

# **Enhancing Student Performance Through Technology-Mediated Formative Feedback in Large Scale University Lectures**

DISSERTATION  
of the University of St.Gallen,  
School of Management,  
Economics, Law, Social Sciences  
and International Affairs  
to obtain the title of  
Doctor of Philosophy in Management

submitted by

**Roman Rietsche**

from

Germany

Approved on the application of

**Prof. Dr. Jan Marco Leimeister**

and

**Prof. Dr. Jan vom Brocke**

Dissertation no. 5012

Difo-Druck GmbH, Untersiemau 2020

The University of St.Gallen, School of Management, Economics, Law, Social Sciences and International Affairs hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St.Gallen, May 18, 2020

The President:

Prof. Dr. Bernhard Ehrenzeller

To my lovely grandma.



## Acknowledgements

This dissertation is the outcome of a three-and-a-half-year venture under the supervision of Prof. Dr. Jan Marco Leimeister at the Institute of Information Management in the University of St.Gallen.

First and foremost, I thank Prof. Dr. Jan Marco Leimeister for his personal support, his supervision of my thesis and inspiring discussions leading to constant critical reflection of my work. Further, I also thank Prof. Dr. Jan vom Brocke for his willingness to co-supervise my thesis and inspiring conversations. I would like to express my deepest gratitude to Prof. Dr. Matthias Söllner for the intense and rewarding collaboration during our joint research projects as well as his relentless efforts to shape me as a researcher and getting the best out of me.

For their untiring advice and support, I would like to specially thank Susanne Gmünder, Dr. Jochen Müller, Dr. Robert Schleich, Dr. Holger Keibel and Prof. Dr. Claude Siegenthaler.

Many others have accompanied me on my way to becoming a researcher. In particular, I would like to thank Prof. Dr. Konstantin Bauman, my long-time research colleague and host at Temple University, for our intense collaboration, endless hours of discussions and joy working together. Furthermore, I would like to thank Dr. Filio Constantinou, my host at the University of Cambridge, for her endless support and all the friends I made during my one-year stay abroad who made this time unforgettable.

Much of my experience was shaped by scientific discourse and collaboration. I was more than lucky to have such inspiring colleagues. Therefore, I am especially grateful to my co-authors and my colleagues at the Institute of Information Management and the University of Kassel. My special thanks go to Marcel Rhyn, Stefan Kleinschmidt, Emanuel Stöckli, Christian Dremel and Arne Grüttner, with whom I have not only composed articles but who also challenged and influenced my work.

Finally I wish to thank my family, mother, father, brother, grandma and all my friends for encouraging and supporting me throughout this journey.

Cambridge, December 2019

Roman Rietsche



# Table of Contents

|                                                        |       |
|--------------------------------------------------------|-------|
| <b>Acknowledgements</b>                                | v     |
| <b>List of Figures</b>                                 | xiii  |
| <b>List of Tables</b>                                  | xv    |
| <b>List of Acronyms</b>                                | xix   |
| <b>Abstract</b>                                        | xxi   |
| <b>Zusammenfassung</b>                                 | xxiii |
| <b>1 Introduction</b>                                  | 1     |
| 1.1 Motivation . . . . .                               | 2     |
| 1.2 Research Questions and Research Methods . . . . .  | 5     |
| 1.3 Structure of the Dissertation . . . . .            | 12    |
| 1.4 Overview of Publications . . . . .                 | 14    |
| 1.4.1 Published Journals . . . . .                     | 14    |
| 1.4.2 Journals Under Review . . . . .                  | 15    |
| 1.4.3 Conference Articles . . . . .                    | 15    |
| <b>2 Theoretical Background and Related Work</b>       | 17    |
| 2.1 Learning in Large-Scale Classes . . . . .          | 18    |
| 2.1.1 Characteristics of Large-Scale Classes . . . . . | 18    |

## TABLE OF CONTENTS

---

|          |                                                                         |           |
|----------|-------------------------------------------------------------------------|-----------|
| 2.1.2    | Technology-Mediated Learning . . . . .                                  | 20        |
| 2.1.3    | Peer Learning . . . . .                                                 | 21        |
| 2.1.4    | Student Performance . . . . .                                           | 23        |
| 2.1.5    | Higher-Order Thinking Skills . . . . .                                  | 25        |
| 2.2      | Assessment for Student Performance . . . . .                            | 27        |
| 2.2.1    | Self-Assessment . . . . .                                               | 27        |
| 2.2.2    | Computer-based Assessment . . . . .                                     | 28        |
| 2.2.3    | Peer Evaluation . . . . .                                               | 28        |
| 2.2.4    | Peer Assessment . . . . .                                               | 29        |
| 2.2.5    | Cognitive Dissonance Theory . . . . .                                   | 29        |
| 2.3      | Feedback on Student Performance . . . . .                               | 30        |
| 2.3.1    | Characteristics of Feedback . . . . .                                   | 31        |
| 2.3.2    | Formative Feedback . . . . .                                            | 32        |
| 2.3.3    | Peer Feedback . . . . .                                                 | 34        |
| 2.3.4    | Hofstede's Cultural Dimension Theory . . . . .                          | 36        |
| 2.3.5    | Ensuring Peer Feedback Quality . . . . .                                | 37        |
| <b>3</b> | <b>System Design and Evaluation</b>                                     | <b>41</b> |
| 3.1      | Introduction . . . . .                                                  | 42        |
| 3.2      | Research Methodology . . . . .                                          | 43        |
| 3.3      | Designing and Evaluating Formative Feedback Tools . . . . .             | 45        |
| 3.3.1    | Step 2: Deriving Meta-Requirements from Scientific Literature . . . . . | 47        |
| 3.3.2    | Step 3: Deriving Requirements from Expert Interviews . . . . .          | 47        |



|          |                                                                                 |           |
|----------|---------------------------------------------------------------------------------|-----------|
| 3.3.3    | Step 4: Deriving Design Principles and Features for the Alpha Version . . . . . | 48        |
| 3.3.4    | Step 5: Evaluation of Alpha Version . . . . .                                   | 52        |
| 3.3.5    | Step 6: Deriving Design Features for the Beta Version . . .                     | 53        |
| 3.4      | Evaluation of Beta Version . . . . .                                            | 54        |
| 3.4.1    | Data Collection . . . . .                                                       | 55        |
| 3.4.2    | Data Preparation . . . . .                                                      | 56        |
| 3.4.3    | Measurement of Constructs . . . . .                                             | 57        |
| 3.4.4    | Data Analysis . . . . .                                                         | 59        |
| 3.5      | Discussion and Contributions . . . . .                                          | 61        |
| 3.5.1    | Limitations and Outlook . . . . .                                               | 65        |
| 3.6      | Conclusion of this Chapter . . . . .                                            | 65        |
| <b>4</b> | <b>Helpfulness of Online Reviews</b>                                            | <b>67</b> |
| 4.1      | Introduction . . . . .                                                          | 68        |
| 4.2      | Research Methodology . . . . .                                                  | 69        |
| 4.3      | Findings . . . . .                                                              | 72        |
| 4.3.1    | Criteria Determining the Helpfulness of Online Reviews .                        | 73        |
| 4.3.2    | Review-Related Factors . . . . .                                                | 77        |
| 4.3.3    | Reviewer-Related Factors . . . . .                                              | 84        |
| 4.3.4    | Reader-Related factors . . . . .                                                | 86        |
| 4.3.5    | Environment-Related Factors . . . . .                                           | 87        |
| 4.4      | Discussion and Contributions . . . . .                                          | 89        |
| 4.4.1    | Limitations and Outlook . . . . .                                               | 90        |

## TABLE OF CONTENTS

---

|          |                                                           |            |
|----------|-----------------------------------------------------------|------------|
| 4.5      | Conclusion of this Chapter . . . . .                      | 91         |
| <b>5</b> | <b>Measuring Peer Feedback Quality</b>                    | <b>93</b>  |
| 5.1      | Introduction . . . . .                                    | 94         |
| 5.2      | Cross-Industry Standard Process for Data Mining . . . . . | 96         |
| 5.3      | Feature Development . . . . .                             | 97         |
| 5.3.1    | Text Features . . . . .                                   | 98         |
| 5.3.2    | Feedback Features . . . . .                               | 101        |
| 5.3.3    | Emotional Features . . . . .                              | 104        |
| 5.4      | Research Methodology . . . . .                            | 113        |
| 5.4.1    | Measurements . . . . .                                    | 113        |
| 5.4.2    | Methods . . . . .                                         | 113        |
| 5.5      | Empirical Study . . . . .                                 | 119        |
| 5.5.1    | Peer Feedback Process . . . . .                           | 119        |
| 5.6      | Analysis and Results . . . . .                            | 120        |
| 5.6.1    | Evaluation Methodology . . . . .                          | 121        |
| 5.6.2    | Results . . . . .                                         | 122        |
| 5.7      | Discussion and Contributions . . . . .                    | 128        |
| 5.7.1    | Contribution and Theoretical Implications . . . . .       | 128        |
| 5.7.2    | Practical Implications . . . . .                          | 131        |
| 5.7.3    | Limitations and Future Research . . . . .                 | 132        |
| 5.8      | Conclusion of this Chapter . . . . .                      | 133        |
| <b>6</b> | <b>Cultural Influences in Peer Feedback</b>               | <b>135</b> |
| 6.1      | Introduction . . . . .                                    | 136        |

|          |                                                                                      |            |
|----------|--------------------------------------------------------------------------------------|------------|
| 6.2      | Research Methodology . . . . .                                                       | 137        |
| 6.3      | Towards Designing and Evaluating Peer Feedback Design Principles . . . . .           | 139        |
| 6.3.1    | Problem Identification . . . . .                                                     | 140        |
| 6.3.2    | Definition of Objectives . . . . .                                                   | 140        |
| 6.3.3    | Design and Development . . . . .                                                     | 142        |
| 6.4      | Results . . . . .                                                                    | 144        |
| 6.5      | Discussion and Contributions . . . . .                                               | 144        |
| 6.5.1    | Limitations and Future Research . . . . .                                            | 149        |
| 6.6      | Conclusion of this Chapter . . . . .                                                 | 149        |
| <b>7</b> | <b>Develop the Peer Feedback Providing Skill</b>                                     | <b>151</b> |
| 7.1      | Introduction . . . . .                                                               | 152        |
| 7.2      | Research Methodology . . . . .                                                       | 154        |
| 7.2.1    | Experimental Design . . . . .                                                        | 154        |
| 7.2.2    | Measurements . . . . .                                                               | 155        |
| 7.2.3    | ITPF Process . . . . .                                                               | 156        |
| 7.3      | Results . . . . .                                                                    | 156        |
| 7.4      | Discussion and Contributions . . . . .                                               | 158        |
| 7.4.1    | Limitations and Future Research . . . . .                                            | 160        |
| 7.5      | Conclusion of this Chapter . . . . .                                                 | 162        |
| <b>8</b> | <b>Synthesis of the Findings</b>                                                     | <b>165</b> |
| 8.1      | Building IT Artifacts for Providing Technology-Mediated Formative Feedback . . . . . | 167        |

## TABLE OF CONTENTS

---

|           |                                                                                                                            |            |
|-----------|----------------------------------------------------------------------------------------------------------------------------|------------|
| 8.2       | Effects of Technology-Mediated Formative Feedback on Student Performance . . . . .                                         | 170        |
| <b>9</b>  | <b>Theoretical Contributions and Implications</b>                                                                          | <b>173</b> |
| 9.1       | Design Knowledge for Research on Formative Feedback IT Artifacts . . . . .                                                 | 175        |
| 9.2       | Effects of Technology-Mediated Formative Feedback on Student Performance . . . . .                                         | 177        |
| 9.3       | Prescriptive Knowledge of Measuring Peer Feedback Quality . . .                                                            | 179        |
| 9.4       | Effects of Technology-Mediated Formative Feedback on the Students' Perceptions of their Skill at Providing Peer Feedback . | 182        |
| <b>10</b> | <b>Practical Implications</b>                                                                                              | <b>185</b> |
| 10.1      | Designing Formative Feedback Tools . . . . .                                                                               | 186        |
| 10.2      | Improving Student Performance . . . . .                                                                                    | 187        |
| 10.3      | Providing Guidelines for Peer Feedback . . . . .                                                                           | 188        |
| <b>11</b> | <b>Future Research</b>                                                                                                     | <b>191</b> |
| 11.1      | A Design Theory of Formative Feedback IT Artifacts . . . . .                                                               | 192        |
| 11.2      | Developing the Feedback Providing Skill Holistically . . . . .                                                             | 193        |
| 11.2.1    | Feedback Writing or Feedback Receiving? . . . . .                                                                          | 193        |
| 11.2.2    | Individualized Recommendations . . . . .                                                                                   | 194        |
| 11.2.3    | How Students Process Feedbacks . . . . .                                                                                   | 195        |
| <b>12</b> | <b>Conclusion</b>                                                                                                          | <b>197</b> |
|           | <b>References</b>                                                                                                          | <b>201</b> |
|           | <b>Appendices</b>                                                                                                          | <b>233</b> |

## List of Figures

|     |                                                                                                               |    |
|-----|---------------------------------------------------------------------------------------------------------------|----|
| 1.1 | Overview of the Research Questions . . . . .                                                                  | 7  |
| 1.2 | Structure of the Dissertation . . . . .                                                                       | 13 |
| 2.1 | Higher-Order Thinking Skills in the Revision of Bloom's Taxonomy                                              | 24 |
| 2.2 | Linking Formative Feedback to the Revision of Bloom's Taxonomy                                                | 34 |
| 3.1 | Three Cycle Design Science Research Process . . . . .                                                         | 45 |
| 3.2 | Overview of the Design Principles and the Features Derived<br>from the Defined (Meta-)Requirements . . . . .  | 46 |
| 3.3 | LOOM's Student View (Multiple Devices) Including Selec-<br>ted Design Principles and Features . . . . .       | 49 |
| 3.4 | Taxonomy Table During the Process of Assigning a<br>Computer-based Assessment to a Learning Outcome . . . . . | 51 |
| 3.5 | Results Research Model A Using Dataset 1 . . . . .                                                            | 57 |
| 3.6 | Results Research Model B Using Dataset 2 . . . . .                                                            | 58 |
| 4.1 | Literature Search Process of Online Review Helpfulness . . . . .                                              | 70 |
| 4.2 | Part 1/2: Co-Topic Analysis of 81 Journal Articles . . . . .                                                  | 75 |
| 4.3 | Part 2/2: Journal Authors of Co-topic Analysis . . . . .                                                      | 76 |

## LIST OF FIGURES

---

|     |                                                                                                               |     |
|-----|---------------------------------------------------------------------------------------------------------------|-----|
| 5.1 | Cross-Industry Standard Process for Data Mining . . . . .                                                     | 97  |
| 5.2 | First Two Levels of the Proposed Feature Hierarchy $\Omega$ Presenting Categories and Subcategories . . . . . | 98  |
| 5.3 | Examples of Sentiment Curves . . . . .                                                                        | 118 |
| 5.4 | Structure of the IT-based Peer Feedback Process . . . . .                                                     | 120 |
| 7.1 | Steps of the IT-based Peer Feedback Process . . . . .                                                         | 157 |
| 7.2 | Results of the Students' Ability to Provide Feedback to Peers . . .                                           | 158 |

