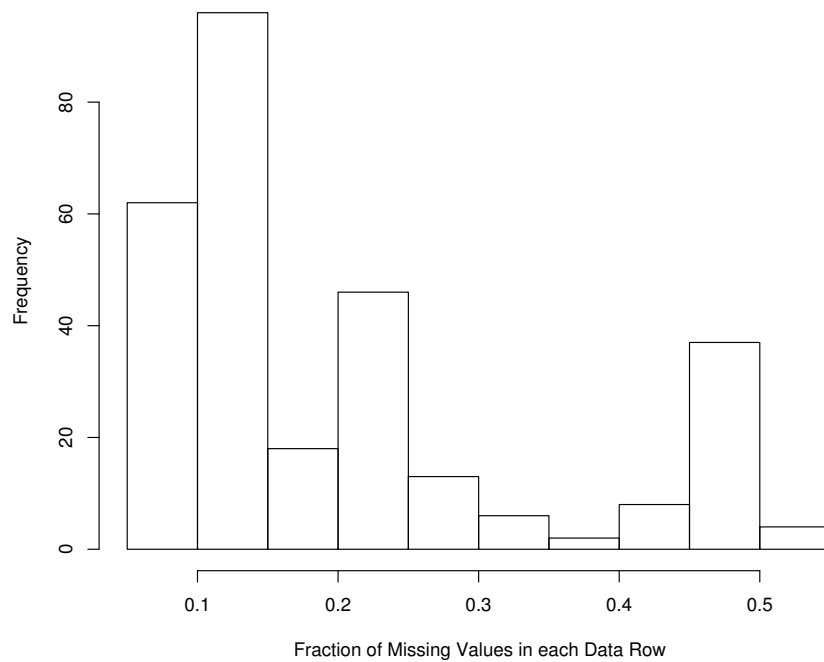


Figure A.13.: Distribution of the Fraction of Missing Values per Data Row (Participant)

Bayesian Imputation (MICE) (van Buuren, 2008) and rfImpute – RandomForest-based imputation (Liaw and Wiener, 2002). All of these algorithms aim to make a prediction for the missing value by using any of the other variables as input.

Hence using any of these imputation algorithms might improve the results but may also introduce additional risks, since the results of the following statistical analysis also relies on the correct behavior of the imputation algorithm.

In experiments with these advanced algorithms, it became clear that the correctness of the imputation behavior is not easy to assess. Given the increased risk introduced by complicated imputation algorithms (leading to reduced robustness of the entire process), I decided to use one of simplest and most predictable imputation methods: **mean imputation**.

In mean imputation, missing values are replaced with the mean of the respective variable. Inserting the mean of a variable effectively replaces the missing value with the most neutral value of that variable. Thus mean imputation will weaken any effects of this variable that are visible in the data. Hence if the statistical analysis finds an effect for a particular variable, the effect must have been in the data and could not have been introduced in the imputation process. Even though mean imputation thus also leads

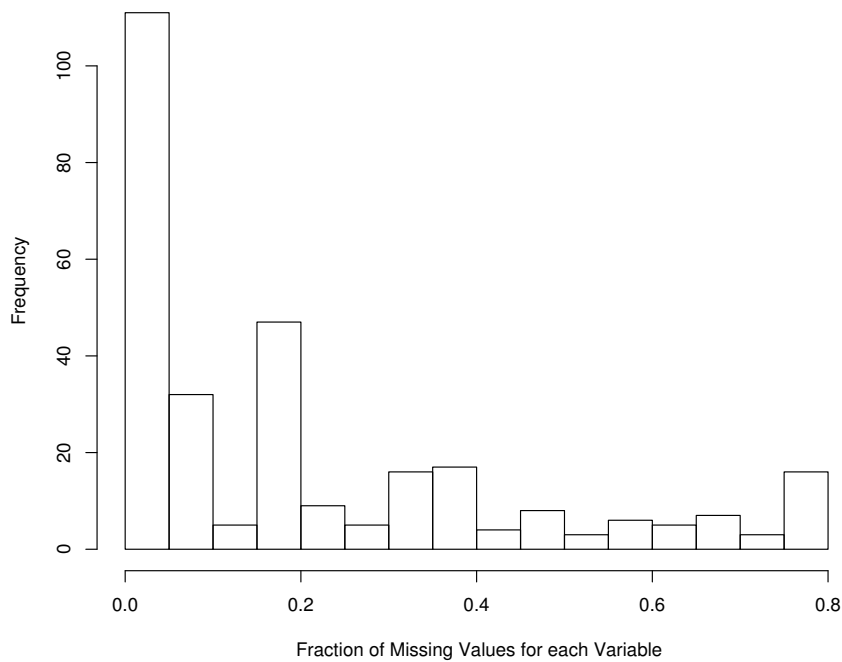


Figure A.14.: Distribution of the Fraction of Missing Values per Variable

to an underestimation of variable effects (i.e., parameter strength in regression), it is a conservative and robust approach. Hence robustness is traded for sensitivity in detecting an effect.

Given its importance, no imputation was used for the outcome variable (the learning index). Instead, the case with a missing value for the outcome variable was removed from the data.

In summary, due to the interactive survey process, there is no single 100% complete dataset. The data has a substantial but non-critical fraction of missing values. Since most algorithms require 100% complete datasets, imputation was performed with one of the most simple, predictable and conservative methods: mean imputation.

A.6. Details on BOGER

A.6.1. Generating Data Frequency Equalized Bootstrapping Samples

For each bootstrapping iteration, random and “equalized” training and test bootstrapping samples are generated with the following steps:

1. The data is split into a few (e.g., 10) bootstrapping sample groups⁶) by labelling each data point (i.e., each data line in the sample database) randomly with bootstrapping sample group number (e.g., from 1 to 10). From these groups, bootstrapping samples will later be assembled. The algorithm parameter *bootstrapping fraction* determines how many groups will be assembled for each iteration to a training bootstrapping sample. The remaining groups will be assembled to a test bootstrapping sample for model fit cross-validation. For a bootstrapping fraction of 0.7, the training sample of iteration 1 would be assembled by seven groups and the test sample of the three remaining groups.
2. To prepare the assembling of the training and test bootstrapping samples for each iteration, BOGER generates all unique combinations of training groups. In the above example, it would generate all unique combinations of seven groups drawn from 10 available groups. The test groups are automatically determined as the remaining groups (in the example, this would be the three non-selected groups for each combination).
3. The algorithm is commonly used with a number of groups that leads to a number of unique combinations that is higher than the requested number of bootstrapping iterations. With more unique group combinations than bootstrapping iterations, BOGER automatically chooses the unique group combinations in order to level out the frequency of use of each group for the training data – as much as possible. The aim of this procedure is that each data point of the sample data is used as equally often for training and testing as any other data point in the bootstrapping iterations – as far as possible. Hence the equalization is not perfect but still good. Thus the genetic solver from section 6.2.3 on page 183 is used to equalize the frequency of each group for training in the bootstrapping iterations as far as possible by optimization. Since the test groups are the exact complement of the training group choice, equalizing the training group choice also equalizes the test group choice.
4. Based on the equalized sub-set of group combinations, for each iteration bootstrapping training and test samples are generated. Since each data point was assigned randomly to a group and the group choice for the training and test bootstrapping samples is frozen at this point, this step is now fully deterministic. Note that despite the deterministic equalization by optimization in the previous step, the resultant composition of the bootstrapping samples (both training and test) is still random, since the assignment of bootstrapping group to data point is random. At this point BOGER has for each of the n_{boot} bootstrapping iterations, a training and a test

⁶The number of bootstrapping sample groups mostly depends on the number of requested bootstrapping iterations (n_{boot}).

bootstrapping sample, which is equalized for data point frequency but still random.

Some bootstrapping algorithms generate training samples that have as many data points as the original sample by using a few data points multiple times. This process is called *bootstrapping with replacement* (Strobl et al., 2007). Such an approach is needed if the model fit result depends directly on the sample size, and thus sample size needs to be kept constant for comparison. Since BOGER's results are not directly dependent on the sample size, *bootstrapping without replacement* is used, leading to a smaller training sample size. In BOGER, combining a training and a corresponding test bootstrapping sample yields the original data sample (only with a new and random sequence).

The user may set the relative sizes of the training and the test bootstrapping samples. Common sizes are: the training bootstrapping sample has 70% of the size of the original sample, while the test sample contains the remaining 30% of the data points.

A.6.2. Implementation Details of BOGER in \mathbb{R}

For use in this study, the BOGER algorithm is implemented in the statistical high-level language \mathbb{R} – using its object-oriented extensions (S4). \mathbb{R} was also used to filter and pre-process the data and in conjunction with L^AT_EX to generate automated reports that facilitated analysis and the debugging of the algorithm.

Reasons for Choosing \mathbb{R} \mathbb{R} was chosen for a number of reasons:

- \mathbb{R} is a flexible modern high-level language (including object-oriented features), which was designed specifically for statistics (R Development Core Team, 2007) and is popular in research on statistics and bio-informatics (Gentleman et al., 2004). Some readers may know MATLAB, Scilab or Octave, which are remotely similar scripted matrix calculation languages. In contrast to statistical software such as SPSS or Stata, \mathbb{R} was designed as a programming language rather than a software package that can be scripted.
- Anything in \mathbb{R} can and must be scripted. Thus any new user faces a steep learning curve, but once overcome, scripting can automate many analysis steps and thus allows for much more experimentation. In addition, it is difficult to use \mathbb{R} mindlessly without understanding it – which is not the case with GUI driven statistical software.
- It is an interpreted language, which removes the need for a compilation step, makes the language and debugging simpler. This makes \mathbb{R} slower than most compiled languages such as C or C++. Yet \mathbb{R} code be optimized for speed by using matrix operations instead of loops and by using linked binary modules that can be written, e.g., in C or C++. Thus \mathbb{R} is still faster than many other scripted languages and software such as MS Excel.