# List of Tables

|     |                                                                                                                           |     |
|-----|---------------------------------------------------------------------------------------------------------------------------|-----|
| 3.1 | Descriptive Statistics of Evaluating the IT artifact . . . . .                                                            | 56  |
| 3.2 | Documentation of the Design Knowledge . . . . .                                                                           | 64  |
| 5.1 | List of all Proposed Features $\mathcal{F}$ Organized into Categories<br>and Subcategories $\Omega$ . . . . .             | 107 |
| 5.2 | Number of Reviews Gathered per Year . . . . .                                                                             | 119 |
| 5.3 | Peer Assessment Items Measuring the Latent Construct of<br>Feedback Quality . . . . .                                     | 121 |
| 5.4 | OLS Regression for <i>1.2 Readability</i> Group of Features from<br><i>1. Text Features</i> Category . . . . .            | 123 |
| 5.5 | OLS Regression for the Features Related to <i>Suggestions</i> from<br><i>2. Feedback Features</i> Category . . . . .      | 124 |
| 5.6 | OLS Regression for <i>Sentiment Curve</i> Subcategory of Features<br>from <i>3. Emotional Features</i> Category . . . . . | 126 |
| 5.7 | Feature Weights in the Random Forest Regression Model . . . . .                                                           | 128 |
| 6.1 | Value Survey Model Results of the Study Participants . . . . .                                                            | 139 |
| 6.2 | Design Principles and their Inherent Cultural Dimension . . . . .                                                         | 143 |
| 6.3 | Comparison of Swiss and Chinese Preferences for the Deve-<br>loped Design Principles . . . . .                            | 144 |

## LIST OF TABLES

---

|     |                                                                                                                                          |     |
|-----|------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 7.1 | Survey Questions Self-Assessment and Peer-Assessment . . . . .                                                                           | 155 |
| 7.2 | Statistical Results of the Students' Ability to Provide Feed-<br>back to Peers . . . . .                                                 | 161 |
| A1  | OLS Regression for Subcategory "1.1 General Text Structure"<br>from 1. <i>Text Features</i> Category . . . . .                           | 235 |
| A2  | OLS Regression for Subcategory "1.2 Readability" from 1.<br><i>Text Features</i> Category . . . . .                                      | 236 |
| A3  | OLS Regression for Subcategory "1.3 Writing Style" from 1.<br><i>Text Features</i> Category . . . . .                                    | 236 |
| A4  | OLS Regression for Subcategory "2.1 Argumentation" from<br>2. <i>Feedback Features</i> Category . . . . .                                | 237 |
| A5  | OLS Regression for Subcategory "2.2 Feedback Points" from<br>2. <i>Feedback Features</i> Category . . . . .                              | 238 |
| A6  | OLS Regression for Subcategory "2.3 Flow of Feedback<br>Points" from 2. <i>Feedback Features</i> Category . . . . .                      | 239 |
| A7  | OLS Regression for Subcategory "2.4 Verification and Elabo-<br>ration" group of features from 2. <i>Feedback Features</i> Category . . . | 240 |
| A8  | OLS Regression for Subcategory "2.5 Function" group of<br>features from 2. <i>Feedback Features</i> Category . . . . .                   | 241 |
| A9  | OLS Regression for Subcategory "3.1 Expressions" from 3.<br><i>Emotional Features</i> Category . . . . .                                 | 242 |
| A10 | OLS Regression for Subcategory "3.2 Nothing to Impro-<br>ve" from 3. <i>Emotional Features</i> Category . . . . .                        | 242 |
| A11 | OLS Regression for Subcategory "3.3 Sentiment Curve" from<br>3. <i>Emotional Features</i> Category . . . . .                             | 243 |



|                                                                                                    |     |
|----------------------------------------------------------------------------------------------------|-----|
| A12 OLS Regression for Subcategory “3.3 Tone” from 3. <i>Emotional Features</i> Category . . . . . | 244 |
|----------------------------------------------------------------------------------------------------|-----|



# List of Acronyms

| <b>Notation</b> | <b>Description</b>                              | <b>Page List</b> |
|-----------------|-------------------------------------------------|------------------|
| AES             | Automated Essay Scoring                         | 37               |
| AVE             | Average Variance Extracted                      | 60               |
| CbA             | Computer-based Assessment                       | 6                |
| CDT             | Cognitive Dissonance Theory                     | 29               |
| CPR             | Calibrated Peer Reviewing                       | 39               |
| CRISP-DM        | CRoss-Industry Standard Process for Data Mining | 11               |
| DFs             | Design Features                                 | 45               |
| DPs             | Design Principles                               | 45               |
| DSR             | Design Science Research                         | 8                |
| EEG             | Electroencephalography                          | 195              |
| EF              | Emotional Features                              | 97               |
| FF              | Feedback Features                               | 97               |
| FFTs            | Formative Feedback Tools                        | 42               |
| fMRI            | functional Magnetic Resonance Imaging           | 195              |
| HCI             | Human Computer Interaction                      | 177              |
| IS              | Information Systems                             | 44               |
| IT              | Information Technology                          | 4                |
| ITPF            | IT-based Peer-Feedback                          | 10               |
| JIF             | Journal Impact Factor                           | 70               |
| LDA             | Latent Dirichlet Allocation                     | 114              |
| LI              | Learning Interest                               | 58               |
| LIWC            | Linguistic Inquiry and Word Count               | 113              |
| LOOM            | Learning Objective and Outcome Manager          | 42               |
| LOs             | Learning Outcomes                               | 45               |
| LUs             | Learning Units                                  | 45               |

## List of Acronyms

---

| Notation | Description                                            | Page List |
|----------|--------------------------------------------------------|-----------|
| MA       | Mastery Approach                                       | 59        |
| MCQs     | Multiple Choice Questions                              | 6         |
| MOOCs    | Massive Open Online Courses                            | 19        |
| MRs      | Meta-Requirements                                      | 47        |
| NLP      | Natural Language Processing                            | 37        |
| OECD     | Organization for Economic Co-operation and Development | 3         |
| PA       | Peer Assessment                                        | 6         |
| PanelOLS | Panel Ordinary Least Squares                           | 11        |
| PI       | Principles of Implementation                           | 63        |
| PLS      | Partial Least Squares                                  | 59        |
| POS      | Part-of-Speech                                         | 115       |
| pp       | percentage points                                      | 3         |
| PSD      | Perceived Skill Development                            | 58        |
| RQs      | Research Questions                                     | 6         |
| SA       | Self-Assessment                                        | 5         |
| SE       | Self-Efficacy                                          | 59        |
| SEM      | Structure Equation Model                               | 59        |
| SRL      | Self-Reported Learning                                 | 58        |
| TF       | Text Features                                          | 97        |
| TF-IDF   | Term Frequency-Inverse Document Frequency              | 115       |
| TML      | Technology-Mediated Learning                           | 20        |
| TP       | Testable Propositions                                  | 62        |
| VIF      | Variance Inflation Factor                              | 60        |
| VSM      | Value Survey Model                                     | 138       |
| VUCA     | Volatility, Uncertainty, Complexity, Ambiguity         | 2         |

## **Abstract**

The characteristics of the world of the 21st century are volatility, uncertainty, complexity and ambiguity. This forces employees to adapt to a new way of working which requires a different kind of skill-set, namely higher-order thinking skills. Since the main goal of universities is to prepare students for the future, they need to provide the conditions for training these skills.

Two factors are essential for achieving this goal: first, a realistic teaching-learning scenario that allows students to apply the skills that need to be developed and second, the provision of rich and meaningful feedback of the students' progress. However, universities are under high pressure due to an increasing number of students and limited budget. Thus, they face the challenge of how to provide students feedback in their learning process, so that they are able to develop their 21st century skills.

This dissertation project addresses this research gap. It aims to develop a teaching-learning scenario which combines the concepts of technology-mediated learning and peer learning to meet the challenges of large-scale classes. This comprises: (1) designing a system to provide technology-mediated formative feedback, (2) evaluating previous research on how to measure feedback quality, (3) deriving characteristics describing high and low quality peer feedback, (4) developing design principles for creating a culturally sensitive peer feedback process, and (5) measuring the effects of the technology-mediated feedback on student performance.

This dissertation contributes to theory in three key ways. Firstly, it generates knowledge on how to design systems that provide technology-mediated formative feedback. Secondly, it identifies a set of empirically grounded guidelines on measuring peer feedback quality. Thirdly, it produces prescriptive knowledge on the effects of technology-mediated formative feedback on student performance.



## **Zusammenfassung**

Die Welt des 21. Jahrhunderts zeichnet sich durch Volatilität, Unsicherheit, Komplexität und Mehrdeutigkeit aus. Dies zwingt Mitarbeiter, sich an eine neue Arbeitsweise anzupassen, die Fähigkeiten auf höheren Komplexitätsstufen erfordert. Da das Hauptziel der Universitäten darin besteht, Studierende auf die Zukunft vorzubereiten, müssen sie die Voraussetzungen für die Entwicklung dieser Fähigkeiten schaffen. Zwei Faktoren sind von wesentlicher Bedeutung: Erstens, ein realistisches Lehr-Lernszenario, das es den Studierenden ermöglicht, die zu entwickelnden Fähigkeiten anzuwenden und zweitens, das Geben von reichhaltigem und aussagekräftigem Feedback über die Fortschritte der Entwicklung. Allerdings stehen die Universitäten aufgrund einer steigenden Zahl von Studierenden und eines oft begrenzten Budgets unter hohem Druck. Daher stehen sie vor Herausforderung, wie sie Studierenden in ihrem Lernprozess Feedback geben können, um die Fähigkeiten die sie im 21. Jahrhundert benötigen, zu entwickeln.

Ziel dieser Dissertation ist die Entwicklung eines Lehr-Lernszenarios, dass die Konzepte des technologiegestützten-Lernens und des Peer-Lernens kombiniert, um den Herausforderungen von grossen Vorlesungen gerecht zu werden. Dies umfasst: (1) den Entwurf eines Systems zur Bereitstellung von technologiegestütztem Feedback, (2) Evaluierung des vorhandenen Wissens und der Literatur zur Messung von Feedbackqualität, (3) die Ableitung von Eigenschaften, die qualitativ hoch- und minderwertiges Peer-Feedback charakterisieren, (4) die Entwicklung von Gestaltungsprinzipien für einen kultursensiblen Peer-Feedback-Prozess und (5) die Messung der Auswirkungen des technologiegestützten Feedbacks auf die Leistung von Studierenden. Die vorliegende Dissertation leistet in dreifacher Hinsicht einen wichtigen Beitrag zur Theorie. Zum einen wird Wissen generiert, wie man Systeme entwirft, die technologiegestütztes formatives Feedback geben. Zweitens identifiziert sie eine Reihe von empirisch fundierten Leitfäden zur Messung von Peer-Feedback Qualität. Drittens generiert sie präskriptives Wissen über die Auswirkungen von technologiegestütztem formativen Feedback auf die Leistungen von Studierenden.





# 1

## Introduction

“Preparing students for jobs that don’t yet exist, using technologies that haven’t been invented, in order to solve problems we don’t even know are problems yet.”

---

— *Richard Riley (2014), Secretary of Education under Clinton*

### 1.1 Motivation

Today's world is changing faster than ever before. Students graduating from university find a world that can be characterized by the acronym *VUCA*, which stands for volatility, *uncertainty*, complexity and *ambiguity* (Bennis and Nanus 1985). Markets and customer preferences are more volatile than in the past, and challenge companies to adapt to changing conditions quickly. Thus, decisions need to be made under higher levels of *uncertainty*, and the problems to solve are increasingly complex and interdependent due to the increasing globalization and interconnectedness of the economy. Furthermore, large amounts of data often lead to *ambiguity* during decision making.

To respond to VUCA conditions, over the last two decades a new movement has emerged which is shaped by an agile mindset (Beck et al. 2001). More and more companies have adapted to this mindset and have started to work in small agile teams. In these teams, members work in short cycles with rapid feedback to respond quickly to changing conditions (Beck et al. 2001).

However, this 'new' way of working requires employees to have a different kind of skill-set, with mainly higher-order thinking skills. Two of the most important skills are (1) being able to reflect on the current state of knowledge and skills and (2) being able to provide feedback to team members to steer them in the right direction (Williams and Cockburn 2003; Highsmith and Cockburn 2001). Self-reflection of one's own current state of knowledge and condition is the foundation for adjusting actions to reach a goal. Reflection entails the comparison of one's own perception with an external perception. If a gap between the internal and external perception is revealed, then feedback is needed to derive action steps for moving forward. Thus, team members need to be able to provide constructive feedback to their peers.

When the purpose of universities is to educate future students working in such agile teams, those skills need to be developed at these institutions. As a consequence the question arises: how can universities prepare students for working in agile teams and help them develop higher-order thinking, in particular the ability to self-reflect and provide feedback?

According to Hattie and Timperley (2007), there are two essential factors for developing higher-order thinking skills: Factor A, a realistic teaching-learning scenario that allows students to apply the skills that need to be developed and Factor B, the provision of rich and meaningful feedback of the students' progress. However, in trying to provide those two boundary conditions, most universities are struggling due to an increasing number of students. The constraints are that the number of students in tertiary education are rising worldwide. For example, focusing on a four-year period from 2014 to 2017, the data shows that the number of students rose in China by 5.26percentage points (pp), Germany by 6.16pp, Switzerland by 3.76pp and worldwide by 4.52pp. The United States of America is an exception, with a decrease of 3.48pp (UNESCO Institute for Statistics 2019).

Moreover, not only are the number of students rising, but universities are also under increasing pressure due to limited financial resources. Government spending<sup>1</sup> on tertiary education decreased from 2013 to 2016 for Germany by 1.46pp and for the United States of America by 3.19pp (no data available for China and worldwide). Switzerland is an exception which increased by 6.42pp (UNESCO Institute for Statistics 2019). The statistics provided by the Organization for Economic Co-operation and Development (OECD) show similar results for a ten-year period. According to this data, the amount of students at universities rose by 15pp in the US and by 29pp in Germany but public spending for education decreased by 7pp in the US and 1pp in Germany between 2005 and 2014 (OECD 2016).

Consequently, the research problem I address in the dissertation is how to create and evaluate a challenging teaching-learning scenario that helps students develop their higher-order thinking skills through providing technology-mediated formative feedback. This is researched in the context of boundary conditions, where the constraints are that lecture sizes keep growing and financial resources stay constant or are even decreasing. These conditions mean that increasing teaching resources by scaling the number of teaching assistants and applying a tutor feedback model such as the one proposed by Marjanovic (2016) is not viable for the majority of universities.