- It is open-source⁷ and runs on many different operating systems. The openness of the source and the licence allows anybody to develop extensions for \mathbb{R} – as simple extension packages or also at the core of \mathbb{R} with the following effect:
- There exist 1,700 extension packages for \mathbb{R} – of which the large majority is open-source as well. The additional features range from simple calculation functions to various choices of interfaces for computer clusters for high performance parallel computing.
- The graphics generating capabilities are very feature rich, flexible and modular enough to allow further extension by the user⁸.
- Rather than displaying many graphics in individual windows and tables again separately, \mathbb{R} code can be embedded in so-called weave `.rnw` files, which mix \mathbb{R} and \LaTeX code. This mixed code is then first run through the Sweave package⁹ (Leisch, 2002), which executes the \mathbb{R} code to perform the calculations and to automatically create PDF and \LaTeX files for the generated figures and tables. The output of this step is a plain \LaTeX file, which contains the documentation contents of the `.rnw` file and refers to the figure PDFs and table `.tex` files. In a final step, these files are compiled by the \LaTeX compiler to a single high-quality PDF file – including all of the advanced document-generation features of \LaTeX , e.g., bibliography references or formulas. Most of the figures and tables in this thesis are automatically generated using these steps. Thus this combination of open-source software packages provides the functionality of flexible and high-quality automated report generation.
- \mathbb{R} code is stored in simple text files and thus will be (human-) readable for many decades without needing to run \mathbb{R} .

Details of the BOGER Implementation The BOGER algorithm (without data preparation or analysis) was implemented in about 4,000 lines of \mathbb{R} code (that is equivalent to about 200 single-spaced pages).

Performance-critical portions of the code are execution speed-optimized by using the built-in (and highly optimized) matrix operations of \mathbb{R} . For an illustration and general impression, see the screenshot in figure A.15 on the next page. (`sapply` is used as accelerated matrix command instead of a for-loop.)

Given the computational intensity of the task, the BOGER algorithm was parallelized to allow for parallel model fitting of different bootstrapping samples on different CPUs or

⁷The open-source licence GPL v2.1 guarantees that one can freely use one's code for a very long time – i.e., until there is nobody anymore maintaining the language for modern computer systems.

⁸The \mathbb{R} user community constantly adds new types of graphs. For example, the user-run [R Graphics Gallery](http://addictedtor.free.fr/graphiques/) at <http://addictedtor.free.fr/graphiques/> features 160 types of graphics.

⁹The author of this thesis has modified and extended the Sweave package.

```

if(bootstrap) {
  if(loadScreening || loadIntermediate)
    screening <- FALSE

  nIter <- ((screening.redundancy + 1) * screening) + 1 + stabilizationStage
  finalIter <- (if(stabilizationStage) c(nIter-1, nIter) else nIter)

  # save user desired bootstrap.samples in case it's loaded from bootstrap*.xdr
  userBootstrap.samples <- bootstrap.samples

  if(!is.logical(screening)) {
    iterCf <- list(bootFrac = rep(bootstrap.frac, nIter),
                  bootSmpl = c(rep(round(bootstrap.frac^2*bootstrap.samples/2)*2, nIter-2),
                                rep(bootstrap.samples, 2)),
                  screenFrac = rep(screening.frac, nIter),
                  paramActive = list())

    iterCf$screenFrac[finalIter] <- 1

    # assign random screening groups to non-core variables (not parameters!)
    varIDsLong <- sort(unique(dataObj@paramVarIDs[dataObj@paramActive]))
    coreVars <- sapply(varIDsLong, function(vid) {
      any(dataObj@coreParams[dataObj@paramVarIDs %in% vid]) })
  }
}

```

Figure A.15.: BOGER \mathbb{R} Code in the Editor Emacs

even different computers. The parallel execution of the model fitter was implemented with the `snowFT` package for \mathbb{R} (Sevcikova and Rossini, 2005), which provides simple cluster computing features.

The interpreted and functional nature of the language allowed another performance optimization: the expression for calculating the prediction of the BOGER mathematical model with a particular term configuration is regenerated and parsed as an \mathbb{R} expression for each iteration. Thus a model with 30 active terms contains also only 30 terms in \mathbb{R} and not 70 terms with parameters set to the neutral position (zero) and 30 truly active parameters.

Software Features of the BOGER Implementation A number of BOGER's software features greatly facilitated the use and the debugging of BOGER:

- There is a history or logging function that stores and can restore any previously tested model, including the test results from a data file.
- This history function also allows the user to restart BOGER at various points within a long calculation run, e.g., if the algorithm has crashed or the results at some point of the development are not as desired.
- The bootstrapping samples are generated randomly but then can be frozen (i.e., stored in a data file), to allow for precise comparison of the algorithm results without

having to wonder if slight changes in the results are due to changes in the algorithm or in the bootstrapping samples.

- The user interaction is provided by a command line. While not the most fashionable form of user interface, it gives the user much flexibility to analyze intermediate results even while running BOGER.
- Each variable can be tagged with multiple variable groups, which greatly facilitates filtering groups of variables by this tag. While tags have no hierarchy, a variable can have multiple tags and thus belong to multiple variable groups.

Thus \mathbb{R} turned out to be a good choice of statistical software, given the complexity of the statistical modelling task and the resulting need for flexibility, custom visualization, speed-optimization as well as iterative analyzing (by automatic analysis report generation).

A.6.3. Flexible Model Fitting - an Interesting Accident

At the end of the full (and 6.5-day long) screening run, an interesting accident occurred¹⁰:

Due to a bug in step 5 of figure 6.1 on page 186, the Pre-Selection Model (step 6 of fig. 6.1) did not contain a much reduced set of parameters (using the by-parameter instability measure) – as desired – but instead contained 502 parameters (i.e., almost all parameters).

Given the large number of parameters that needed to be optimized for fit by the genetic algorithm, the time to fit an individual bootstrapping model rose from 5–10 min to 10–24 hrs.

The large number of variables (502) compared to a sample of 292 (see section A.5.1 on page 302) leads to a very flexible model – despite the parametric nature of BOGER’s math model (eq. 6.1 on page 182). As theoretically explained in section 4.2.2 on page 123, this high model flexibility leads to a high risk of overfitting – which is confirmed by the model fit results shown in table A.2.

	Training Fit	Test Fit
R_{abs}	0.163	0.134
R_{abs}	0.322	0.223
R_{abs}	0.428	0.205

Table A.2.: Accident Model Fit Summary

¹⁰The accident was later rectified by restarting a debugged version of BOGER with the screening data generated so far (stored with the history feature) just before step 5 in figure 6.1 on page 186. Hence the accident did not affect the final results.

Thus as expected, with too many variables compared to the sample size, also BOGER performs poorly. Yet with a reasonably small number of parameters, BOGER shows much less overfitting – as the final model fitting results with approximately 4% overfitting in table 6.4 on page 200 show.

A.6.4. Residuals

Even though the BOGER algorithm does not require the assumption of homoscedacity¹¹, residual¹² plots are useful to detect non-linearities. Figures A.16 on the next page and A.17 on page 313 show the same residuals as scatterplots and as distributions (frequency plots). Overall, the residuals appear independent of the independent variables and often even normally distributed, which suggests that the BOGER model has treated the non-linearities in the data with sufficient accuracy.

¹¹*Homoscedacity* means the variance of the residuals is constant for (i.e., not a function of) any of the independent variables.

¹²The vector of residuals is the vector of differences between an individual model estimation and the corresponding true (i.e., surveyed) value. This residual vector is plotted against the corresponding independent variables in scatterplots and distribution (i.e., frequency) plots.

A.6. Details on BOGER

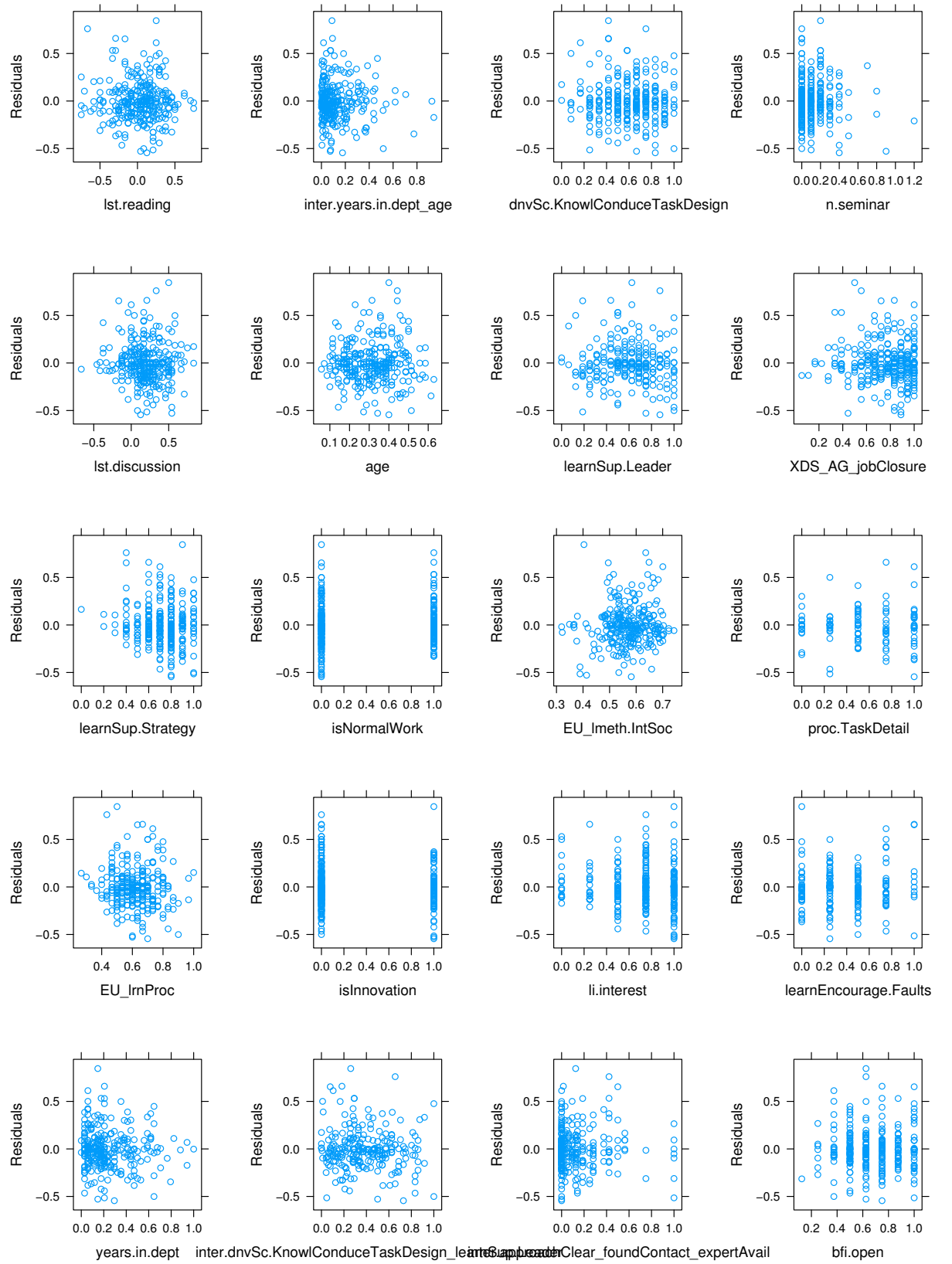


Figure A.16.: Residual Scatterplots for each Independent Variable (in the final model)

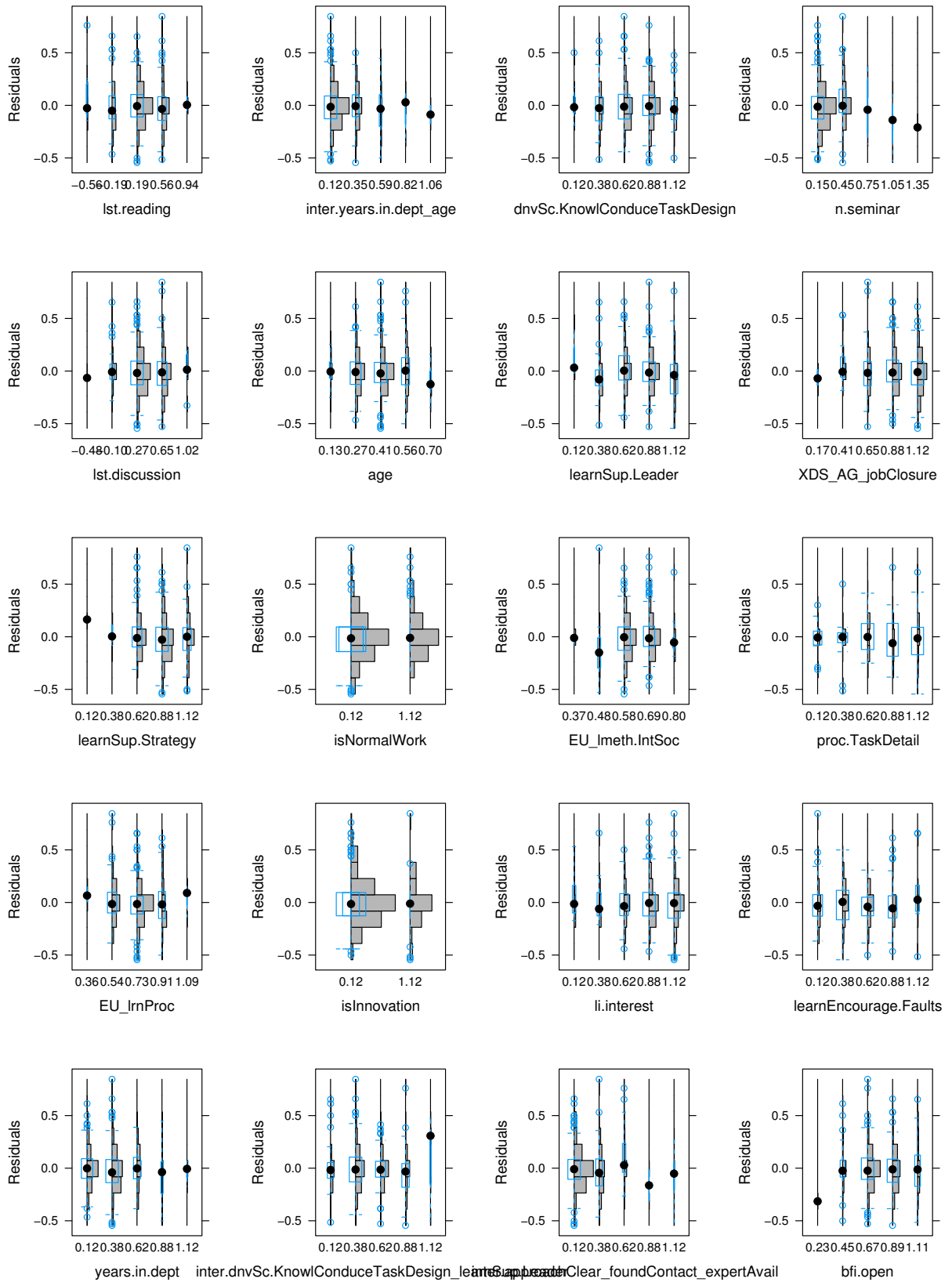


Figure A.17.: Residual Distributions for each Independent Variable (in the final model)

A.6.5. Empirical Robustness of Model Fit Measures

This section provides an empirical analysis of the robustness of R^2 vs. R_{abs} and the different biases of the training and test estimates by comparing the model fit measures as distributions over the different individual bootstrapping models.

The BOGER model fit results in table 6.4 on page 200 for the survey data are somewhat surprising, since the R^2 estimate for the individual models and the internal test data is higher than the corresponding R^2 estimate for the training data. That would imply underfitting¹³, i.e., the model fits the data worse than the real underlying statistical process. The equivalent estimates for R_{abs} , however, suggest the opposite: the internal test fit and thus the predictive power is lower than the fit to the training data, and thus the model slightly overfits the data.

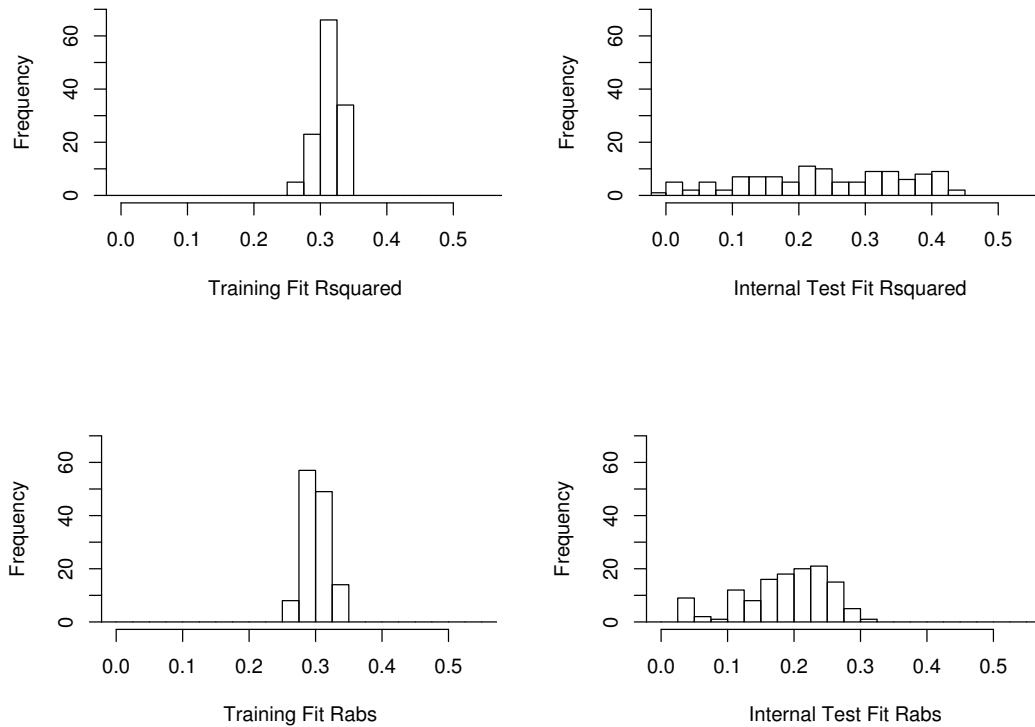


Figure A.18.: Distributions of the different R^2 and R_{abs} estimates – for all fitted individual models

Hence the question arises: which measure – R^2 or R_{abs} – is more accurate and robust?

¹³Underfitting is the opposite of overfitting. For more details on overfitting and the conditions under which it occurs, see section 4.2.2 on page 123.

Figure A.18 on the preceding page, showing the distributions of the two measures over the different individual models in the bag, gives a clear answer. For the internal test data, R^2 (top row of graphs) has a rather uniform distribution without any visible peak, while R_{abs} (bottom row of graphs) has a distribution that resembles a skewed normal distribution with much less variance than that of R^2 . The variance for R_{abs} and R^2 estimated based on the training data is very similar, however. These qualitative results remain the same even when the independent variables in the model are changed or when the number of individual models in the bag is increased.

Thus these results are further evidence for the claim from section 6.2.7 on page 195 that R_{abs} is more robust than R^2 for measuring model fit using the internal test data. As mentioned in section 6.2.7 on page 195, the square function in R^2 (purposely) amplifies outliers, i.e., those samples with a large deviation from the model and thus large residuals. R_{abs} , using the absolute value function, treats all samples in the same manner, independent of the magnitude of their residual. Since the effect is not strong on the larger training dataset but occurs only on the internal test dataset (only 30% the size of the training dataset), R_{abs} seems to be more robust and accurate than R^2 only when the sample size is very small.

Note that figure A.18 on the facing page shows the model fit estimates for all fitted individual models based on different bootstrapping data but based on the same set of independent variables. As described in section 6.2.6 on page 192, the bagged model contains only a subset of “good” models: the best 25% of all bootstrapping models. The distributions of the model fit estimates for the “good” models have much more defined peaks and less variance – as figure A.19 on the following page shows¹⁴. Thus any measure estimated by the average of the “good” model’s fit will be rather accurate with both measures R^2 and R_{abs} .

Furthermore, the comparison of figures A.18 on the preceding page and A.18 on the facing page illustrates the effectiveness of the model filtering strategy: the bagged model quality increases by filtering for the “good” bootstrapping models.