<sup>1</sup> In constant US dollar, considering inflation

Importantly, the term feedback is used in two ways in the dissertation, first, as the information given to students to develop a skill and, second, as the skill of providing the feedback itself. In the first context, in which formative feedback is given to develop a skill, formative feedback usually consists of two parts: the internal assessment and the external assessment of one's own perception of the current state of knowledge or skill development. In the second context in which feedback is a 21st century skill (i.e., providing feedback to another person), the focus lies on written textual feedback (Trilling and Fadel 2009).

To address the described problem, one solution is to use Information Technology (IT). IT can help to leverage scaling effects and address the problem of large-scale classes. Hence, the objective is to design and create an IT-enhanced teaching-learning scenario in which students are able to a) develop their reflection and feedback skills and b) receive technology-mediated formative feedback during the learning process.

Formative feedback, which is a specific form of feedback, is the core element in this teaching-learning scenario (Kluger and DeNisi 1996; Hattie and Timperley 2007; Shute 2008; Price et al. 2010). The objective of formative feedback is to provide feedback multiple times during the learning process, which the students are able to use to immediately adjust their actions (Dawson et al. 2016). Thus, formative feedback itself is seen as a process rather than a one-time event as is the case with summative feedback. The purpose of summative feedback is to provide an evaluation of understanding, for example at the end of a semester, usually in the form of a grade. Formative feedback used in this dissertation is always technology-mediated.

Hattie (2008) synthesized 800 meta-analyses of the effects of e.g. feedback, pre-knowledge and assessment, on student performance, while 23 were specifically on the effects of feedback on student performance. The results revealed that formative feedback is one of the top influential factors for developing higher-order thinking skills. Formative feedback unveils the gap between the students' internal and external assessment of their own perception of their current state of knowledge (Shute 2008). Their awareness of this gap and efforts to close it help students to develop their self-reflection skills.

Hence, the aim of this dissertation is to leverage the potentials of modern IT to design, develop and evaluate a teaching-learning scenario that addresses two main questions. First, how can universities provide students with a teaching-learning scenario to develop higher-order thinking skills such as self-reflection and the skill of providing feedback<sup>2</sup>? Second, how can formative feedback be provided during the learning process to develop these two skills?

## 1.2 Research Questions and Research Methods

I developed an IT-enhanced teaching-learning scenario for this dissertation which consists of three phases: *pre-phase*, *during-phase* and *post-phase*. These three phases link to the teaching intervention within the teaching-learning scenario such as a lecture and hence describe the phases before, during and after the lecture.

In each of these phases students receive formative feedback with the objective to enhance student performance holistically. The pre-phase comprises the period before the teaching intervention (e.g. a lecture, a seminar or an exercise session). The objective of this phase is that students receive formative feedback to ensure preparedness, for example, in flipped classroom concepts.

The during phase is the period when the teaching intervention takes place (e.g. a two hour lecture). In this phase students receive formative feedback to deepen the learned material in class.

The post phase is the period after the teaching intervention. Students receive formative feedback to develop higher-order thinking skills and deepen long-term knowledge. It needs to be noted that a lecturer can define hard boundary conditions for pre- and post-phases, however, this is not always the case and both phases can even overlap. Nevertheless, each phase has distinct purposes.

Formative feedback provided in each of the phases consists of two components: a) *internal assessment* and b) *external assessment*. *Internal assessment* is conceptualized as Self-Assessment (SA). The SA is an assessment made by the student of his or her current state of knowledge or skill (Boud 1991). SA is

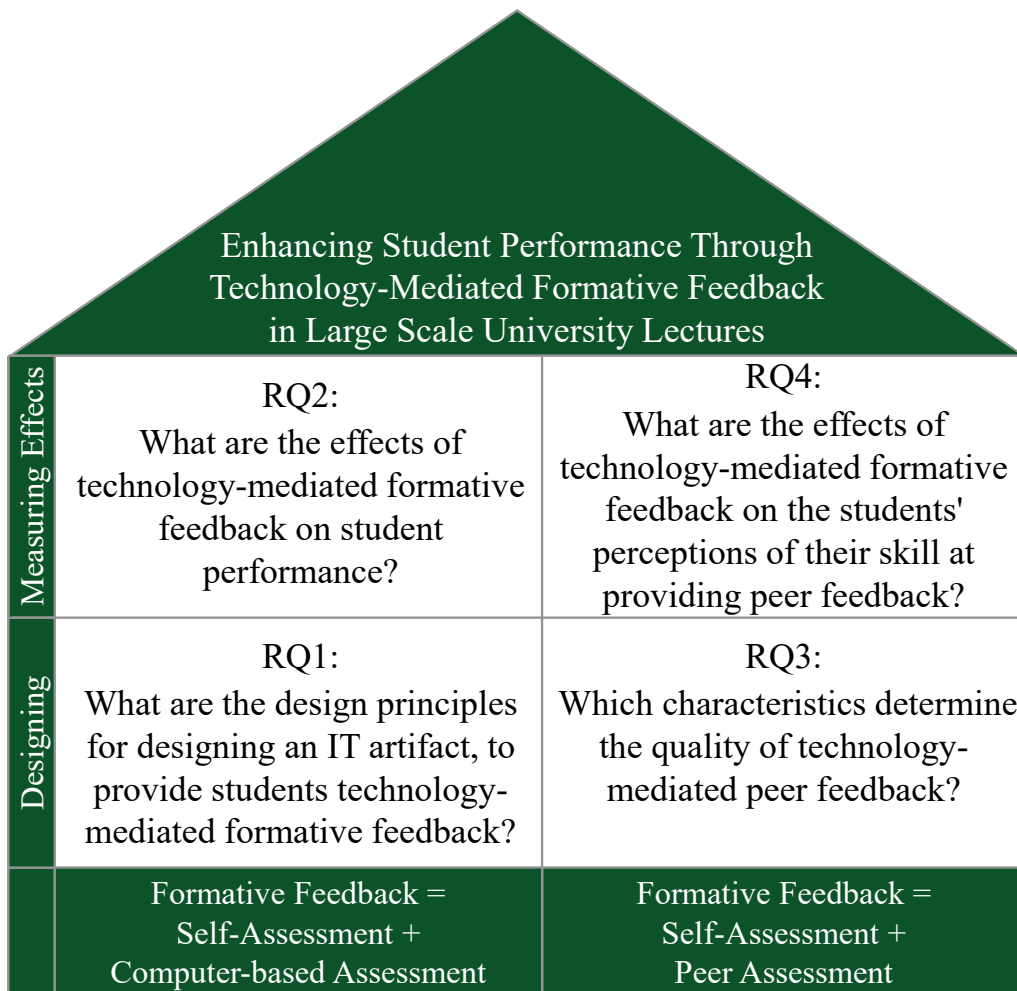
<sup>2</sup> This skill will be more simply referred to as ‘feedback providing skill’ in this dissertation

measured through items of the construct current state of knowledge or current state of skill level developed.

The *external assessment* can be automated by using Computer-based Assessment (CbA, Higgins and Bligh 2006) or can be provided by a human by using Peer Assessment (PA, Falchikov 2001). CbA includes two assessment types: Multiple Choice Questions (MCQs) and flow diagrams. For both assessment types, students receive an automated evaluation of the correctly and incorrectly answered questions. When PA is used, this includes the evaluation of the students' current skill level developed. PA is measured through items of the construct peer feedback quality.

Overall the dissertation uses this developed teaching-learning scenario and tries to answer four Research Questions (RQs). The inter-relationship of these questions are shown in Figure 1.1. Research question one and two conceptualize formative feedback as SA and CbA. Research question one focuses on the design of the IT artifact that provides formative feedback to enhance student performance. Research question two is related to measurement effects when using an instantiation of the IT artifact. Hence, this second question tries to evaluate whether there is a statistically significant effect of formative feedback SA and CbA on student performance.

Research question three and four conceptualize formative feedback specifically as SA and PA. Research question three investigates how to measure the characteristics of high and low peer feedback quality. The answer to this question builds the foundation for developing an algorithm which automatically assesses peer feedback quality. Finally, research question four tries to provide insights into whether there is a statistically significant effect when students receive formative feedback on their ability to provide feedback, in particular, whether the students' perceived skill in providing feedback increases when receiving formative feedback.



**Figure 1.1:** Overview of the Research Questions

### **RQ1: What are the design principles for designing an IT artifact, to provide students technology-mediated formative feedback?**

There are many learning tools for higher education in the market. Popular ones are Blackboard<sup>3</sup>, Canvas<sup>4</sup> and Moodle<sup>5</sup>. However, most of the IT tools have in common that the purpose is to provide students with learning materials such as lecture slides and additional readings, offering forums to share knowledge or providing a platform to organize assignments (Dobre 2015). Some IT tools have additionally integrated an add-on module to support the creation of MCQs.

However, less is known about how formative feedback tools should be designed, what their components should be and the optimal relationship between them.

The aim of the first research question is to identify what the requirements and design principles should be for building an IT-tool that provides formative feedback to enhance student's performance where the formative feedback consists of the results of a SA and CbA.

The methodology I used to answer this research question followed Design Science Research (DSR) and in particular I used the three-cycle DSR approach developed by Hevner (2007). This approach provides a scientific method to solve a set of practical problems that researchers and practitioners experience in their own practice. Furthermore, I used the framework of Gregor and Jones (2007) to document the design knowledge of the IT artifact which is the outcome of the iterative DSR process.

### **RQ2: What are the effects of technology-mediated formative feedback on student performance?**

Hattie and Timperley (2007) synthesized the results of 23 meta-analyses including 1,287 studies in education. The data revealed that feedback had an average effect size<sup>6</sup> of 0.79 on student performance, whereby the mean effect size of all (e.g. frequency of testing, mentoring, individualized instructions) effects was 0.4. When considering subclass formative feedback the effect size is even 0.9,

---

<sup>3</sup> [www.blackboard.com](http://www.blackboard.com)

<sup>4</sup> [www.canvas.net](http://www.canvas.net)

<sup>5</sup> [www.moodle.com](http://www.moodle.com)

<sup>6</sup> Effect size is the importance of the magnitude of differences (Hattie 2008):

$$Effectsize = [Mean_{treatment} - Mean_{control}] / SD$$



which makes formative feedback to the top three factors influencing student performance.

However, before feedback can be provided, an assessment needs to take place which evaluates the current state of knowledge or skill developed. Previous research shows that carrying out a SA of one's current state of knowledge, improves educational performance (Nicol and Macfarlane-Dick 2006). According to Sadler (1989) students who carried out a SA of their own work improved their learning of the course being studied and this SA provided a foundation for lifelong learning. However, SA by itself can lead to an overestimation of one's skills. Kruger and Dunning (1999) found out that overestimation is often related to a lack of meta-cognitive skills among less skilled participants. Hence, SA should be paired with an external assessment in order to foster awareness and understanding of how to control learning. This should, consequently, help to develop self-regulated learning skills which are essential for the development of higher-order thinking skills (Nicol and Macfarlane-Dick 2006).

There are studies showing the effectiveness of internal and external assessment building formative feedback on student performance. However, there are two gaps in the literature: a) most of the studies have been carried out in small or medium sized lectures at schools (Hattie and Yates 2013) and b) the existing studies have not investigated the simultaneous use of internal assessment such as SA and external assessment such as CbA.

The aim of the second research question is to investigate the effects of formative feedback on students performance. Therefore I created a research model to explain the cause and effect relationship of a formative feedback tool on perceived and objective student performance. As the formative feedback tool I used the instantiated IT artifact based on the design knowledge which was the outcome of RQ1.

The methodology I used to answer this research question was quantitative data analysis using an experimental study. In the first step, I created the research model based on theory from scientific literature on perceived student performance (Alavi 1994) and objective student performance (Gupta et al. 2010). The research model was then tested in an experimental field study within the context of a large-scale class involving 189 first year masters' students.

### **RQ3: Which characteristics determine the quality of technology-mediated peer feedback?**

The higher-order thinking skill, ability to provide feedback, is a skill which is of particular importance for students working in agile teams. Feedback on a person's performance or behaviour from a fellow team member can help the person to adjust actions and steering them in the right direction (Whitworth and Biddle 2007). According to Hattie and Timperley (2007), there are two essential factors for developing higher-order thinking skill such as the ability to provide feedback: Factor A, a realistic teaching-learning scenario that allows students to apply the skills that need to be developed and Factor B, the provision of rich and meaningful feedback of the students' progress.

The Factor A is addressed by using the concept of peer feedback in which students provide each other feedback. The process of peer feedback is well known in academia, especially in the context of publishing academic journals (Ziman 1974). To determine the quality of a scientific paper, each paper undergoes a peer-feedback process. The paper is sent to three anonymous reviewers who evaluate its quality and provide feedback on the advantages and disadvantages of the paper and suggestions for improvement (Meadows 1998). In this current study this process is operationalized as IT-based Peer-Feedback (ITPF). However, Factor B remains challenging, if one wants to provide individualized formative feedback, the first step is to understand what are characteristics of peer feedback determine its quality and how to measure them automatically.

In the academic literature the characteristics determining high and low quality peer feedback have been defined in a very generic way. There is no comprehensive study which has operationalized the peer feedback characteristics and measured them empirically using a real world data set (Brookhart 2008; Shute 2008).

The aim of the third research question was to contribute prescriptive knowledge that extend the literature on feedback. The answer to this question consists of three components. Firstly, the analysis of related work on how to measure feedback quality. Secondly, through the generation of a list of 90 features of feedback, and evaluated statistical effects on feedback quality. This list is based

on characteristics from a) scientific literature and b) real world data. The real world data set contained 7,036 instances of peer feedback, which were collected in a quasi experimental setting from 2015 to 2018. Thirdly, through deriving design principles for designing cultural sensitive peer feedback.

The methodologies I used to answer this research question was firstly, a literature review, secondly a quantitative data analysis and thirdly a DSR approach. The structured literature review analysis 81 journal articles. In the quantitative data analysis I followed the CRoss-Industry Standard Process for Data Mining (CRISP-DM, Chapman et al. 2000). In an iterative process, the literature was evaluated and a framework containing three main categories with additional sub-categories and feature groups was derived. The feature groups contained deductively operationalized features from the literature and inductively operationalized features from analyzing the real word data set. I used *Panel Ordinary Least Squares (PanelOLS)* for measuring the effect of each feature on the category, sub-category and feature group level. Furthermore, I used the machine learning approach *random forest* to measure the overall importance of the features. In the DSR approach I derived requirements from scientific literature to develop cultural sensitive design principles for peer feedback. The design principles were tested in 16 semi-structured interviews.

#### **RQ4: What are the effects of technology-mediated formative feedback on the students' perceptions of their skill at providing peer feedback?**

Formative feedback is one of the most important factors for developing higher-order thinking skills such as the ability to provide peer feedback (Hattie and Timperley 2007). Besides providing objective formative feedback on the students ability to provide feedback, also the perceived skill development is important (Alavi 1994). For the perceived skill development the formative feedback consists of an internal assessment and external assessment of the student's own perceived skill level.

While the internal assessment is conceptualized as SA, that is the student's own perceived skill level, and the external assessment is conceptualized as PA, that is the estimated skill level from the perspective of the student's peers. If the judgements of SA and PA differ significantly, this can cause cognitive

dissonance in the student's mind and initiate self-reflection processes. This reflection helps the student to identify areas where improvement is needed.

There is a gap in the literature concerning the effects of providing a student with these specific form of formative feedback: internal assessment (e.g. SA) and external assessment (e.g. PA) on the perceived development of the ability to provide feedback. To the best of my knowledge, there is currently no research which has explored the effects when these two assessments are presented on the students' ability to provide feedback.

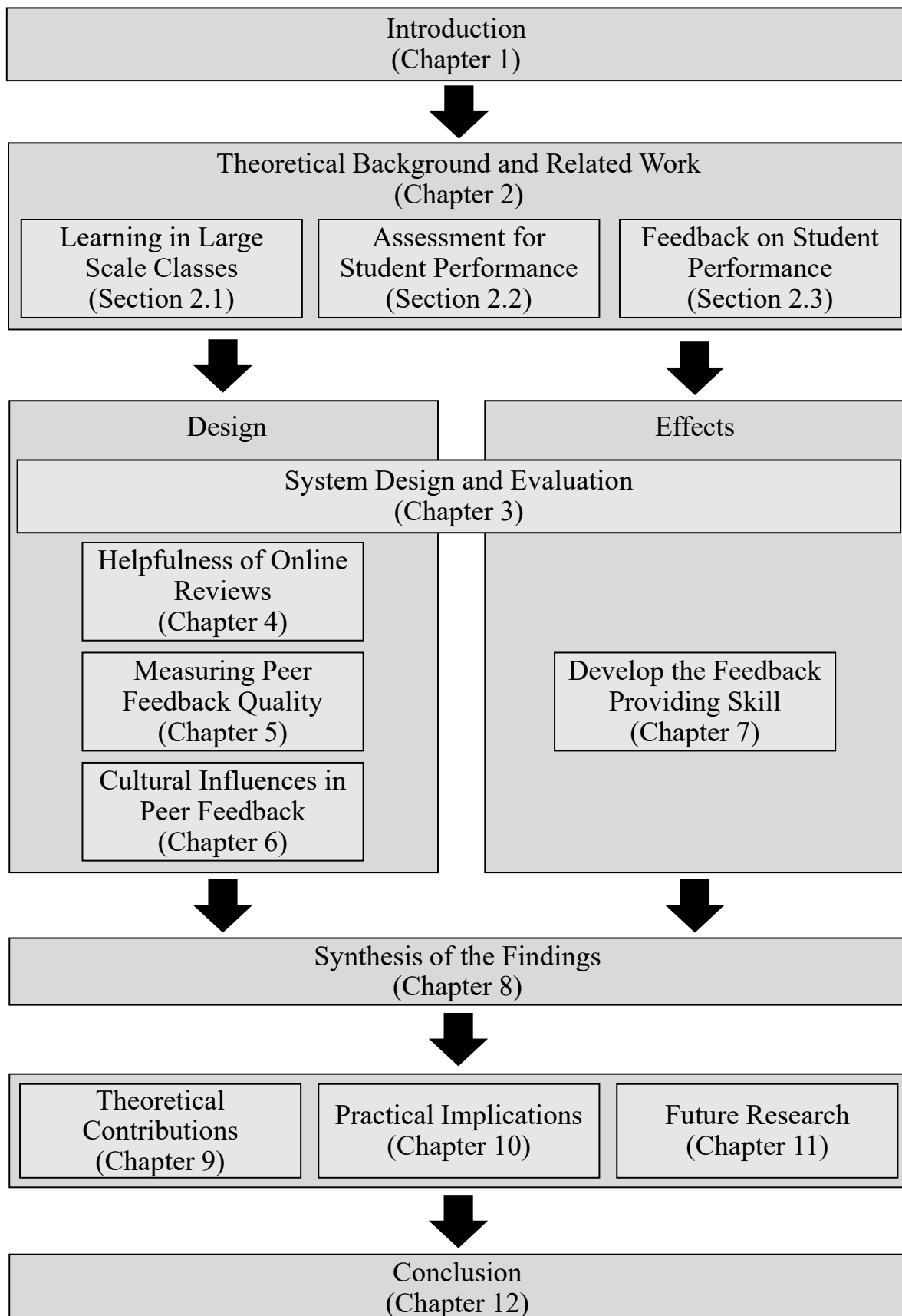
The objective of the fourth research question was to carry out an experiment to investigate empirically the relationship between formative feedback and the perceived development of the skill of providing feedback, in particular, whether the students' perceived ability to provide feedback improves over time when given formative feedback.

The methodology I used to answer this research question was a quantitative data analysis using an experimental design setting. I carried out a quasi-experiment with a one-group pre-test post-test design (Millsap and Maydeu-Olivares 2009) in a large-scale university class on business model innovation at Master's level using ITPF. The study consisted of seven ITPF iterations that took place in a weekly manner over a period of one semester. Between 73 and 101 students participated in each ITPF. The longitudinal data (seven observations) were analyzed by performing comparison analysis between SA and PA scores using one-sided, paired t-test.

### 1.3 Structure of the Dissertation

In line with the dissertation's objectives and to answer all of the underlying research questions, the structure of this dissertation is as follows (see Figure 1.2 for an illustration of the structure).

In the following Chapter, I will provide the theoretical background and discuss related work from the academic literature about how students learn in large-scale classes. This includes a definition of how assessment and feedback are understood and used in this dissertation as well as their characteristics. In Chapter 3, I will present a study that develops a design knowledge to design

**Figure 1.2:** Structure of the Dissertation

formative feedback IT artifacts. Furthermore, I discuss the results based on an evaluation of the instance used in a field experiment. Chapter 4 will discuss research on online review helpfulness which is a highly related research stream to peer feedback. I will illustrate how this knowledge can be used for measuring peer feedback quality in Chapter 5. The conducted study in this Chapter shows how I derive characteristics determining peer feedback quality based on the scientific literature and a real-world data set. Furthermore, I will present an approach for measuring those characteristics automatically using text mining techniques. Chapter 6 presents a experimental study on the effect of students' cultural differences on peer feedback. This includes design principles for developing culturally sensitive IT artifacts. In Chapter 7, I will present a study on how formative feedback consisting of an internal and an external assessment can have a positive impact on students' skills of providing peer feedback. Chapter 8 will provide a synthesis of findings of the presented studies in this dissertation and show the interrelationship between my research questions. Chapter 9 and 10 will discuss the theoretical contribution and the practical implications of the set of research carried out. Chapter 11 provides an overview of possible feature research, while Chapter 12 provides a conclusion of this dissertation.

## 1.4 Overview of Publications

During my early academic carrier, I have authored and co-authored the publications listed below. Since most publications are related to the subject and content of the dissertation, there will be an overlap between them. I will provide a note at the beginning of each chapter highlighting publications that significantly influenced the content of the specific chapter.

### 1.4.1 Published Journals

**Seufert, S., Meier, C., Soellner, M., and Rietsche, R. 2019.** "A Pedagogical Perspective on Big Data and Learning Analytics: A Conceptual Model for Digital Learning Support," *Technology, Knowledge and Learning* (24:4), pp. 599-619.

### 1.4.2 Journals Under Review

**Rietsche, R., Bauman, K., Soellner, M., and Leimeister, J. M.** “What Makes Great Feedback? Understanding and Measuring Technology-Mediated Peer Feedback Quality”.

- Journal: Information Systems Research (ISR)
- Status: Under review
- Submission Date: December 19, 2019

### 1.4.3 Conference Articles

**Lechler, R., Stöckli, E., Rietsche, R., and Uebersnickel, F. 2019.** “Looking Beneath the Tip of the Iceberg: The Two-Sided Nature of Chatbots and their Roles for Digital Feedback Exchange,” in Proceedings of the European Conference on Information Systems (ECIS), Stockholm, Sweden.

**Rietsche, R., Frei, D., and Soellner, M. 2019.** “Not all Reviews are Equal – a Literature Review on Online Review Helpfulness,” in Proceedings of the European Conference on Information Systems (ECIS), Stockholm, Sweden.

**Rietsche, R., and Söllner, M. 2019.** “Insights into Using IT-Based Peer Feedback to Practice the Students Providing Feedback Skill,” in Proceedings of the Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, USA.

**Wambsganss, T., and Rietsche, R. 2019.** “Towards Designing an Adaptive Argumentation Learning Tool,” in Proceedings of the International Conference on Information Systems (ICIS), Munich, Germany.

**Buholzer, F., Rietsche, R., and Söllner, M. 2018.** “Knowing what students Like? Developing a Cultural Sensitive Peer Assessment Process in MOOCs,” in Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI), Lüneburg, Germany.

**Rietsche, R., Duss, K., Persch, J. M., and Soellner, M. 2018.** “Design and Evaluation of an IT-based Formative Feedback Tool to Foster Student Per-

formance,” in Proceedings of the International Conference on Information Systems (ICIS), San Francisco, California, USA.

**Rietsche, R., Lehmann, K., Haas, P., and Söllner, M. 2017.** “The Twofold Value of IT-Based Peer Assessment in Management Information Systems Education,” in Proceedings of the International Conference on Wirtschaftsinformatik (WI), St. Gallen, Switzerland.

**Rietsche, R., and Söllner, M. 2017.** “The obvious and hidden values of IT-based peer assessment in management education,” Paper presented at the Academy of Management Annual Meeting (AOM). Atlanta, Georgia, USA.

**Rietsche, R., Söllner, M., and Seufert, S. 2017.** “Digital formative learning assessment tool - Towards helping students to take ownership of their learning,” in Proceedings of the European Conference on Information Systems (ECIS), Guimarães, Portugal.

**Kopf, S., Schön, D., Guthier, B., Rietsche, R., and Effelsberg, W. (eds.) 2015.** A real-time feedback system for presentation skills, Association for the Advancement of Computing in Education (AACE), Waynesville, NC.

**Rietsche, R., Russell, G., and Karduck, A. P. 2011.** “Creation of an Online Virtual Cisco Router Learning Environment,” in Proceedings of the 14th IASTED international conference on Computers and advanced technology in education, Cambridge, UK.



# 2

## **Theoretical Background and Related Work**

“Nanos gigantum humeris  
insidentes. Discovering truth  
by building on previous  
discoveries”

---

— *Bernard of Chartres (12th  
Century), Philosopher, Scholar, and  
Administrator.*

### 2.1 Learning in Large-Scale Classes

This dissertation<sup>1</sup> focuses on how formative feedback can be used to improve student learning when formative feedback consists of multiple components, such as the student's internal assessment of his or her own knowledge and an external assessment of the student's current state of knowledge. An external assessment can be, for example, computer generated or given by a peer. For clarity, I will interpret and use the terms *assessment* and *feedback* in the following way. *Assessment* is conceptualized as the action of evaluating while *feedback* is conceptualized as the action of communicating the results of this evaluation. Therefore, in this context, assessment is a prerequisite for feedback. Since, the overall assumption of this dissertation implies of how this formative feedback can be provided in large-scale classes, theoretical concepts have been chosen accordingly. In the last three decades theoretical concepts such as technology-mediated learning and peer learning emerged to provide an answer for conditions of large-scale classes and often steady or decreasing financial resources. Thus, both concepts form the foundation of my research.

#### 2.1.1 Characteristics of Large-Scale Classes

Increasing lecture sizes is a worldwide phenomenon which also affects the university in which I am writing this dissertation. In 2018 the University of St.Gallen had 1647 students<sup>2</sup> (Universität St.Gallen 2019). This led to lecture sizes of over 800 students per lecture. However, the question is when is a lecture large-scale? There are no definitive values for this (Mulryan-Kyne 2010).

Instead, research has shown that the effects of a large-scale class are dependent on the context. For example, if a science laboratory lecture was designed for 30 students, but 50 attend, it might be very challenging to teach the larger class as effectively (Biggs 1999). Nevertheless, it might not cause as many issues to teach 230 students in a management lecture using presentation slides even if the

---

<sup>1</sup> Parts of this chapter have been published in proceedings of peer-reviewed conferences (Rietsche and Söllner 2019; Rietsche et al. 2019; Buholzer et al. 2018; Rietsche et al. 2018) or are currently under review in academic journals (Information Systems Research). The content has been reformatted, restructured, modified, and extended to provide a comprehensive theoretical background for this dissertation.

<sup>2</sup> First and second semester of undergraduates studies

class was originally designed for 200 students. Furthermore, it depends on the students' course level and the learning objectives (Mulryan-Kyne 2010). For example, if early bachelor students need to learn the fundamentals of management and achieve the learning outcomes of remembering or understanding in the learning taxonomy, 200 students might be a manageable large-scale class. However, if Master's students need to apply their knowledge and skills and develop higher-order thinking skills, a lecture size of 100 students might already be too large to achieve this goal.

According to a study conducted in Haiti, a lecture with 50 to 80 students is already considered to be a large lecture (Renaud et al. 2007). Yang (2008); Nicol and Boyle (2003); Carbone and Greenberg (1998) define a lecture with 100 students as a very large lecture, which already leads to difficulties such as how to engage with students and provide them with feedback. Fortes and Tchantchane (2010) use the term very large classes for 150-200 students and Felder (1997) for 150-300 students. A study in New Zealand in 2007 stated that at their university, 30 courses consisted of over 500 students while a third even exceed 1000 students (Exeter et al. 2010). Probably one of the largest classes are Massive Open Online Courses (MOOCs) which have up to over 150,000 registered students (Allen and Seaman 2015).

Overall, most researchers acknowledge that large-scale classes are problematic due to multiple reasons. The main issues are that large-scale classes lead to poor student engagement, less interaction and a lack of feedback (Fortes and Tchantchane 2010; Biggs 1999; Gibbs 1992). The latter issue is of particular concern especially when it comes to developing higher-order thinking skills, for example the skill of providing high quality feedback, critical thinking and self-reflection including how to use learned methods/information. To develop these kinds of skills, a student needs to be a 'deep' learner, apply theories, put their skills into action and reflect on their current state of knowledge (Biggs 2011; Anderson et al. 2001; Bloom et al. 1984).

Thus, if classes sizes exceed a threshold in which lecturers are not able to provide feedback, alternative approaches need to be adopted. Two concepts that are useful in understanding how to achieve alternative forms of feedback are technology-mediated learning (Alavi and Leidner 2001) and peer learning (Nicol and Boyle 2003). Both concepts complement each other and serve

to address the full spectrum of learning goals in the revised Bloom learning taxonomy (Krathwohl 2002).