Moreover, the two figures empirically support the claim regarding different biases and thus the claim for robustness of BOGER’s model fit estimate from section 6.2.7 on page 195. Given the lower sample size of the test data, the test model fit has a higher variance and thus is more affected by the sampling bias than the training data, while the training data is more affected by biases linked to overfitting (as argued in other sections).

Finally, the low variance of both the training and the test model fit estimates for the filtered “good” models (in figure A.19 on the following page) further supports the robustness of the model predictive power estimation method described in section 6.2.7 on page 195 – leading to a positive quality assessment of the final BOGER model in

¹⁴Since the “good” models have been selected by a combination of internal training data fit and internal test data fit, it is not very surprising that these graphs have more defined peaks and less variance.

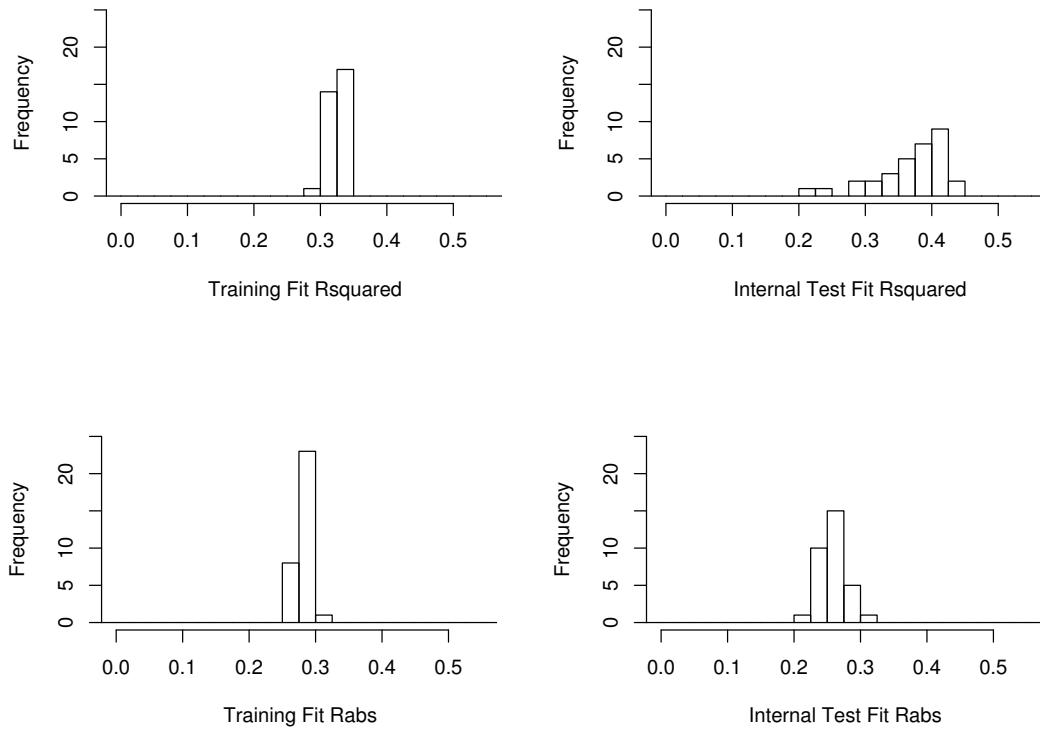


Figure A.19.: Distributions of the different R^2 and R_{abs} estimates – only for the individual filtered (“good”) models in the bag

section 6.3.2 on page 199.

Note that BOGER internally exclusively uses R_{abs} – see section 6.2.7 on page 195.

Bibliography

- Abele, A. E., Stief, M., and Andrä, M. S. (2000). Zur ökonomischen erfassung beruflicher selbstwirksamkeitserwartungen - neukonstruktion einer bsw-skala. *Zeitschrift für Arbeits- und Organisationspsychologie A&O*, 44(3):145–151.
- Abou-Zeid, E.-S. (2002). A knowledge management reference model. *Journal of Knowledge Management*, 6(5):486–499.
- Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16:3–9.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.
- Al-Mashari, M., Zairi, M., and Ginn, D. (2005). Key enablers for the effective implementation of QFD: a critical analysis. *Industrial Management & Data Systems*, 105(9):1245–1260.
- Anderson, D. R. and Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management*, 66(3):912–919.
- Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management*, 64(4):912–923.
- Anderson, J. R. (1988). *Kognitive Psychologie – Eine Einführung*. Spektrum der Wissenschaft, Heidelberg.
- Anderson, J. R. (1990). *Cognitive psychology and its implications*. W. H. Freeman/Times Books/ Henry Holt & Co., 3rd edition.
- Anderson, N. R. and West, M. A. (1998). Measuring climate for work group innovation: Development and validation of the team climate inventory. *Journal of Organizational Behavior*, 19(3):235–259.
- Andriessen, D. (2004). *Making Sense of Intellectual Capital – Designing a Method for the Valuation of Intangibles*. Elsevier Butterworth-Heinemann. ISBN: 0-7506-7774-0.
- Andriessen, D. (2006). On the metaphorical nature of intellectual capital: A textual analysis. *Journal of Intellectual Capital*, 7(1). Special issue: 'Becoming Critical'.

- Arbnor, I. and Bjerke, B. (1997). *Methodology for Creating Business Knowledge*. Sage Publications.
- Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260.
- Argote, L. (1999). *Organizational Learning. Creating, Retaining and Transferring Knowledge*. Kluwer Academic Publishers. ISBN 0-387-22581-1.
- Argote, L., McEvily, B., and Reagans, R. (2003). Managing knowledge in organizations: An integrative framework and review of emerging themes. *Management Science*, 49(4):571–582.
- Argyris, C. (2002a). Double-loop learning, teaching, and research. *Academy of Management Learning & Education*, 1(2):206–218.
- Argyris, C. (2002b). Teaching smart people how to learn. *Reflections*, 4(2):4–15.
- Association, A. P. (2002). *Publication Manual of the American Psychological Association*. American Psychological Association.
- Atkinson, A. C. and Riani, M. (2007). Building regression models with the forward search. *Journal of Computing & Information Technology*, 15(4):287–294.
- Augier, M. and Knudsen, T. (2004). The architecture and design of the knowledge organization. *Journal of Knowledge Management*, 8(4):6–20.
- Backhaus, K., Erichson, B., Plinke, W., and Weiber, R. (2006). *Multivariate Analysemethoden*. Springer, 11th edition.
- Badke-Schaub, P. and Frankenberger, E. (2004). *Management kritischer Situationen: Produktentwicklung erfolgreich gestalten*. Springer, Berlin. ISBN 3-540-43175-6.
- Badke-Schaub, P., Neumann, A., Lauche, K., and Mohammed, S. (2007). Mental models in design teams: a valid approach to performance in design collaboration? *CoDesign*, 3(1):5–20.
- Badke-Schaub, P. and Strohschneider, S. (1998). Complex problem solving in the cultural context. *Travail Humain*, 61(1):1–28.
- Baltes, P. B. and Staudinger, U. M. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology*, 50(1):471–.
- Bandura, A. (1997). Self-efficacy. *Harvard Mental Health Letter*, 13(9):4–8.

- Bartlett, C. A. and Ghoshal, S. (1990). Matrix management: Not a structure, a frame of mind. *Harvard Business Review*, 68(4):138–145.
- Becker, H. (2006). *Phänomen Toyota : Erfolgsfaktor Ethik*. Springer, Berlin.
- Beer, M., Eisenstat, R. A., and Spector, B. (1990). Why change programs don't produce change. *Harvard Business Review*, 68(6):158–166.
- Berings, M. G. M. C., Poell, R. F., and Simons, P. R.-J. (2005). Conceptualizing on-the-job learning styles. *Human Resource Development Review*, 4(4):373–400.
- Bernstein, D. M., Atance, C., Meltzoff, A. N., and Loftus, G. R. (2007). Hindsight bias and developing theories of mind. *Child Development*, 78(4):1374–1394.
- Berson, Y., Nemanich, L. A., Waldman, D. A., Galvin, B. M., and Keller, R. T. (2006). Leadership and organizational learning: A multiple levels perspective. *Leadership Quarterly*, 17(6):577–594.
- Best, W. D. (2006). Critical chain project management flies.
- Bevilacqua, M., Ciarapica, F., and Giacchetta, G. (2009). Critical chain and risk analysis applied to high-risk industry maintenance: A case study. *International Journal of Project Management*, 27(4):419–432.
- Bevilacqua, M., Ciarapica, F. E., and Giacchetta, G. (2008). Value stream mapping in project management: A case study. *Project Management Journal*, 39(3):110–124.
- Bierhals, R., Schuster, I., Kohler, P., and Badke-Schaub, P. (2007). Shared mental models – linking team cognition and performance. *CoDesign*, 3(1):75–94.
- Binner, H. F. (2008). Wissensbewahrung – Eine Gesellschaftliche Herausforderung. published on the web in 'Wissensmanagement Online'.
- Blackler, F., Crump, N., and McDonald, S. (2000). Organizing processes in complex activity networks. *Organization*, 7(2):277–300.
- Boisot, M. and Canals, A. (2004). Data, information and knowledge: Have we got it right? *Journal of Evolutionary Economics*, 14(1):43–67.
- Boldini, A., Russo, R., and Avons, S. E. (2004). One process is not enough! a speed-accuracy tradeoff study of recognition memory. *PBR*, 11(2):353–361.
- Borkenau, P. and Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Hogrefe, Göttingen.

- Bornemann, M. and Alwert, K. (2007). The german guideline for intellectual capital reporting: method and experiences. *Journal of Intellectual Capital*, 8(4):563–576.
- Bornermann, M. (1999). Potential of value systems according to the vaic method. *International Journal of Technology Management*, 18(5-8):463–476.
- Bou-Llusar, J. C., Escrig-Tena, A. B., Roca-Puig, V., and Beltrán-Martín, I. (2009). An empirical assessment of the EFQM excellence model: Evaluation as a TQM framework relative to the MBNQA model. *Journal of Operations Management*, 27(1):1–22.
- Box, G. (1994). Statistics and quality improvement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(2):209–229.
- Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley & Sons. ISBN-13: 978-0471810339.
- Brambor, T., Clark, W. R., and Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, 14(1):63–82.
- Brehmer, B. (2005). Micro-worlds and the circular relation between people and their environment. *Theoretical Issues in Ergonomics Science*, 6(1):73–93.
- Brehmer, B. and Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior*, 9(2):171–184.
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754.
- Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24(2):123–140.
- Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383.
- Breiman, L. (1998). Arcing classifier. *Annals of Statistics*, 26:801–849.
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.
- Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.
- Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review / Revue Internationale de Statistique*, 60(3):291–319.