### 2.1.2 Technology-Mediated Learning

The development of Technology-Mediated Learning (TML) has gained increasing attention since the late 1990s. Consequently, several researchers have conducted studies in the area of TML to investigate the potential of technology for different teaching and learning settings. TML is defined as “an environment in which the learner’s interactions with learning materials, peers, and/or instructors are mediated through advanced information technology” (Alavi and Leidner 2001, p. 2). Technology offers several opportunities including being a tool that can help manage large-scale classes. For example, it can aid in coordinating group work, planning assessments, supporting collaborative learning in transcontinental teams, or, as in this research, facilitate teaching-learning scenarios and peer feedback (Gupta and Bostrom 2013; Alavi et al. 1997). With the ongoing transformation of higher education, including calls for increasing technology-based personalization of learning processes, we can expect that the importance of understanding the effects of, as well as designing effective, TML scenarios will remain high or even increase in the future (Lucas 2014; Christensen et al. 2011; Klačnja-Milićević et al. 2011).

For example, Bauman and Tuzhilin (2018) illustrated how machine learning approaches can help to provide personalized recommendations to students, thus which learning materials should be used next to fill existing knowledge gaps. Additionally, other technology-based scaffolds can be used to individualize learning processes and foster learning outcomes such as problem-solving (Janson et al. 2019).

Thus, TML supports the provision of formative feedback in large-scale classes for lower and higher-order thinking, by facilitating the provision of internal and external assessment to improve student performance. TML is especially important for scaling peer learning concepts.

### 2.1.3 Peer Learning

Peer learning approaches have become increasingly popular recently, partly because of their potential to overcome the challenge of an increasingly higher student to lecturer ratio, but also because of several learning-related advantages (Boud et al. 2014). Peer learning is of importance for developing higher-order thinking skills. When comparing peer learning to traditional teaching-learning scenarios, the key difference is that in peer learning, “students [are] learning from and with each other in both formal and informal ways” (Boud et al. 2014, p. 4).

Hence, peer learning is of special importance for skill development because students need to put skills into action. For example, students need to provide feedback to their peers which enables them to develop their ability to provide feedback, and in collaborative settings students exchange best practices and tacit knowledge such as in a task of developing a business model (Boud et al. 2014; Nonaka and Takeuchi 1995). Research on the latter was carried out by Chan et al. (2014) who used peer learning to increase productivity and knowledge transfer in sales personnel. Their research shows that weekly sales increased by using peer learning concepts such as observing successful sales techniques of peers and direct peer teaching. The concepts of technology-mediated learning and peer learning were combined in this dissertation to design the IT-based peer feedback (ITPF) teaching-learning scenario which is explained in more detail in Section 2.3.

Peer learning builds on the fundamental educational idea of *constructivism*, which is a theory about how people learn. It emphasizes on the importance of existing individual knowledge, beliefs and skills which are build on in the experience of learning (Seel 2011). According to Poplin (1988), constructivists posit that learning is a process whereby new meanings are created (constructed) by the learner within the context of their current knowledge. Hence, learning is an active process of constructing rather than acquiring knowledge and teaching is a process of supporting that construction rather than communicating knowledge (Cunningham and Duffy 1996). If the student acquires the strategies that meet the objective, then learning has occurred and the measurement of this learning can only be done through estimation with observation or dialogue (Glaserfeld

1989). Hence, constructivism is the concept that underpins peer learning. In the teaching-learning scenario used in this dissertation, students worked together to develop higher-order thinking skills, while the setting was technology-mediated, asynchronous and anonymous. Students created submissions and their peers evaluated the quality of them. This peer learning involved three interactions: student-content interaction, student-lecturer interaction, and student-student interaction (Moore 1989). These interaction types are of importance, because interaction provokes learning activities that demand an exchange between content, lecturers and students.

The *student-content* type of interaction can be described as the “internal didactical conversation” that students engage in when they “talk to themselves” about the information and ideas they encounter in papers, textbooks, lectures slides and other educational materials (Holmberg 1986). Hence, initiating the process of intellectually interacting with the content is the basis for changing the students’ understanding, perspective or cognitive structures of their mind (Moore 1989). In developing higher-order thinking skills, student-content interaction is an important factor and takes place when the student is in the creator and reviewer role. In the creator role students need to create a submission and in the reviewer role students need to evaluate the quality of the submission and write constructive feedback.

The *student-lecturer* interaction, the second type of interaction, has three main goals. Firstly, the lecturer provides access and prepares learning materials for the students so that they can construct their own knowledge (Cunningham and Duffy 1996). Secondly, the lecturer seeks to enhance and maintain the students’ interest in the taught topic and to motivate the student to learn, including via self-direction (Moore 1989). Thirdly, the lecturer provides a work-environment and best practices on how to work together in peer learning, for example by providing instructions and examples on how to write good feedback.

The third type of interaction is the *student-student* interaction whereby students interact in pairs or in groups to construct their own knowledge. This facilitates the development of higher-order thinking skills that are needed in business contexts such as critical thinking skills by enabling students to reflect on each other’s ideas as well as providing the opportunity for them to provide high quality feedback to their peers (Dai et al. 2010).

In the teaching-learning scenario used in this dissertation, the *student-student* interaction is the foundation of how students construct their own knowledge. Through critically reviewing their peers' submission, the (student) reviewers need, on the one hand, to reflect on their own knowledge and, on the other hand, to think about how to structure the review so as to convey their thoughts on how to improve the work of their peers. Thus, the reviewers need to put themselves into the position of the peer student to communicate the ideas as convincingly as possible so that the student can most benefit from them. The three discussed interaction types give important insights into why interaction is necessary for student performance and skill development.

#### 2.1.4 Student Performance

The aim of formative feedback provided through technology-mediated learning and peer learning is to overall improve student performance. There are different definitions of student performance, which depend on the purpose of teaching and learning. The most widely used definition in the context of technology-mediated learning is based on Gupta et al. (2010) who defined four categories of learning goals indicating student performance: *skill-based*, *cognitive*, *affective* and *metacognitive* goals.

*Skill-based* learning goals focus on the ability of students to use learned methods such as a business process modelling language. *Cognitive* learning goals focus on the ability and judgment of students. For example, do the students understand the concept of the lecture and are they able to choose the right tool for the given business problem? *Affective* goals focus on the emotional aspects of the students' behavior such as their motivational knowledge (e.g., usefulness of the IT tool) or satisfaction with the learning process (Söllner et al. 2017). *Metacognitive* or self-regulated learning goals refers to the students' knowledge of their own learning and information processing (Gupta et al. 2010).

The learning goals underpinning student performance can be evaluated from two different perspectives. The final exam of a course can be used to gain insight into the objective student performance, while survey items can be created to gather information about students' perceptions of their level of performance. In this dissertation for the development of formative feedback tools, skill and cognitive

## 2. Theoretical Background and Related Work

|                     |                         | Cognitive Process Dimension                                                                                      |            |       |                                                                                                                                                                                                          |          |        |
|---------------------|-------------------------|------------------------------------------------------------------------------------------------------------------|------------|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|--------|
|                     |                         | Remember                                                                                                         | Understand | Apply | Analyze                                                                                                                                                                                                  | Evaluate | Create |
| Knowledge Dimension | Factual Knowledge       |                                                                                                                  |            |       | <div>Higher Order Thinking Skills:</div> <ul style="list-style-type: none"><li>Critical thinking</li><li>Communication and Collaboration such as ability to provide high quality peer feedback</li></ul> |          |        |
|                     | Conceptual Knowledge    |                                                                                                                  |            |       |                                                                                                                                                                                                          |          |        |
|                     | Procedural Knowledge    |                                                                                                                  |            |       |                                                                                                                                                                                                          |          |        |
|                     | Metacognitive Knowledge | <div>Higher Order Thinking Skills:</div> <ul style="list-style-type: none"><li>Self regulated learning</li></ul> |            |       |                                                                                                                                                                                                          |          |        |

**Figure 2.1:** Higher-Order Thinking Skills in the Revision of Bloom's Taxonomy (Krathwohl 2002)

goals were analyzed in terms of both objective and perceived performance. Affective and metacognitive goals, however, were analyzed with regard to perceived performance only because they cannot be measured by evaluating exam results (Chemers et al. 2001).

The early work of Bloom (1956) was the first attempt to categorise and formalize learning goals into a taxonomy. The overall idea was to create a taxonomy which would form the basis for course curriculum creation and for the assessment of learning goals to determine whether the set goals were actually achieved. Nearly 50 years later, Krathwohl (2002) published a revised version of Bloom's Taxonomy, which is shown in Figure 2.1). The taxonomy consists of six cognitive process dimensions and four knowledge dimensions. The first three cognitive processes dimensions (remember, understand and apply) are often associated with lower-order thinking dimensions, while the next three dimensions (analyze, evaluate and create) are associated with higher-order thinking dimensions. The taxonomy provides the possibility to assign learning goals and skills to each of the dimensions in the framework. However, it needs to be stated that this framework is only based on cognitive and meta-cognitive dimensions and does not take affective learning goals into account.

The main focus of this dissertation is on the higher-order cognitive process dimensions as well as the meta-cognitive dimension (see Figure 2.1, since both are important in the development of higher-order thinking skills. The meta-cognitive dimension in particular reflects the students' ability for self-



regulated learning and perceived learning success (Rietsche et al. 2018; Butler and Winne 1995; Alavi 1994). Formative feedback stimulates self-reflection and self-regulated learning processes in the learners mind and those processes are an essential factor for developing higher-order thinking skills (Nicol and Macfarlane-Dick 2006; Butler and Winne 1995; Boud 1990).

### **2.1.5 Higher-Order Thinking Skills**

Higher-order thinking skills are skills that are at the higher level (4-6) of the Krathwol and Bloom taxonomy (Krathwohl 2002). Although there is no single definition that encapsulates higher-order thinking, these skills can be described by their attributes (Resnick 1987). Higher-order thinking is non-algorithmic, meaning that there is not a single true solution or path. These skills tend to be complex, entailing multiple steps and processes. For example, abstracting problems is a higher-order skill which requires dividing the problem into smaller subsets, evaluating these complex problems and judging the advantages and disadvantages of each possible solution. Those steps are often part of the overall skill of critical thinking (Ennis 2011). Furthermore, higher-order thinking involves the ability to communicate the results of one's own critical thinking process. This is often the case when someone needs to provide feedback about another person's performance. According to Nicol et al. (2014), creating high quality peer feedback is a fundamental graduate skill and should be taught in higher education curricula.

Overall, one of the underpinning principles of higher-order thinking is self-regulation (Butler and Winne 1995). This can be defined as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment” (Pintrich 2000, p. 453).

Thus, self-regulated learning involves the ability to assess one's current state of knowledge or skill level and the gap from the desired goal. This initiates thinking processes in the students' mind in order to create a plan and develop ideas on how to close this gap. In the context of providing feedback, the feedback provider needs to interpret the submission, judge its quality and develop suggestions

for improvement. The production of explanations, by generating feedback comments, about what makes the work strong or weak initiates self-regulation processes within the feedback provider's mind. They have to reflect on what they themselves have and have not already understood in order to provide helpful suggestions to the student (Nicol et al. 2014). This is similar to peer tutoring, whereby the act of constructing explanations for the peers leads the tutors to rehearse, evaluate and hence improve their own understanding of the topic. Roscoe and Chi (2008) named this effect *reflective knowledge building*. This learning effect, hence, links back to the idea of constructivism in peer learning. Students not only construct new knowledge by receiving feedback and trying to make meaning out of it, but also by constructing knowledge for themselves through critically analyzing their peers' submissions and providing them with feedback.

A decade ago, educators, companies, organizations, policymakers and many more collaborated to develop a framework which is called the Partnership for 21st Century Skills (Trilling and Fadel 2009). This framework builds on the 21st Century Knowledge-and-Skills Rainbow containing: Life and Career Skills; Learning and Innovation Skills; Information, Media and Technology Skills; and Core Subjects and 21st Century Themes. This dissertation concentrates on the first set of skills Learning and Innovation Skills. The three main skills in this set are: *Critical thinking and problem solving* (expert thinking); *Communication and collaboration* (complex communicating); and *Creativity and innovation* (applied imagination and invention) (Trilling and Fadel 2009).

For *Critical Thinking*, students need to be able to analyze problems, and make judgements and decisions. This reflects the idea of peer learning in which students need to apply these skills to their peers' work. For *Communication and Collaboration*, students need to be able to articulate thoughts and ideas effectively. In peer learning contexts, students need to be able to formalize feedback, including making critiques and suggestions. Students need to be able to write this feedback in a way that is most helpful to their peers and that avoids putting the feedback receiver into a defensive state which may lead them to reject the comments. For *Creativity and Innovation*, students need to be able to use the learned concepts and create novel ideas. In peer learning, this is dependent on the task. In the teaching-learning scenario in this dissertation, students need

to put themselves in the position of being an entrepreneur and develop a new business idea. Furthermore, students learn to receive feedback and assimilate it so they can derive actions to improve their work.

Overall, to develop higher-order thinking skills, formative feedback is necessary to identify gaps and areas of improvement and to initiate self-regulated learning processes that enable students to reflect on possible ways of closing these gaps. The basis for formative feedback is the assessment of an internal and an external view of one's current state of knowledge or skill development.

## 2.2 Assessment for Student Performance

Assessment is the step before being able to provide feedback. Hence, assessment is conceptualized as the action of evaluating, for example the quality of text, quality of feedback or the current state of knowledge. While the result of this evaluation is usually a number, the assessment itself can be *internal* or *external* (McDonald and Boud 2003; Boud et al. 1999; Black and Wiliam 1998). It is an *internal assessment* when a student is self-assessing his or her own current state of knowledge or skill level and it is an *external assessment* when the evaluation is not carried out by the student such as by a peer or a computer instead. In this dissertation, the formative feedback provided to the students consists of a) an *external* or b) an *external* and *internal* assessment. In b) the gap between both is used to initiate self-regulated learning processes.

### 2.2.1 Self-Assessment as Internal Assessment

Self-assessment (SA) can be defined as the involvement of students in identifying standards and/or criteria to apply to their work and making judgments about the extent to which they actually meet these criteria and standards (Boud 1991). Previous research shows that reflection based on SA contributes to student performance (Nicol and Macfarlane-Dick 2006). For example, McDonald and Boud (2003) showed that training on how to carry out SA improved student performance in the final exam. According to Sadler (1989), students carrying out SA of their own work improved their learning in the course being studied, and this also provided a foundation for lifelong learning. However, SA by itself can lead to an overestimation of one's skills. Kruger and Dunning (1999) found

that overestimation often goes along with a lack of meta-cognitive and, hence, self-regulation skills among less skilled participants. Therefore, SA should be paired with objective assessment to foster awareness and understanding of how to control one's own learning and consequently develop self-regulation skills which are essential for the development of higher-order thinking skills (Nicol and Macfarlane-Dick 2006).

### **2.2.2 Computer-based Assessment as External Assessment**

Computer-based Assessment (CbA) is an automated objective evaluation of performance using algorithms or a set of rules (Higgins and Bligh 2006). In higher education, the main advantage of CbA is efficiency, as it saves the lecturers time and is less costly for the universities. Once the initial programming is completed, computers are ultimately able to provide unlimited assessments (Thelwall 2000). Furthermore, CbA and the communication of the results (feedback) allow immediate interventions in the learning process with the objective of closing the gap between the current and the pursued state of knowledge or skill level. Students are, therefore, able to adapt their learning habits in a timely manner to reach their aimed performance goal (van der Kleij et al. 2012). If designed correctly, CbA can be unbiased, non-judgemental and accurate, being able to ignore irrelevant characteristics of students (Mason and Bruning 2001). These advantages make CbA valuable for large-scale classes.