- Brown, J. S. and Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovating. *Organization Science*, 2(1):40–57.
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press., 2nd edition.
- Burkhard, C. (2006). *TQM-Trend-Matrix – Methode zur prognostischen Analyse unternehmensspezifischer Wirkungen von TQM-Maßnahmen*. PhD thesis, Technischen Universität Cottbus.
- Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.
- Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.
- Butler, D. L. and Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3):245–282.
- Cabrera, A., Collins, W. C., and Salgado, J. F. (2006). Determinants of individual engagement in knowledge sharing. *International Journal of Human Resource Management*, 17(2):245–264.
- Cacioppo, J. T. and Petty, R. E. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, 119(2):197–254.
- Carlile, P. R. (2002). A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organization Science*, 13(4):442–455.
- Carlile, P. R. (2004). Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries. *Organization Science: A Journal of the Institute of Management Sciences*, 15(5):55–69.
- Carlile, P. R. and Rebentisch, E. S. (2003). Into the black box: The knowledge transformation cycle. *Management Science*, 49(9):1180–1195.
- Cattell, R. B. (1971). *Abilities: Their Structure, Growth, and Action*. Houghton Mifflin, Boston, MA.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466.

- Chatfield, C. (2002). Confessions of a pragmatic statistician. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(1):1–21.
- Chiou, W.-B. and Wan, C.-S. (2007). The dynamic change of self-efficacy in information searching on the internet: Influence of valence of experience and prior self-efficacy. *Journal of Psychology*, 141(6):589–603.
- Clark, J. (2005). Explaining learning: From analysis to paralysis to hippocampus. *Educational Philosophy & Theory*, 37(5):667–687.
- Cleary, P. D. and Kessler, R. C. (1982). The estimation and interpretation of modifier effects. *Journal of Health and Social Behavior*, 23(2):159–169.
- Coffey, J. W. and Hoffman, R. R. (2003). Knowledge modeling for the preservation of institutional memory. *Journal of Knowledge Management*, 7(3):38 – 52.
- Cohen, W. M. and Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1):128–152.
- Collins, J. (2001a). *Good to Great: Why Some Companies Make the Leap... and Others Don't*. Collins Business, 1st edition. ISBN: 978-0066620992.
- Collins, J. (2001b). Level 5 leadership. *Harvard Business Review*, 79(1):66–76.
- Colonia-Willner, R. (1999). Investing in practical intelligence: Ageing and cognitive efficiency among executives. *International Journal of Behavioral Development*, 23(3):591–614.
- Colquitt, J. A., LePine, J. A., and Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, 85(5):678–707.
- Cook, S. D. and Brown, J. S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science*, 10(4):381–401.
- Cross, R., Parker, A., Prusak, L., and Borgatti, S. P. (2001). Knowing what we know: Supporting knowledge creation and sharing in social networks. *Organizational Dynamics*, 30(2):100–120.
- Crossan, M. M., Lane, H. W., and White, R. E. (1999). An organizational learning framework: From intuition to institution. *Academy of Management Review*, 24(3):522–537.

- Dahl, T. I., Bals, M., and Turi, A. L. (2005). Are students' beliefs about knowledge and learning associated with their reported use of learning strategies? *British Journal of Educational Psychology*, 75(2):257–273.
- Dalmedico, A. D. (2004). Early developments of nonlinear science in soviet russia: The andronov school at gor'kiy. *Science in Context*, 17(1-2):235–265.
- Davenport, T. H. and Beck, J. C. (2001). *The Attention Economy: Understanding the New Currency of Business*. Accenture.
- Davenport, T. H., Harris, J. G., De Long, D. W., and Jacobson, A. L. (2001). Data to knowledge to results: Building an analytic capability. *California Management Review*, 43(2):117–138.
- Davenport, T. H., Leibold, M., and Voelpel, S. C. (2005). *Strategic Management in the Innovation Economy*, chapter 1. Publicis Corporate Publishing.
- De Long, D. W. and Fahey, L. (2000). Diagnosing cultural barriers to knowledge management. *Academy of Management Executive*, 14(4):113–127.
- Debowski, S., Wood, R. E., and Bandara, A. (2001). Impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic search. *Journal of Applied Psychology*, 86(6):1129–1141.
- Deming, W. E. (1985). Transformation of western style of management. *Interfaces*, 15(3):6–11.
- DeMocker, J. (1998). 'knowledge management' packages focus on search tools. *Internet World*, 4(8):16–16.
- D'Eredita, M. A. and Barreto, C. (2006a). How does tacit knowledge proliferate? an episode-based perspective. *Organization Studies* (01708406), 27(12):1821–1841.
- D'Eredita, M. A. and Barreto, C. (2006b). How does tacit knowledge proliferate? an episode-based perspective. *Organization Studies*, 27(12):1821–1841.
- Devor, R. E., Chang, T.-H., and Sutherland, J. W. (1992). *Statistical Quality Design and Control: Contemporary Concepts and Methods*. Prentice Hall, 1st edition.
- Dewhurst, S. A., Holmes, S. J., Brandt, K. R., and Dean, G. M. (2006). Measuring the speed of the conscious components of recognition memory: Remembering is faster than knowing. *Consciousness and Cognition*, 15(1):147–162.
- Dienes, Z. and Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, 22(5):735–808.

- Dodgson, M., Gann, D. M., and Salter, A. (2007). The impact of modelling and simulation technology on engineering problem solving. *Technology Analysis & Strategic Management*, 19(4):471–489.
- Dörner, D. (2005). Verstehen verstehen. *Zeitschrift für Psychologie / Journal of Psychology*, 213(4):187–192.
- Dörner, D., Schaub, H., and Strohschneider, S. (1999). Komplexes Problemlösen – Königsweg der Theoretischen Psychologie? *Psychologische Rundschau*, 50(4):198–205.
- EFQM (2001). *Assessor Training Modules*. European Foundation for Quality Management, Bussels Representative Office, Avenue des Pleiade 15, B-1200 Brussels.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407 – 499.
- Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.
- Ehrlich, C. (2006). The EFQM-model and work motivation. *Total Quality Management & Business Excellence*, 17(2):131–140.
- Elsbach, K. D., Barr, P. S., and Hargadon, A. B. (2005). Identifying situated cognition in organizations. *Organization Science*, 16(4):422–433.
- Ericsson, K. and Lehmann, A. (1996). Expert and exceptional performance: Evidence and maximal adaptation to task constraints. *Annual Review of Psychology*, 47(1):273–.
- Ericsson, K. A. (2005). Recent advances in expertise research: a commentary on the contributions to the special issue. *Applied Cognitive Psychology*, 19(2):233–241.
- Ericsson, K. A., Prietula, M. J., and Cokely, E. T. (2007). The making of an expert. *Harvard Business Review*, 85(7/8):114–121.
- Ertl, B., Kopp, B., and Mandl, H. (2008). Supporting learning using external representations. *Computers & Education*, 51(4):1599–1608.
- Esser, H. (2002). *Soziologie. Spezielle Grundlagen 1: Situationslogik und Handeln*, volume 1. Campus Verlag.
- European Foundation for Quality Management (2001). Assessor training modules. Technical report, EFQM.

- European Foundation for Quality Management (2003). Introducing excellence. Technical report, EFQM.
- Evans, J. R. and Jack, E. P. (2003). Validating key results linkages in the baldrige performance excellence model. *Quality Management Journal*, 10(2):7–25.
- Fahey, L. and Prusak, L. (1998). The eleven deadliest sins of knowledge management. *California Management Review*, 40(3):265–276.
- Faraway, J. J. (1992). On the cost of data analysis. *Journal of Computational and Graphical Statistics*, 1(3):213–229.
- Faraway, J. J. (2002). Practical regression and anova using r. Technical report, University of Michigan.
- Fischer, S., Drosopoulos, S., Tsen, J., and Born, J. (2006). Implicit learning–explicit knowing: A role for sleep in memory system interaction. *Journal of Cognitive Neuroscience*, 18(3):311–319.
- Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *Proc. 15th International Conf. on Machine Learning*, pages 144–151. Morgan Kaufmann, San Francisco, CA.
- Frone, M. R. and Russell, M. (1995). Job stressors, job involvement and employee health: A test of identity theory. *Journal of Occupational & Organizational Psychology*, 68(1):1–11.
- Gentleman, R. C., Carey, V. J., Bates, D. M., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Lawrence Erlbaum Associates, Hillsdale, NJ. (Original work published 1979).
- Gioia, D. A. and Chittipeddi, K. (1991). Sensemaking and sensegiving in strategic change initiation. *Strategic Management Journal*, 12(6):433–448.
- Gittelman, M. and Kogut, B. (2003). Does good science lead to valuable knowledge? biotechnology firms and the evolutionary logic of citation patterns. *Management Science*, 49(4):366–382.
- Glisby, M. and Holden, N. (2003). Contextual constraints in knowledge management theory: The cultural embeddedness of nonaka’s knowledge-creating company. *Knowledge and Process Management*, 10(1):29–36.

- Goldberg, J. S. and Cole, B. R. (2002). Quality management in education: Building excellence and equity in student performance. *Quality Management Journal*, 9(4):8–22. ASQ.
- Goldratt, E. M. (1997). *Critical Chain*. Gower Publishing Ltd. ISBN-13: 978-0884271536.
- Grünwald, P. D. (2007). *The Minimum Description Length Principle*, chapter 17 MDL in Context, pages 523–595. MIT Press.
- Gupta, R., Duff, M. C., Denburg, N. L., Cohen, N. J., Bechara, A., and Tranel, D. (2009). Declarative memory is critical for sustained advantageous complex decision-making. *Neuropsychologia*, 47(7):1686–1693.
- Guyon, I. (2007). Introduction to feature selection. by internet video lecture.
- Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., editors (2006). *Feature Extraction, Foundations and Applications*. Physica-Verlag, Springer.
- Haas, M. and Hansen, M. (2005). When using knowledge can hurt performance: The value of organizational capabilities in a management consulting company. *Strategic Management Journal*, 26(1):1–24.
- Habermas, J. (1989). *The Theory of Communicative Action, Volume 2: Lifeworld and System: A Critique of Functionalist Reason*, chapter 6. Beacon Press, Boston. Translated by Thomas McCarthy.
- Hacker, W. and Wetzstein, A. (2004). Verbalisierende reflexion und lösungsgüte beim entwurfsdenken. *Zeitschrift für Psychologie / Journal of Psychology*, 212(3):152–166.
- Hansen, M. T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, 44(1):82 – 112.
- Hansen, M. T. and Haas, M. R. (2001). Competing for attention in knowledge markets: Electronic document dissemination in a management consulting company. *Administrative Science Quarterly*, 46(1):1–28.
- Hansen, M. T., Nohria, N., and Tierney, T. (1999). What’s your strategy for managing knowledge? *Harvard Business Review*, 77(2):106–116.
- Hargadon, A. and Fanelli, A. (2002). Action and possibility: Reconciling dual perspectives of knowledge in organizations. *Organization Science*, 13(3):290–302.
- Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, 35(1):1–97.

- Heinrich, S. and Kohlenberg, D. (2008). Performance excellence bei der hannover rückversicherungs ag. *ZfCM - Zeitschrift für Controlling & Management*, Sonderheft 3:51–59.
- Herroelen, W., Leus, R., and Demeulemeester, E. (2002). Critical chain project scheduling: Do not oversimplify. *Project Management Journal*, 33(4):48–61.
- Hintzman, D. L. and Caulton, D. A. (1997). Recognition memory and modality judgments: A comparison of retrieval dynamics. *Journal of Memory & Language*, 37(1):1–.
- Hitchcock, C. (Fall 2007). Probabilistic causation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Stanford University.
- Hofer, B. K. (2001). Personal epistemology research: Implications for learning and teaching. *Educational Psychology Review*, 13(4):353–383.
- Hofer-Alfeis, J. (2000). Icap knowledge management ansatz. Technical report, Siemens.
- Hofer-Alfeis, J. (2003). Effective integration of knowledge management into the business starts with a top-down knowledge strategy. *Journal of Universal Computer Science*, 9(7):719–728.
- Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.
- Hothorn, T., Hornik, K., and Zeileis, A. (2008). *party: A Laboratory for Recursive Part(y)itioning*. Uni Erlangen-Nürnberg, WU Wien. Documentation of the party package from the R-Project.
- Hussi, T. (2004). Reconfiguring knowledge management combining intellectual capital, intangible assets and knowledge creation. *Journal of Knowledge Management*, 8(2):36–52.
- Hüther, G. (2006). *Bedienungsanleitung für ein menschliches Gehirn*. Vandenhoeck & Ruprecht, 1st edition.
- Jacobson, A. and Prusak, L. (2006). The cost of knowledge. *Harvard Business Review*, 84(11):34–34.
- Jaeggi, S. M., Buschkuhl, M., Etienne, A., Ozdoba, C., Perrig, W. J., and Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective & Behavioral Neuroscience*, 7(2):75–89.
- Ji, L.-J., Schwarz, N., and Nisbett, R. E. (2000). Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin*, 26(5):585–593.

- John, O. P., Donahue, E. M., and Kentle, R. L. (1991). The big five inventory-versions 4a and 54. Technical report, University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA, USA.
- Judge, T. A. and Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology*, 87(4):797–807.
- Kane, A. A., Argote, L., and Levine, J. M. (2005). Knowledge transfer between groups via personnel rotation: Effects of social identity and knowledge quality. *Organizational Behavior & Human Decision Processes*, 96(1):56–71.
- Kapetanios, G. (2007). Variable selection in regression models using nonstandard optimisation of information criteria. *Computational Statistics & Data Analysis*, 52(1):4–15.
- Kaplan, R. S. and Norton, D. P. (2007). Using the balanced scorecard as a strategic management system. *Harvard Business Review*, 85(7/8):150–161.
- Kaufman, D. (2002). Turning search into knowledge management. *Electronic Library*, 20(1):49–54.
- Kearney, E., Gebert, D., and Voelpel, S. C. (2009). When and how diversity benefits teams: The importance of team members’ need for cognition. *Academy of Management Journal*, 52(3):581–598.
- Kobayashi, M. and Sakata, S. (1990). Mallows’ Cp criterion and unbiasedness of model selection. *Journal of Econometrics*, 45(3):385–395.
- Kuhn, G. and Dienes, Z. (2005). Implicit learning of nonlocal musical rules: Implicitly learning more than chunks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1417–1432.
- Kujala, J. and Lillrank, P. (2004). Total quality management as a cultural phenomenon. *Quality Management Journal*, 11(4):43–55.
- Kulik, C. T., Oldham, G. R., and Langner, P. H. (1988). Measurement of job characteristics: Comparison of the original and the revised job diagnostic survey. *Journal of Applied Psychology*, 73(3):462–466.
- Kwan, M. and Balasubramanian, P. (2003). Process-oriented knowledge management: a case study. *Journal of the Operational Research Society*, 54(2):204–212.
- Lam, W. and Chua, A. (2005). Knowledge management project abandonment: An exploratory examination of root causes. *Communications of AIS*, 2005(16):723–743.

- Lamotte, G. and Carter, G. (2000). Are the balanced scorecard and the EFQM excellence model mutually exclusive or do they work together to bring added value to a company? Technical report, PACEPerformance.
- Lane, F. C. (2001). *Ships for Victory*. John Hopkins University Press.
- Lee, P. M. (1997). *Bayesian Statistics*. Arnold, London / John Wiley & Sons Inc., New York, 2nd edition. ISBN 0471 19481 6.
- Leibold, M., Voelpel, S. C., and Tekie, E. B. (2004). Managerial levers in cultivating new mental space for business innovation. *South African Journal of Business Management*, 35(4):61–71.
- Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In Härdle, W. and Rönz, B., editors, *Compstat 2002 – Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.
- Li, F., Runger, G. C., and Tuv, E. (2006). Supervised learning for change-point detection. *International Journal of Production Research*, 44(14):2853–2868.
- Lian, Y.-H. and Van Landeghem, H. (2007). Analysing the effects of lean manufacturing using a value stream mapping-based simulation generator. *International Journal of Production Research*, 45(13):3037–3058.
- Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.
- Liker, J. (2004). *The Toyota Way: Fourteen Management Principles from the World's Greatest Manufacturer*. McGraw-Hill Professional. ISBN-13: 978-0071392310.
- Loftus, E. F. (2003). Make-believe memories. *American Psychologist*, 58(11):867–873.
- Luenberger, D. G. (1998). *Investment Science*. Oxford University Press, New York. ISBN 0-19-510809-4.
- Lukacs, P. M., Thompson, W. L., Kendall, W. L., Gould, W. R., Doherty, P. F., Burnham, K. P., and Anderson, D. R. (2007). Concerns regarding a call for pluralism of information theory and hypothesis testing. *Journal of Applied Ecology*, 44(2):456 – 460.
- Lutz, R. W. and Buhlmann, P. (2006). Conjugate direction boosting. *Journal of Computational & Graphical Statistics*, 15(2):287–311.
- Malik, F. (2008). *Strategie des Managements komplexer Systeme*, volume 10. Haupt Verlag Bern. ISBN 978-3-258-07396-5.

- Mallows, C. (1973). Some comments on Cp. *Technometrics*, 15(4):661–676.
- Manier, D., Apetroaia, I., Pappas, Z., and Hirst, W. (2004). Implicit contributions of context to recognition. *Consciousness and Cognition*, 13(3):471–483.
- Mann, S. (2001). Wearable computing: Toward humanistic intelligence. *IEEE Intelligent Systems*, 16(3):10–15.
- Mann, S. (2005). Sousveillance and cyborglogs: A 30-year empirical voyage through ethical, legal, and policy issues. *Presence: Teleoperators & Virtual Environments*, 14(6):625–646.
- Mann, S. and Barfield, W. (2003). Introduction to mediated reality. *International Journal of Human-Computer Interaction*, 15(2):205–208.
- Mann, S. and Fung, J. (2002). Eyetap devices for augmented, deliberately diminished, or otherwise altered visual perception of rigid planar patches of real-world scenes. *Presence: Teleoperators & Virtual Environments*, 11(2):158–175.
- Mathieu, J. E., Goodwin, G. F., Heffner, T. S., Salas, E., and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85(2):273–283.
- Matson, E. and Prusak, L. (2003). The performance variability dilemma. *MIT Sloan Management Review*, 45(1):39–44.
- Maurer, T. J., Weiss, E. M., and Barbeite, F. G. (2003). A model of involvement in work-related learning and development activity: The effects of individual, situational, motivational, and age variables. *Journal of Applied Psychology*, 88(4):707–724.
- McBride, D. M. and Doshier, B. A. (1997). A comparison of forgetting in an implicit and explicit memory task. *Journal of Experimental Psychology: General*, 126(4):371–392.
- McCann, L. and Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics & Data Analysis*, 52(1):249–257.
- Mebane, W. R. and Sekhon, J. S. (2007). Genetic optimization using derivatives: The rgenoud package for r. downloaded from the web on 23.04.2008.
- Mertz, E. (2007). *The Language of Law School: Learning to "Think Like a Lawyer"*. Oxford University Press. ISBN-13: 978-0195182866.
- Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 147(3):389–425.