The most commonly used types of tests in CbA are multiple-choice questions (MCQ, Scalise and Gifford 2006). However, using MCQs often only allows the evaluation of lower-order thinking dimensions in the revised Bloom taxonomy (Krathwohl 2002). This suggests that CbA may have limitations when assessing higher-order thinking learning skills. This may be especially the case when the complexity of the tasks increase such as when creativity is involved such as for writing an essay or creating a business model (Scalise and Gifford 2006).

### **2.2.3 Peer Evaluation as External Assessment**

Peer evaluation is when a student judges the quality of the submission of a peer qualitatively without providing a grade. The term peer evaluation is often used interchangeably with the term peer feedback (Nicol et al. 2014). However, I use

the term peer evaluation as the task of judging the advantages and disadvantages of a peer's submission, whereas peer feedback is the task of communicating the results of this evaluation. Hence, the evaluation is the prerequisite for peer feedback.

#### **2.2.4 Peer Assessment as External Assessment**

In the literature, peer assessment is often characterized as students grading the work or performance of their peers using relevant criteria (Falchikov 2001). In this dissertation, PA is used to refer to the action of students evaluating the quality of qualitative feedback provided by their peers. Hence, PA does not refer to evaluating the submission itself.

Overall, the assessment used in this dissertation consists of an internal and an external part and an identification of the gap between both in order to initiate self-regulated learning processes. If the gap is either positive or negative, it causes so-called cognitive dissonance in the students' mind.

#### **2.2.5 Cognitive Dissonance Theory**

From an educational perspective, cognitive dissonance can be defined as an uncomfortable internal state occurring when the presented unfamiliar or contradictory information conflicts with existing knowledge (Festinger 1962). The theory of cognitive dissonance holds that individuals have a motivational drive to resolve this dissonance by either changing their beliefs, attitudes and behaviors, or rationalizing their beliefs, attitudes, and behaviors.

Presenting students with the results of their internal and external assessment could lead to such cognitive dissonance. Empirical studies of cognitive dissonance have shown that the need to resolve this dissonance is extremely motivating for students and activates cognitive processes until the dissonance is resolved (Elliot and Devine 1994). One of the key aspects of the Cognitive Dissonance Theory (CDT) is the relationship between the level of cognitive dissonance and the motivation to resolve it. Students might become quickly bored with a level of dissonance that is too easily resolved but might be frustrated with a level of dissonance that is too high (Festinger 1962). According to Piaget (1985), the dissonance is seen as an essential trigger for the learning

process, leading students to engage in problem-solving activities and construct new knowledge structures. As an added benefit to the learning process, the motivational aspects of resolving cognitive dissonance create an environment where students are continually exposed to content-relevant information, facilitating deeper processing. Thus cognitive dissonance can be produced during the delivery of feedback from the internal and external assessment.

### 2.3 Feedback on Student Performance

Assessment and especially the gap between the internal and external evaluation of students' current state of knowledge or skill level are the prerequisites of providing feedback for developing higher-order thinking skills. According to Hattie and Timperley (2007), feedback is one of the most influential factors affecting student performance, supported by the results of his synthesis of 23 meta-analyses including 1,287 studies in education. The data revealed that feedback had an average effect size<sup>3</sup> of 0.79 on student performance, whereby the mean effect size of all (e.g. frequency of testing, mentoring, individualized instructions) effects was 0.4. When considering subclass formative feedback the effect size was even 0.9, which makes formative feedback one of the top three factors influencing student performance.

One of the most cited definitions of feedback is from Hattie and Timperley (2007), who conceptualized feedback as “information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding. A teacher or parent can provide corrective information, a peer can provide an alternative strategy, a book can provide information to clarify ideas, a parent can provide encouragement, and a learner can look up the answer to evaluate the correctness of a response. Feedback thus is a “onsequence” of performance.” (Hattie and Timperley 2007, p. 81)

This definition inherits the idea that feedback should correct something which is wrong. Hence, the purpose of the feedback is to assess a state and evaluate its strengths and weaknesses once at the end of the carried out task. The feedback is not seen as something given along the learning process to incrementally improve

---

<sup>3</sup> Effect size is the importance of the magnitude of differences (Hattie 2008):

$$Effectsize = [Mean_{treatment} - Mean_{control}] / SD$$

performance and support self-reflection over time. Thus, this definition does not explicitly contain the idea that feedback can have multiple purposes such as motivation, initiation of self-regulated learning processes or provision of suggestions for improvement in the future.

The conceptualization of the purpose of feedback and how it should be provided has changed over the last two decades. Feedback is no longer seen as a one-time event but rather as a process in which students have an active role to play (Dawson et al. 2016). The shift towards feedback as a process subsequently has effects on the roles of lecturers and students. The consequence of this new conceptualization is that it has changed the driver of the feedback; providing feedback should be driven by the student rather than by the educator. Thus, feedback involves a multitude of players and necessarily involves the student making use of information to achieve change (Dawson et al. 2016).

The new definition conceptualizes feedback “as a process through which learners make sense of information from various sources and use it to enhance their work or learning strategies. This definition goes beyond notions that feedback is principally about teachers informing students about strengths, weaknesses and how to improve, and highlights the centrality of the student’s role in sense-making and using comments to improve subsequent work. Information may come from different sources (e.g. peers, teachers, friends, family members or automated computer-based systems) to support a student’s self-evaluation of progress.” (Carless and Boud 2018, .p 1315)

The objective of feedback used in this dissertation is to provide students with formative feedback to help them improve during the learning process.

### **2.3.1 Characteristics of Feedback**

Feedback is highly multidimensional and can be characterized, for example, by its content, purpose, function, timeliness, specificity and subjectivity. However, in general feedback should contain two main components: on the one hand, verification and, on the other hand, elaboration (Kulhavy and Stock 1989). Verification describes the judgment of whether the response is either “right” or “wrong”. The right/wrong decision is the first element of information encoded in feedback. The elaboration part provides information on why the response

was either correct or incorrect (Kulhavy and Stock 1989). In addition to these two components, feedback should answer three questions (Hattie and Timperley 2007): *Where am I going?*, *How am I going?*, and *Where to next?*. The first question aims at defining what the learning goals are. More precisely, it addresses the goals related to the task or performance. The judgment concerning the achievement of a learning goal may occur on many dimensions, such as directly “passing a test” or “completing an assignment”. The second question asks about the progress being made towards meeting the goal. This involves providing information in relation to a task or performance goal. Usually, this is related to prior performance and/or to success or failure in a particular task. Feedback is effective when it consists of information regarding the progress and/or how to proceed with a certain task. The third question asks about which activities need to be undertaken in order to achieve better student performance. This helps to guide students by providing advice that can lead to greater opportunities for learning. This could involve more self-reflection in the learning process, greater fluency and automaticity, deeper understanding, more strategies and processes regarding the completion of tasks or more information about what is and what is not understood.

These three questions have mainly arisen from lecturer’s perspectives. However, when peer learning is involved, answering the first question might be rather factitious, since the learning goals are set by the lecturer and not by the peer. Thus, often this question might only be dealt with when the perception of the learning goal differs substantially between the feedback provider and the creator of the submission on which feedback has been provided.

### 2.3.2 Formative Feedback

Formative feedback is a subclass of feedback and becomes important when students develop skills (Yorke 2003). Formative feedback can be defined as feedback that is provided throughout the students’ learning process, and aims to provide opportunities for improvement before the final assessment (Brown 2013). Formative feedback does not contribute to the final grade but rather provides support while students are developing and improving their skills.



In this dissertation the formative feedback was assembled in a way that addressed each domain of the revised Bloom learning taxonomy. Figure 2.2 shows the four types:

*L1* covers the Cognitive Dimensions: Remember, Understand, Apply and Analyze; and the Knowledge Dimensions: Factual, Conceptual and Procedural. The formative feedback exclusively based on the external assessment using CbA, which aims to improve for example students' ability to remember facts and concepts or to analyze them.

*L2* covers the same Cognitive Dimensions but the target Knowledge Dimension is Meta-Cognition. This type of formative feedback is comprised of SA and CbA, which aims to improve self-regulated learning by using the theoretical concept of cognitive dissonance.

*H1* covers the Cognitive Dimensions, evaluate and create, and Knowledge Dimensions, Factual, Conceptual and Procedural. The formative feedback is comprised exclusively of peer feedback, which is in contrast to formative feedback based on CbA. This type of peer feedback is a written qualitative judgment of the submission by a human. The objective is to develop higher-order thinking skills such as the students' ability to evaluate the strengths and weaknesses of business models and providing suggestions for improvement.

*H2* covers the same Cognitive Dimensions but the Knowledge Dimension is, once again, Meta-Cognition. The formative feedback is comprised of SA and PA to improve higher-order thinking skills such as the ability to provide feedback to peers. Hence, SA is the internal assessment of the student's own perception of how good he or she is in providing high quality feedback and the PA is the external assessment of how good the student's peer thinks that he or she is at providing high quality feedback. Cognitive dissonance theory is also used here as the underpinning theoretical concept.

## 2. Theoretical Background and Related Work

|                     |                          | Cognitive Process Dimension                                     |            |       |         |                                                                |        |
|---------------------|--------------------------|-----------------------------------------------------------------|------------|-------|---------|----------------------------------------------------------------|--------|
|                     |                          | Remember                                                        | Understand | Apply | Analyze | Evaluate                                                       | Create |
| Knowledge Dimension | Factual Knowledge        | Formative Feedback: CbA                                         |            |       |         | Formative Feedback: Peer Evaluation                            |        |
|                     | Conceptual Knowledge     |                                                                 |            |       |         |                                                                |        |
|                     | Procedural Knowledge     | L1 Task type: Multiple Choice Questions and Flow Diagrams       |            |       |         | H1 Task type: Open Ended Questions                             |        |
|                     | Meta-cognitive Knowledge | L2 Formative Feedback: SA + CbA<br>Theory: Cognitive Dissonance |            |       |         | H2 Formative Feedback: SA + PA<br>Theory: Cognitive Dissonance |        |

**Figure 2.2:** Linking Formative Feedback to the Revision of Bloom's Taxonomy (Krathwohl 2002). Self-assessment (SA), Computer-based Assessment (CbA), Peer Assessment (PA)

### 2.3.3 Peer Feedback

One way of engaging students actively in the feedback process is by using peer feedback (Liu and Carless 2006). In the literature, the terms 'peer feedback' and 'peer reviewing' are both used to describe the same concept (Nicol et al. 2014). In the dissertation I will use the term peer feedback since this term includes both components: Firstly, evaluating and secondly, communicating the results of the evaluation. Nicol et al. (2014) defines peer feedback as students evaluating and making judgements about the work of their peers and constructing written feedback. These students have two roles, a) being the feedback provider and b) being the feedback receiver. This leads to higher engagement amongst the students since they play an active role in the feedback process.

In role A (feedback provider), the students need to critically evaluate the quality of the submission and understand the peers' work in order to give feedback. Additionally, the students need to think about how to write the feedback in the best possible way so that it is helpful for the other person (Nicol et al. 2014). In role B (feedback receiver), the students need to learn how to assimilate and critically evaluate the received feedback in order to come up with actions for improving their own work (Cho and MacArthur 2010).

Peer feedback has the benefit of nearly unlimited scaling, especially in large-scale classes at universities or MOOCs which often consist of hundreds or

even thousands of students (Fortes and Tchanchane 2010). Overall research has shown that peer feedback is beneficial in at least three ways. Firstly, students perceive qualitative peer feedback as more valuable than lecturer's grades (Sadler and Good 2006). Secondly, students value peer feedback as more understandable and helpful than lecturer feedback, because it is written in more accessible language (Nicol et al. 2014; Falchikov 2005; Topping 1998). Thirdly, a controlled study has shown that when multiple feedback providers are involved in evaluating the quality of a submission, feedback receivers are given more diverse feedback and a higher quantity of suggestions in contrast to when the lecturer alone provides feedback (Cho et al. 2010; Topping 1998).

To leverage the potential of IT for using peer feedback, the design of *ITPF* as a teaching-learning scenarios emerged. *ITPF* builds on two fundamental concepts, namely technology-mediated learning and peer learning. Hence, *ITPF* describes the situation in which peers provide feedback to each other using technology. Managing peer feedback can be very time consuming, as this includes tasks such as the allocation of feedback providers and receivers, which increase more time with growing classes. Hence, using technology, including automated processes, to facilitate this process can help the lecturer save valuable time. Additionally, if *ITPF* is used in a way in which the feedback receiver has to use the received feedback to improve his or her own version, then this can have a direct influence on student performance. This contrasts to the situation when a lecturer provides feedback on an essay, which often takes weeks and students find it hard to use this feedback especially to apply to an essay which is often written months later (Nicol et al. 2014).

Furthermore, using technology has the advantage of providing an anonymous environment, since real names can be hidden and substituted with pseudonyms. According to Ballantyne et al. (2002) and Tahir (2012), students have difficulties in criticizing their peers, who are often their friends. Students perceive moral hazards and are afraid that their friends are not comfortable with them giving feedback. Thus, to overcome this peer pressure, research emphasises that peer feedback should be carried out anonymously (Hsia et al. 2016).

Based on the advantages of peer feedback described above, it has become the quasi standard in MOOCs, not least because MOOCs often consist of thousands of students and peer feedback is the only possible solution to scale

to large cohorts. However, MOOCs are not only large-scale but they also inclusive, leading to classes that usually consist of students with diverse cultural backgrounds. Thus, international classes of students using peer feedback consist of feedback providers with different native languages, cultures, world views and preferences regarding the giving as well as receiving of feedback (Suen 2014). Hence, another important step when designing teaching-learning scenarios using peer feedback is to be able to capture these differences.

### 2.3.4 Hofstede's Cultural Dimension Theory

Culture in the context of learning can be defined as the system of social morals, behavioral standards, symbols, world views and beliefs that provide meaning and structure to a person's life (Seel 2011). One of the most widely used frameworks to capture cultural dimensions is based on the research conducted by Hofstede and Hofstede (2005). This framework consists of five dimensions: (1) Power Distance, (2) Individualism vs. Collectivism, (3) Masculinity vs. Femininity, (4) Uncertainty Avoidance and (5) Long-Term Orientation.

*Power Distance (1)* describes the extent to which members of a society accept that power is distributed unequally. Societies with a large power distance accept hierarchical orders in which every person has a place. In countries with a low power distance people demand justifications for inequalities and strive to equalize the distribution of power within the society.

In *Individualistic (2)* cultures only loosely knit social frameworks exist and people are expected to take care only of themselves and their immediate families. This is the opposite in *Collectivist (2)* cultures, where tightly knit frameworks exist in society and people cannot only rely on their relatives but also expect members of a particular in-group to look after them.

*Masculine (3)* cultures value achievement and success. Societies in which this dimension dominates are in general more competitive. In contrast, in *Feminine (3)* cultures dominant values are cooperation, caring for others and quality of life. Societies in which this dimension is stronger are also generally more consensus oriented.

*Uncertainty (4) avoidance* describes the extent to which people of a society feel threatened by ambiguity and uncertainty. Countries with strong uncertainty

avoidance tend to hold rigid codes of belief and behavior and are intolerant of unorthodox behavior. Countries with weak uncertainty avoidance maintain a more relaxed attitude in which practice counts more than principles.

Countries with *Long-Term Orientation* (5) hold a future-orientated perspective. They encourage thrift and efforts in modern education as a way to prepare for the future. For countries with short-term orientation the short term point of view is the prominent perspective. They prefer to maintain time-honored traditions and norms while viewing societal change with suspicion.

Overall, this framework builds a foundation when designing cultural sensitive IT artifacts and hence can also be used for the design of IT-based peer feedback. However, besides the advantages of peer feedback and the consideration of cultural differences of peer feedback providers and receivers, one problem still remains. Since peer feedback is provided by students and not experts (i.e., not the lecturer), then the question arises as to whether this feedback is of high quality?

### 2.3.5 Ensuring Peer Feedback Quality

There are at least two possible approaches to ensure peer feedback quality. Firstly, automatic evaluation can be used, which includes for example the evaluation of the quality of the written feedback text using Natural Language Processing (NLP) or using voting mechanisms such as for online reviews. Secondly, calibration techniques can be used. In this case, the student's skill of providing feedback is calibrated to the skill of an expert such as a lecturer.

#### 2.3.5.1 Automatic Evaluation

*Automated Essay Scoring (AES)* follows the idea of automatically assessing texts, although the dependent variable is not the helpfulness of the text but rather the grade of an essay. AES automatically assesses the quality of an essay using text mining algorithms. Attali and Burstein (2006) built a tool called e-rater which automatically assesses structure, grammar, lexical complexity and vocabulary usage of essays. For example, the structure feature comprises the existence of an introductory paragraph, at least three paragraphs in the main body and a conclusion paragraph. Latest research on AES ranges from

using feature-based machine learning to deep learning and neuronal networks. However, as a consequence of using deep learning, it is unknown which exact features result in high or low essay grades (Farag et al. 2018). This is the reason why in our research we rely on feature-based machine learning and not on neuronal networks. Furthermore, AES provides scores using rubrics instead of “rich and meaningful” feedback. This is problematic, since scores and standardized rubrics often do not help students to derive actions for improving their skills (Hattie and Timperley 2007; Black and Wiliam 1998).

*Online reviews* are “a type of product information created by users based on personal usage experience which can serve as [information] to help consumers identify the products that best match their idiosyncratic usage conditions” (Chen and Xie 2008, p. 1). Review helpfulness is defined as the extent to which the customer perceives that a review has helped them to make the right decision in the purchasing process (Mudambi and Schuff 2010). Voting systems and filter options are two systems that are used nowadays to identify helpful reviews (Cao et al. 2011). Either the voting system asks the reader to rate a review as helpful or not helpful or it simply works with up- and down-votes. Different filter options are available to show, for example, the most recent or most helpful reviews (i.e., reviews with the most positive votes).

Many researchers of online review helpfulness differentiate between two product types a) search goods and b) experience goods. Search goods can be generally described as products which have characteristics that can quite easily be evaluated before purchasing them (Luan et al. 2016). Experience goods can be generally described as products which have characteristics that are quite difficult to evaluate prior to buying them (Baek et al. 2013). It is very important to differentiate between the two types of products because the product type influences the effect of some factors on perceive helpfulness.

Besides the differentiation between product types, it is important to understand the concept of the negativity bias. This concept deals with the psychological effect of negative versus positive information. According to the negativity bias, the effect of negative information is much stronger than the effect of positive information. A great example of the negativity bias is the damage to Volkswagen’s reputation after the emissions scandal in 2015 (Neat 2015). Negative information outweighs the positive information, because negative

information reveals more about the personality and character of someone than positive information does (Wu 2013).

#### 2.3.5.2 Calibration Techniques

Calibration Techniques such as Calibrated Peer Reviewing (CPR) are often used in massive open online courses, which may contain thousands of students working in a peer learning setting, providing feedback to each other (Balfour 2013). In CPR a new student has to grade an essay which the lecturer has already graded. The two grades will be compared and in cases where the student's grade is significantly different to the lecturer's grade, the student will need to grade further essays until his or her grade converges to the lecturer grade.

However, applying CPR requires a lot of effort from the lecturer, since the idea of the approach is to calibrate the peer feedback based on the feedback the lecturer would provide. Thus, this characteristic does not help deal with the increasing scarcity of teaching resources at most universities. Furthermore, just as for AES, CPR often relies on pre-defined scoring schemes to provide feedback (Balfour 2013). Thus, learners receive scores rather than rich qualitative feedback, which is not ideal as the latter has been found to have a more beneficial effect on skill development (Hattie and Timperley 2007; Black and Wiliam 1998).





# 3

## **System Design and Evaluation**

## 3.1 Introduction

In this chapter<sup>1</sup> I will present the first stage of developing a design theory for designing Formative Feedback Tools (FFTs). To show the applicability of the design, I created an instance based on the class of FFTs. This instance (IT artifact) is evaluated in a field experiment and called Learning Objective and Outcome Manager (LOOM). Thus, the term FFTs represents the general class of formative feedback tools while LOOM represents a instantiation of the class of FFTs.

FFTs should provide lecturers with a structured way to design their teaching-learning scenarios and create assessments with which to evaluate whether their students meet the defined learning outcomes and objectives. Furthermore, FFTs should provide lecturers with an overview of the students' performance on different levels of aggregation. This provides valuable insights such as which learning outcomes have been understood well and which might need further clarification.

For students, FFTs should provide them with the possibility to assess their own, current state of knowledge during the learning process, before they take part in the final assessment such as an exam. Hence, this should enable students to adjust their actions in a timely manner. Furthermore, through an assessment of their internal and external perception of their current state of knowledge, students are able to develop self-regulated learning skills.

In this dissertation, formative feedback is composed of four different types. Figure 2.2 in Section 2.3.2 shows these four types: *L1*, *L2*, *H1* and *H2*. *L* and *H* are related to the cognitive dimensions. *L* represents the lower-order thinking skills and *H* the higher-order thinking skills. The numbers *1* and *2* are related to the knowledge dimensions. Number *1* represents the factual, the conceptual and the procedural knowledge dimensions and number *2* represents the meta-cognitive knowledge dimension. The formative feedback used in this chapter is based on *L1* and *L2*.

---

<sup>1</sup> A previous version of this chapter (Sections 3.2 to 3.5) has been published in Rietsche et al. (2018). The content has been reformatted and updated to the latest state of scientific discussion at the time of submission of this dissertation.

Overall, the aim of this chapter is to answer research question one and two of this dissertation:

*RQ1: What are the design principles for designing an IT artifact to provide students technology-mediated formative feedback?*

*RQ2: What are the effects of technology-mediated formative feedback on student performance?*

To answer research question one, I use Design Science Research (DSR) and in particular follow the three-cycle DSR approach developed by Hevner (2007). To document the design knowledge which emerges during the design process of the IT artifact, I use the framework for design knowledge developed by Gregor and Jones (2007). To answer research question two, which requires measuring the effects of formative feedback on student performance, the instantiation (LOOM) of the design knowledge is tested in a one-group field test with 194 students at Master's level.

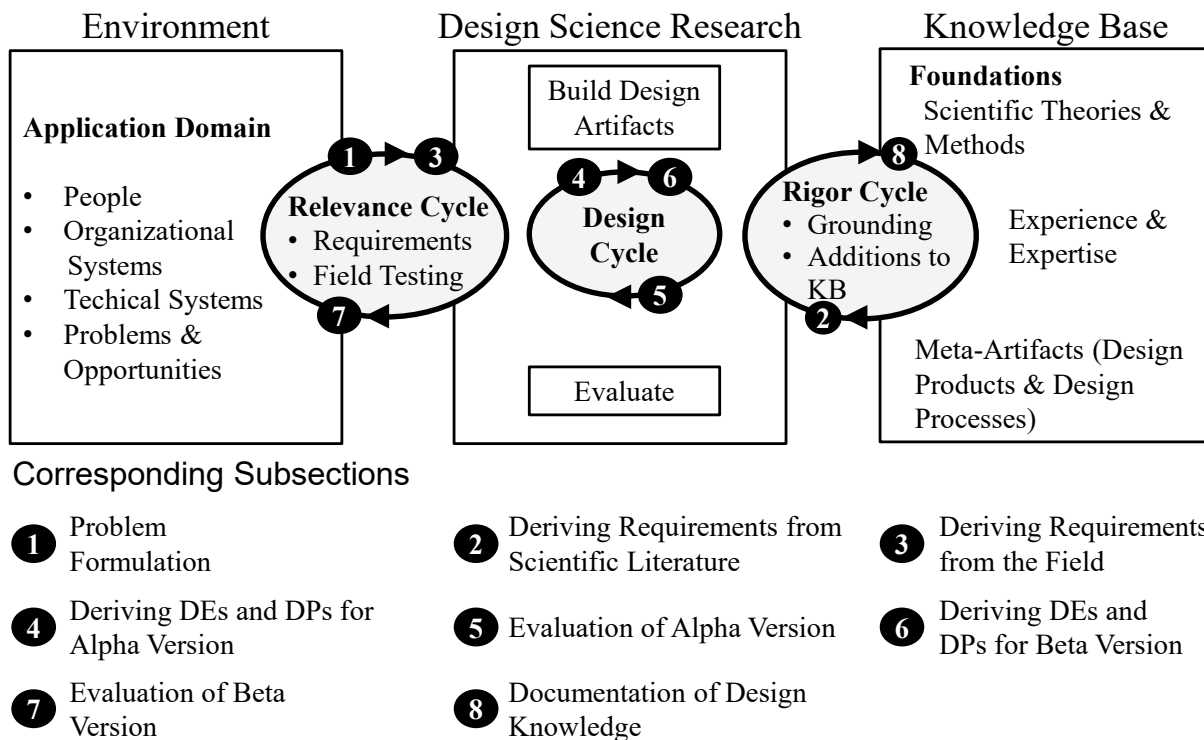
The remainder of this chapter is structured as follows. In the next section I describe the research methodology and how I follow the DSR process to design FFTs. In particular, it addresses how I derive design principles and features for FFTs from scientific literature and expert interviews. Furthermore, I explain how I evaluate the alpha version of the instance LOOM and derive additional design principles and features for the beta version. I follow the structure of Gregor and Hevner (2013) and separate the evaluation and discussion of this process into two distinct sections. The evaluation section shows in detail the outcomes of the one-group field test used to evaluate the beta version. The following section discusses the results, presents a documentation of the design knowledge in accordance with Gregor and Jones (2007), discusses limitations of the study and, finally, provides an outlook for future research. The final section provides a conclusion to this chapter.

## 3.2 Research Methodology

To achieve my research goal, I used the three cycle DSR approach of Hevner (2007). I followed this approach because I wanted to a) use a scientific method to solve a set of practical problems that researchers and practitioners experience

in their own practice and b) contribute to the existing body of knowledge by designing and evaluating a new research artifact and documenting the design knowledge according to Gregor and Jones (2007). Moreover, to ensure that FFTs addressed all important aspects to increase student performance, I followed a theory-driven design approach by grounding my research on the formative feedback theory (Black and Wiliam 2009) and CDT (Festinger 1962). In this chapter, I will present the details of all phases advocated by Hevner (2007) for the development of FFTs.

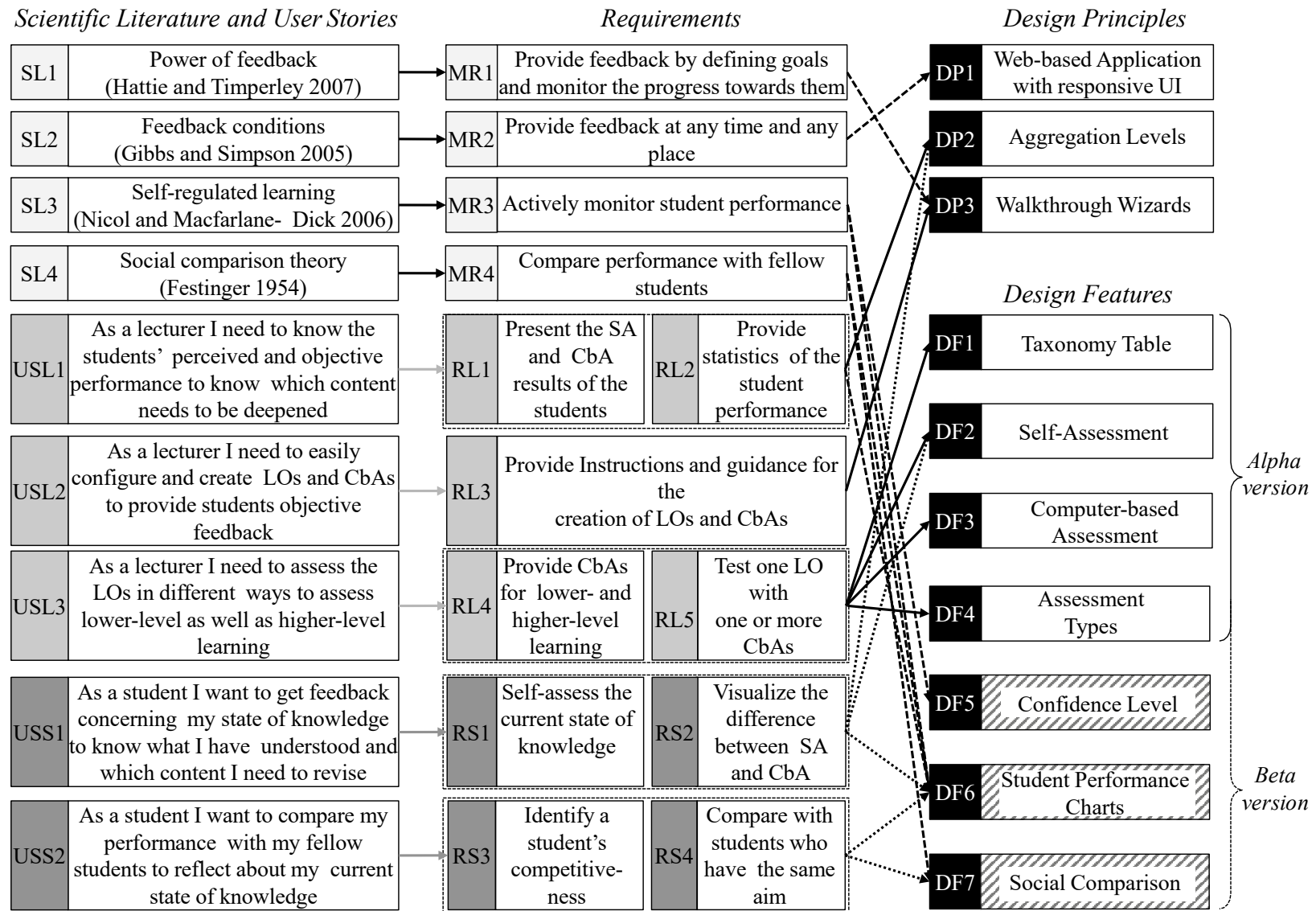
Figure 3.1 shows the steps that have been carried out as well as their order and the corresponding subsections which contain detailed information. The first step includes the problem formulation. The practical driven motivation of the problem was discussed in detail in the introduction section. Based on the problem definition, in the second and third step, I derived user stories and requirements for the design of FFTs from a) scientific literature and b) six semi-structured interviews with lecturers and students. The length of the interviews was timed between 40 and 60 minutes. In the fourth step, I derived design features based on the requirements for the alpha version and implemented them. In the fifth step the evaluation of the alpha version was conducted. In the sixth step, the design features of step four had been revised based on the evaluation results of step five. The outcome of this step was the beta version. In the seventh step, the beta version was evaluated in a one-group field test with 194 students of a Master course in Management Information Systems (IS). In the eighth step, the design knowledge is documented according to the eight components for documenting a design theory which was developed by Gregor and Jones (2007). To improve readability, I discuss step seven and eight in separate sections. Hence, step seven “evaluation of the beta-version” can be found in Section 3.4 and step eight “documentation of design knowledge” in Section 3.5 Discussion. For the evaluation of the FFT design knowledge I created an instance called LOOM. I followed the evaluation framework proposed by Venable et al. (2016). This framework provides the advantage that, it describes a systematic approach of evaluating each step of the artifact design process. I performed an artificial ex ante evaluation of the design features of the alpha version and a naturalistic ex post evaluation of the design features of the beta version.