- Mischel, W. (1977). *Personality at the crossroads: Current issues in interactional psychology*, chapter The interaction of person and situation., pages 333–352. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.
- Nonaka, I. (1991). The knowledge-creating company. *Harvard Business Review*, 69(6):96–104.
- Nonaka, I. and Takeuchi, H. (1995). *The Knowledge-Creating Company. How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, New York.
- Nonaka, I., Toyama, R., and Konno, N. (2000). SECI, ba and leadership: a unified model of dynamic knowledge creation. *Long Range Planning*, 33(1):5–34.
- Nonaka, I., von Krogh, G., and Voelpel, S. (2006). Organizational knowledge creation theory: Evolutionary paths and future advances. *Organization Studies*, 27(8):1179–1208.
- North, K. (2002). *Wissensorientierte Unternehmensführung*. Gabler, 3rd edition.
- O'Donnell, D. and Henriksen, L. B. (2002). Philosophical foundations for a critical evaluation of the social impact of ict. *Journal of Information Technology (Routledge, Ltd.)*, 17(2):89–99.
- O'Donnell, D., Porter, G., McGuire, D., Garavan, T. N., Heffernan, M., and Cleary, P. (2003). Creating intellectual capital: a habermasian community of practice (cop) introduction. *Journal of European Industrial Training*, 27(2-4):80–87.
- Okhuysen, G. A. and Eisenhardt, K. M. (2002). Integrating knowledge in groups: How formal interventions enable flexibility. *Organization Science: A Journal of the Institute of Management Sciences*, 13(4):370–386.
- Opgen-Rhein, R. and Strimmer, K. (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8(Suppl 2):S3.
- Orlikowski, W. J. (2002). Knowing in practice: Enacting a collective capability in distributed organizing. *Organization Science: A Journal of the Institute of Management Sciences*, 13(3):249–264.
- Orlikowski, W. J. and Hofman, D. J. (1997). An improvisational model for change management - the case of groupware technologies. *Sloan Management Review*, 38(2):11–21.
- Orr, J. E. (1996). *Talking About Machines: An Ethnography of a Modern Job*. Cornell University Press. ISBN-13: 978-0801483905.

- Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284.
- Pearl, J. (2003). Statistics and causal inference: A review. *TEST*, 12(2):281–345.
- Piaget, J. (2003). Part i: Cognitive development in children: Piaget: Development and learning. *Journal of Research in Science Teaching*, 40:S8–s18.
- Platek, S. M. and Kemp, S. M. (2009). Is family special to the brain? An event-related fMRI study of familiar, familial, and self-face recognition. *Neuropsychologia*, 47(3):849–858.
- Polanyi, M. (1966). *The Tacit Dimension*. Routledge & Kegan Paul, London.
- Polanyi, M. and Prosch, H., editors (1975). *Personal Knowledge*. University of Chicago Press.
- Porac, J. and Shapira, Z. (2001). On Mind, Environment, and Simon’s Scissors of Rational Behavior. *Journal of Management and Governance*, 5(3-4):206–211.
- Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, 133(2):227–244.
- Project Management Institute (1996). *A Guide to the Project Management Body of Knowledge: PMBOK Guide*. PMI Publishing Division, 3rd edition. ISBN: 1-880410-12-5.
- Prusak, L. (2005). The madness of individuals. *Harvard Business Review*, 83(6):22–22.
- R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Racsmány, M. and Conway, M. A. (2006). Episodic inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(1):44–57.
- Rammstedt, B. and John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.
- Rand, G. K. (2000). Critical chain: the theory of constraints applied to project management. *International Journal of Project Management*, 18(3):173–177.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press. ISBN 0-262-18253-X.

- Rasmussen, J. (2001). The importance of communication in teaching: a systems-theory approach to the scaffolding metaphor. *Journal of Curriculum Studies*, 33(5):569–582.
- Raz, T. (2003). A critical look at critical chain project management. *Project Management Journal*, 34(4):24–32.
- Reagans, R., Argote, L., and Brooks, D. (2005). Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. *Management Science*, 51(6):869–881.
- Reimann, R. and Dörner, D. (2004). Die Auswirkung von selbstadressierten Fragen auf die Entwurfsqualität beim Konstruieren. *Zeitschrift für Psychologie / Journal of Psychology*, 212(1):1–9.
- Rivkin, J. W. (2000). Imitation of complex strategies. *Management Science*, 46(6):824–.
- Robert, C. P. (2005). *Monte Carlo Statistical Methods*. Springer, Berlin, 2nd edition. ISBN-13: 978-0387212395.
- Rolf, B. (2004). Two theories of tacit and implicit knowledge. In *Autumn Meeting of the SIG Philosophy and Informatics*.
- Roßnagel, C. S. (2001). Revealing hidden covariation detection: Evidence for implicit abstraction at study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Vol 27(5) Sep 2001, ., 25(5):1276–1288.
- Roßnagel, C. S. (2008). *Mythos: “alter” Mitarbeiter - Lernkompetenz jenseits der 40?!* Beltz Verlag, Weinheim, Basel, 1st edition.
- Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, 29(3):343–367.
- Rueda, M., González, S., and Arcos, A. (2005). Indirect methods of imputation of missing data based on available units. *Applied Mathematics & Computation*, 164(1):249–261.
- Rusjan, B. (2005). Usefulness of the EFQM excellence model: Theoretical explanation of some conceptual and methodological issues. *Total Quality Management & Business Excellence*, 16(3):363–380.
- Salter, A. and Gann, D. (2003). Sources of ideas for innovation in engineering design. *Research Policy*, 32(8):1309–1325.
- Samra-Fredericks, D. (2000). Doing ‘boards-in-action’ research - an ethnographic approach for the capture and analysis of directors’ and senior managers’ interactive routines. *Corporate Governance - An International Review*, 8(3):244–257.

- Sandow, D. and Allen, A. M. (May 2005). The nature of social collaboration: How work really gets done. *Reflections: The SoL Journal*, 6:1–14.
- Sarkar, D. (2008). *lattice: Lattice Graphics*. R package version 0.17-4.
- Schmidt, D. W. (2008). Balanced scorecard & EFQM-modell. *ZfCM Controlling & Management*, Sonderheft 3:2–11.
- Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology*, 82(3):498– 504.
- Schommer, M., Calvert, C., Gariglietti, G., and Bajaj, A. (1997). The development of epistemological beliefs among secondary students: A longitudinal study. *Journal of Educational Psychology*, 89(1):37–41.
- Schreyögg, G. and Geiger, D. (2005). Zur Konvertierbarkeit von Wissen - Wege und Irrwege im Wissensmanagement. *Zeitschrift für Betriebswirtschaft (ZfB)*, ZfB 75(H. 5):433–454.
- Schreyögg, G. and Geiger, D. (2007). The significance of distinctiveness: A proposal for rethinking organizational knowledge. *Organization*, 14(1):77–100.
- Schwarz, N. and Bienias, J. (1990). What mediates the impact of response alternatives on frequency reports of mundane behaviors? *Applied Cognitive Psychology*, 4(1):61–72.
- Schwarzer, R. (2000). *Streß, Angst und Handlungsregulation*. Kohlhammer, Stuttgart, 4rd edition.
- Schwarzer, R. and Jerusalem, M. (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen*. Schwarzer, Ralf and Jerusalem, Matthias.
- Schön, D. A. (1992). II. Philosophical Perspectives and Practice: The Crisis of Professional Knowledge and the Pursuit of an Epistemology of Practice. In *Praxiologies & the Philosophy of Economics - Praxiology*, pages 163–185. Transaction Publishers.
- Scott, R. B. and Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology / Learning, Memory & Cognition*, 34(5):1264–1288.
- Senge, P. M. (2004). Learn to innovate. *Executive Excellence*, 21(6):3–4.
- Sengupta, K., Abdel-Hamid, T. K., and Van Wassenhove, L. N. (2008). The experience trap. *Harvard Business Review*, 86(2):94–101.
- Senturia, S. D. and Wedlock, B. D. (1993). *Electronic Circuits and Applications*. Krieger Publishing Company.

- Serrano, I., Ochoa, C., and Castro, R. D. (2008). Evaluation of value stream mapping in manufacturing system redesign. *International Journal of Production Research*, 46(16):4409–4430.
- Sevcikova, H. and Rossini, A. J. (2005). *snowFT: Fault Tolerant Simple Network of Workstations*. R package version 0.0-2.
- Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, 17(3):1176–1197.
- Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product / 50th Anniversary Commemorative Issue*. Amer Society for Quality. ISBN-13: 978-0873890762.
- Siegler, R. S. (2000). Unconscious insights. *Current Directions in Psychological Science*, 9(3):79–83.
- Siegler, R. S. (2005). Robert S. Siegler: Award for Distinguished Scientific Contributions. *American Psychologist*, 60(8):767–778.
- Siegler, R. S. and Chen, Z. (2008). Differentiation and integration: guiding principles for analyzing cognitive change. *Developmental Science*, 11(4):433–448.
- Siegler, R. S. and Svetina, M. (2006). What leads children to adopt new strategies? a microgenetic/cross-sectional study of class inclusion. *Child Development*, 77(4):997–1015.
- Sims, A. C., Bowles, J., Crosby, P. B., Gale, B. T., Hammond, J., Reimann, C. W., Deming, W. E., Crawford-Mason, C., Clausing, D., Galvin, R. W., Hockman, K. K., Pifer, P., Cooper, G. E., Leach, K. E., McKeown, K., Peck, D., Shiba, S., Peterson, D. E., Irwin, B. M., and Rickard, N. E. (1992). Does the baldrige award really work?
- Sobel, M. E. (2005). Discussion: ‘the scientific model of causality’. *Sociological Methodology*, 35(1):99–133.
- Soule, M. (1987). *Viable Populations for Conservation.*, chapter Where do we go from here?, page 175–183. Cambridge University Press, Cambridge, UK.
- Spender, J.-C. (1996). Making knowledge the basis of a dynamic theory of the firm. *Strategic Management Journal*, 17:45–62.
- Starbuck, W. H. (2004). Why I stopped trying to understand the real world. *Organization Studies (Sage Publications Inc.)*, 25(7):1233–1254.
- Steers, R. M., Mowday, R. T., and Shapiro, D. L. (2004). Introduction to special topic forum: The future of work motivation theory. *The Academy of Management Review*, 29(3):379–387.

- Sternberg, R. J. (1997). Managerial Intelligence: Why IQ Isn't Enough. *Journal of Management*, 23(3):475–494.
- Sternberg, R. J. (2008). Increasing fluid intelligence is possible after all. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6791–6792.
- Sternberg, R. J. and Hedlund, J. (2002). Practical intelligence, g, and work psychology. *Human Performance*, 15(1/2):143–160.
- Stewart, T. R. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*, chapter Improving reliability of judgmental forecasts., pages 81–106. Norwell, MA: Kluwer Academic Publishers.
- Stewart, T. R. and Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting*, 13(7):579–599.
- Steyn, H. (2001). An investigation into the fundamentals of critical chain project scheduling. *International Journal of Project Management*, 19(6):363–370.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25.
- Sun, R., Slusarz, P., and Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112(1):159–192.
- Sutherland, I. E. (1964). Sketch pad a man-machine graphical communication system. In *DAC '64: Proceedings of the SHARE design automation workshop*, pages 6.329–6.346, New York, NY, USA. ACM.
- Svenmarck, P. and Dekker, S. (2003). Decision support in fighter aircraft: from expert systems to cognitive modelling. *Behaviour & Information Technology*, 22(3):175–185.
- Szulanski, G., Cappetta, R., and Jensen, R. J. (2004). When and how trustworthiness matters: Knowledge transfer and the moderating effect of causal ambiguity. *Organization Science*, 15(5):600–613.
- Teece, D. J. (2000). Strategies for managing knowledge assets: the role of firm structure and industrial context. *Long Range Planning*, 33(1):35–54.

- Thomas, D. R., Zhu, P., and Decady, Y. J. (2007). Point estimates and confidence intervals for variable importance in multiple linear regression. *Journal of Educational and Behavioral Statistics*, 32(1):61–91.
- Thomas, W. I. and Znaniecki, F. (1927). *The Polish Peasant in Europe and America*, volume 1. Knopf, New York.
- Thuiller, W., Midgley, G. F., Rougeti, M., and Cowling, R. M. (2006). Predicting patterns of plant species richness in megadiverse south africa. *Ecography*, 29(5):733–744.
- Tsoukas, H. (2005a). Afterword: why language matters in the analysis of organizational change. *Journal of Organizational Change Management*, 18(1):96–104.
- Tsoukas, H. (2005b). *Complex Knowledge: Studies in Organizational Epistemology*. Oxford University Press.
- Tsoukas, H. and Chia, R. (2002). On organizational becoming: Rethinking organizational change. *Organization Science*, 13(5):567–582.
- Tsoukas, H. and Vladimirou, E. (2001). What is organizational knowledge? *Journal of Management Studies*, 38(7):973–994.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company. ISBN 0-201-07616-0.
- Uhl-Bien, M., Marion, R., and McKelvey, B. (2007). Complexity leadership theory: Shifting leadership from the industrial age to the knowledge era. *The Leadership Quarterly*, 18(4):298–318.
- van Buuren, S. (2008). Mice imputation package for r. <http://web.inter.nl.net/users/S.van.Buuren/mi/html/mice.htm>.
- van de Ven, A., Angle, H., and Poole, M. (2000). *Research on the Management of Innovation: The Minnesota Studies*. Oxford University Press.
- van der Laan, M. (2006). Statistical inference for variable importance. *International Journal of Biostatistics*, 2:1008–1008.
- Van Eerde, W. and Thierry, H. (1996). Vroom’s expectancy models and work-related criteria: A meta-analysis. *Journal of Applied Psychology*, 81(5):575–586.
- Vera, D. and Crossan, M. (2004). Strategic leadership and organizational learning. *Academy of Management Review*, 29(2):222–240.
- Vicente, K. J. (2003). Beyond the lens model and direct perception: Toward a broader ecological psychology. *Ecological Psychology*, 15(3):241–267.

- Voelpel, S. and Meyer, J. (2006). Haridimos tsoukas: Complex knowledge: Studies in organizational epistemology. *Organization Studies* (01708406), 27(10):1562–1568.
- Voelpel, S. C., Dous, M., and Davenport, T. H. (2005). Five steps to creating a global knowledge-sharing system: Siemens' sharenet. *Academy of Management Executive*, 19(2):9–23.
- von Krogh, G. and Grand, S. (2000). *Knowledge Creation: A Source of Value*, chapter Knowledge Creation Theories explained and justified, pages Part 1, Chpt. 1. Palgrave Macmillan.
- von Krogh, G. and Venzin, M. (1995). Anhaltende wettbewerbsvorteile durch wissensmanagement. *Die Unternehmung*, 6:417 – 436.
- Vroom, V. H. (1964). *Workand motivation*. Wiley, New York.
- Walgenbach, P. and Beck, N. (2000). Von statistischer qualitätskontrolle über qualitätssicherungssysteme hin zum total quality management – die institutionalisierung eines neuen managementkonzepts. *Soziale Welt*, 3:325–354.
- Walsham, G. (2001). Knowledge management: The benefits and limitations of computer systems. *European Management Journal*, 19(6):599–608.
- Ward, V. (1998). Mapping meta knowledge: A cartographic approach to finding "knowledge" about knowledge. *Knowledge Management Review*, 1(5):10 – 16.
- Weick, K. E. (1993). The collapse of sensemaking in organizations: The mann gulch disaster. *Administrative Science Quarterly*, 38(4):628–652.
- Weick, K. E., Sutcliffe, K. M., and Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, 16(4):409–421.
- Weiss, L. M., Capozzi, M. M., and Prusak, L. (2004). Learning from the internet giants. *MIT Sloan Management Review*, 45(4):79–84.
- Wengraf, T. (2001). *Qualitative Research Interviewing*. Sage Publications.
- West, S. G. (2006). Seeing your data: Using modern statistical graphics to display and detect relationships. In Bootzin, R. R. E. and McKnight, P. E. E., editors, *Strengthening research methodology: Psychological measurement and evaluation.*, pages 159–182. American Psychological Association.
- Wexler, M. N. (2001). The who, what and why of knowledge mapping. *Journal of Knowledge Management*, 5(3):249–263.

- Wilkinson, G. and Dale, B. (1999). Models of management system standards: a review of the integration issues. *International Journal of Management Reviews*, 1(3):279–298.
- Winne, P. H. (1995). Self-regulation is ubiquitous but its forms vary with knowledge. *Educational Psychologist*, 30(4):223–229.
- Winship, C. and Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25(1):659–507.
- Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann. The official reference to the WEKA machine learning software environment.
- Wong, W. L. P. and Radcliffe, D. F. (2000). The tacit nature of design knowledge. *Technology Analysis & Strategic Management*, 12(4):493–512.
- Woodward-Kron, R. (2008). More than just jargon – the nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes*, 7(4):234–249.
- Wysocki, R. K. (2006). *Effective Project Management: Traditional, Adaptive, Extreme*, volume 4. Wiley & Sons. ISBN 978-0470042618.
- Yang, J.-B. (2007). How the critical chain scheduling method is working for construction. *Cost Engineering*, 49(4):25–32.
- Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472):1202–1214.
- Zhang, P. (1992a). Inference after variable selection in linear regression models. *Biometrika*, 79(4):741–746.
- Zhang, P. (1992b). On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, 87(419):732–737.
- Zhang, P. (1993). Model selection via multifold cross validation. *Annals of Statistics*, 21(1):299–313.
- Zusho, A., Pintrich, P. R., and Goppola, B. (2003). Skill and will: the role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education*, 25(9):1081–1095.

Index

- PIA-model
 - principle variant, 31
 - with visualization, 36
- Actors approach, 90
- augmented reality, 37
- background knowledge
 - of the lifeworld, 42
- balanced scorecard, 64
- bias
 - in perception, 40, 66
- BOGER, 179
- bootstrapping, 132, 189
- casual ambiguity, 73
- change management, 68, 79
- classifier algorithms, 108
- classifier models, 108
- collinearity, 117
- correlation analysis, 165
- cross validation, 131
- crystalline intelligence, 41
- data, 77
- definition
 - of on-the-job learning, 79
- expected value
 - general mathematical definition, 121
 - mean estimator, 122
- expected value models, 108
- experience, 39, 42
- externalization, 70
- feature selection, 35
- feedback
 - in learning, 47
- filtering, 31
- fluid intelligence, 41
- Humanistic Intelligence, 37
- imputation, 304
- information, 77
- instability
 - of a parameter, 190
- integrating information, 34, 67
- intrinsic motivation, 239
- iterative research approach, 94
- Kaizen, 62
- knowledge, 77
- knowledge justification, 38
- KPI - key performance indicator, 61, 64
- learning
 - implicit, 53
 - explicit, 53
- learning index, 150
- lens model, 33
- logistic regression, 108
- Matsushita bread baking example, 71

- mental models, 41
- metaphor, 44
- missing values, 303
- mode shape graphics, 215
- model error, 115
- model fit, 114
- model fit
 - Breiman's predictive error estimate, 195
 - coefficient of determinance R^2 , 120
 - Breiman's predictive error estimate, 133
 - test data model fit, 196
 - training data model fit, 196
- model fit
 - absolute sum coefficient R_{abs} , 120
- Muda, 61
- multi-collinearity, 117
- Mura, 61
- Muri, 61
- noise
 - in the data, 166
- non-factors, 260
- null-hypothesis testing, 208
- organizational factors
 - in the survey, 141
- outlier removal, 302
- overfitting, 123
- perspective taking, 30
- Predictive Error, 114
- prior knowledge, 41
- project
 - status, 59
 - management, 58
- RADAR, 64
- resampling, 131, 189
- sample size, 138, 303
- sampling bias, 129
- sampling bias, 120, 132
- SECI model by Nonaka, 69
- self-efficacy
 - for search activities, 51
- Self-Regulated Learning (SRL) model, 48
- sense making, 35
- socially constructed knowledge, 43
- software
 - for decision support by visualization, 37
- spurious correlation, 110
- stability
 - of a parameter, 190
- statistical inference, 105
- statistical significance, 189
- stochastic processes, 107
- structural equations modeling
 - non-linear, 109
- subjectivity, 42
- survey-reduction levels, 154
- t-test by Student, 208
- tacit knowledge
 - as defined by Nonaka, 69
 - as defined by Polanyi, 54
- thomas theorem in sociology, 39
- toyota production system (TPS), 60
- TQM, 62
- truth, 38, 74
- value
 - of knowledge, 75
 - of organizational research, 89
 - of this study, 91
- variable importance
 - general application, 206
 - implementation details, 209