**Figure 3.1:** Three Cycle Design Science Research Process by (Hevner 2007)

### 3.3 Designing and Evaluating Formative Feedback Tools

Based on the problem formulation (step one) in the Introduction Section 3.1, this section discusses how I derived the (meta-)requirements as well as the Design Principles (DPs) and Design Features (DFs), relevant for the development of formative feedback tools. All insights I gathered, are summarized in Figure 3.2. The basic structure of FFTs include different levels such as courses, Learning Units (LUs), Learning Outcomes (LOs) and CbA. Every course in FFTs can have multiple LUs. A LU normally represents one lesson and the topics covered in it. In turn, every LU can have multiple LOs. A LO defines an objective pursued by the students within a LU and is usually formulated in one sentence. Every LO can involve multiple CbAs checking the knowledge of students. However, a CbA is not necessarily assigned exclusively to one LO. It can also be integrated in other LOs which may belong to different LUs or courses. The lecturers weight CbAs when assigning them to a LO by allocating a number between 10 and 100 percent.



**Figure 3.2:** Overview of the Design Principles and the Features Derived from the Defined (Meta-)Requirements

### **3.3.1 Step 2: Deriving Meta-Requirements from Scientific Literature**

In this section, I gather Meta-Requirements (MRs) from scientific literature for the fundamental development of FFTs (Gregor and Jones 2007). MR1 deals with the three questions a feedback should answer (see Section 2.3.1) developed by Hattie and Timperley (2007). The MR2 deduced from scientific literature deals with the observation that if students receive feedback often and regularly, it promotes better monitoring and self-regulation of their learning progress (Bitzer et al. 2016; Gibbs and Simpson 2005). Hence, FFTs should enable students to get feedback at any time and any place. Moreover, students need the possibility to actively monitor their performance over time (MR3). FFTs should provide an overview of the student performance aggregated at LO, LU and course level. MR4 addresses the possibility of students to compare their performance with the performance of fellow students to resolve uncertainties about their own abilities (Festinger 1954). However, FFTs should provide the possibility to disable social comparison mechanisms if a student's demand for competitiveness is low. According to Nicol and Macfarlane-Dick (2006), feedback on the student performance leads to an improved awareness.

### **3.3.2 Step 3: Deriving Requirements from Expert Interviews**

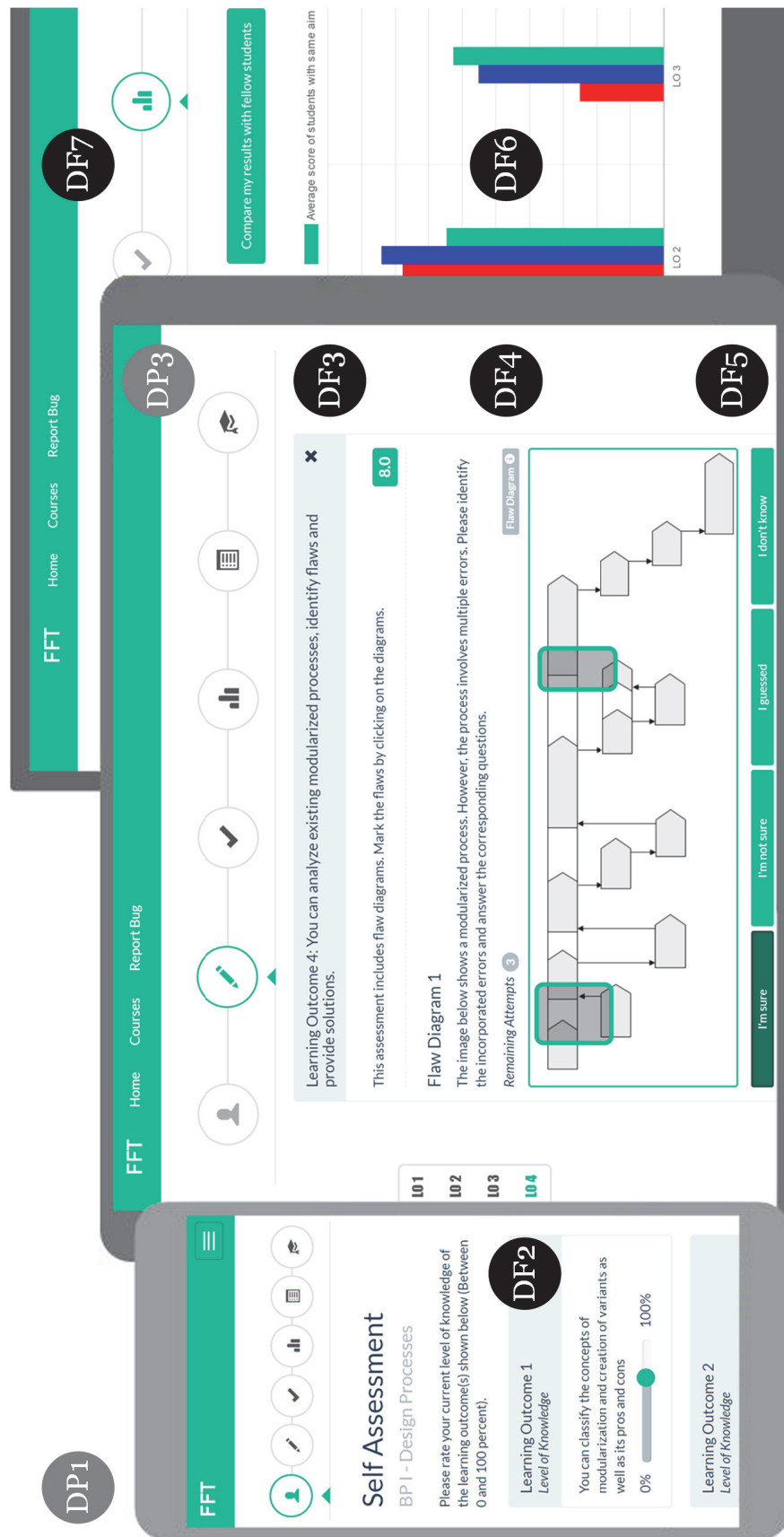
After defining a set of meta-requirements, I identified user stories and requirements originating from specific field problems of lecturers and students. The first user story of lecturers (USL1) reflects the lack of information of the student performance during the semester. Lecturers have the requirement that FFTs should provide statistical information of the student performance throughout the semester (RL1 and RL2). The second user story (USL2) represents the fact that FFTs should be easy to use and self-explanatory. Hence, FFTs should provide wizards and instructions to minimize the time for adopting the tool (RL3). The third user story (USL3) addresses the fact that lecturers want to develop the students' action competences. Therefore, FFTs should provide the possibility to create LOs and assign multiple weighted CbAs (RL5) for lower-level as well as higher-level learning (RL4). The first user story of students (USS1) represents their need to know what their current state of knowledge is, and which content needs to be deepened. Hence, FFTs should provide the

possibility to carry out SA (RS1) and CbA to the LOs and graphically present the student performance as well as differences between SA and CbA (RS2). The second user story (USS2) indicates the demand of students to compare their student performance with the student performance of fellow students. FFTs should provide the possibility to compare the results of individual students to the class average (RS4) based on their competitiveness (RS3).

#### **3.3.3 Step 4: Deriving Design Principles and Features for the Alpha Version**

The first Design Principle (DP) is overarching and must be considered in the implementation of every other DP and Design Feature (DF). It defines that the emerging tool should be a web-based application with a responsive user interface in order to be available for lecturers and students from anywhere at any time independent of the operating system in use (DP1). To meet this requirement, FFTs should be implemented as a web-based application accessible via browser. The front-end of FFTs should be developed with the latest web technologies including Hypertext Markup Language version 5, Cascading Style Sheets and JavaScript as well as the JavaScript library jQuery. Moreover, students should be able to efficiently use FFTs from any web-enabled device regardless of its screen size. To achieve a fully responsive user interface, front-end web framework Bootstrap 3 which offers a fluid grid system was used. As a result, users can access FFTs with different devices such as notebooks, tablets or smartphones running on different operating systems such as Windows, macOS, iOS or Android. Figure 3.3 shows the front-end of the instance LOOM on a notebook, tablet and smartphone. Students face this view when conducting CbAs to a LU.





**Figure 3.3:** LOOM's Student View (Multiple Devices) Including Selected Design Principles and Features

FFTs present assessment results at LO, LU and course level. This leads to the benefit that lecturers can analyze student performance at different aggregation levels (DP2) to evaluate if certain content was understood or needs to be repeated. Students can do the same individually for themselves and simultaneously compare their performance to the class average in order to disclose potential knowledge gaps.

For the user roles lecturer and student, walkthrough wizards (DP3) in FFTs were incorporated. FFTs automatically starts a wizard when lecturers create a new LO. This wizard aims to introduce lecturers to the structure of FFTs offering them a simple way to effectively use the tool (Rietsche et al. 2017b). It guides the lecturers from the creation of the LO, through the assignment of CbAs, up to the creation of new or the modification of existing CbAs. These three visualized steps building on one another provide a simple way to use FFTs and help lecturers to better understand and manage the relationship between LOs and CbAs. For students, FFTs assessment process involves multiple steps. To maintain a clear structure, to allow for an intuitive use of FFTs and to reduce the cognitive load of students, a wizard guides the students through the steps of the process. Specifically, the process consists of SA, CbA, feedback, social comparison, review and completion.

Formative feedback plus CbA can address learning objectives on various levels. Bloom's Taxonomy (Bloom et al. 1984) provides hierarchical models which allow the categorization of learning objectives according their complexity. In the cognitive domain, the Taxonomy defines six major categories of learning. These categories are ordered from easy to difficult while the expertise of each subjacent category is essential for the understanding of the next higher one (Krathwohl 2002). In the revision of Bloom's Taxonomy, Krathwohl (2002) re-defined these six categories of learning and show how they intersect with distinct levels of knowledge. Therefore, the revised Taxonomy is a two-dimensional framework. The two dimensions "Cognitive Process" and "Knowledge" and their intersections can be visualized within the Taxonomy Table see Figure 3.4 (also see Section 2.1.4). The Taxonomy Table allows the classification of assessments and the goal-oriented provision of feedback. Moreover, it reveals untapped potential in the way knowledge is transferred.

Taxonomy Table 1

show/hide assistant

| Knowledge Dimension     | Remember | Understand | Apply | Analyze | Evaluate | Create |
|-------------------------|----------|------------|-------|---------|----------|--------|
| Factual Knowledge       |          |            |       | ✓       | ✓        |        |
| Conceptual Knowledge    |          |            |       | ✓       | ✓        |        |
| Procedural Knowledge    |          |            |       | ✓       | ✓        | ✗      |
| Metacognitive Knowledge |          |            |       | ✓       | ✓        | ✗      |

Assigned Assessments 2

Status

Computer-based Assessment Name

Computer-based Assessment Type

Weighting [%]

Delete

☒

BPMN Model

Flaw Diagram

100

Add Existing Cba

+ Add New Computer based Assessment

**Figure 3.4:** Taxonomy Table During the Process of Assigning a Computer-based Assessment to a Learning Outcome

When creating a LO, lecturers use the Taxonomy Table (DF1) to define which objectives they want to achieve. For lecturers unfamiliar with the revised Taxonomy, FFTs offers an assistant that determines the right placements in the table by posing a few cumulative questions. In the process of assigning CbAs to a LO, the lecturers then continuously see in the Taxonomy Table if the selected CbAs correspond with the defined objectives. If the user tries to create a LO without meeting all the objectives, a warning is being displayed. The highlighted area in the Taxonomy Table (see Figure 3.4) represents the learning objectives the lecturer wants to achieve with the LO that is being created. The black check marks in the same table indicate which learning objectives are actually achieved with the assigned CbAs (see Figure 3.4). If there are pursued objectives that remain uncovered by the assigned CbAs, FFTs warn the lecturer by filling the affected cells with an orange cross. The classification of objectives, which a CbA is covering, depends to a certain degree on the question type included in the CbA. In the context of LOOM, primarily MCQs were used to address the cognitive process dimensions “remember” and “understand” and, thus, to target lower order thinking (Attali and van der Kleij 2017).

In flaw diagrams, items of type “Substitution/Correction” (Scalise and Gifford 2006), students are asked to find and revise errors. According to Krathwohl

(2002) these cognitive processes mainly take place within the dimensions “analyze” and “evaluate” and target higher order thinking. Those classifications are set as default when adding a new assessment. However, the specific classification within the Taxonomy Table can vary from CbA to CbA even if they are of the same type. Therefore, lecturers can adjust the default classification of individual CbAs based on the characteristics of the involved questions. In FFTs, feedback consisting of SA and CbA is based on the LUs which are covered throughout the semester. That way, students have enough time to use the formative feedback provided by FFTs to improve themselves before the final exam takes place. SA (DF2), the first step of the assessment process, forces students to assess themselves in terms of their level of knowledge. One slider for each LO allows the students to define their current understanding of the corresponding learned content within a range from 0 to 100 percent. After the students assessed themselves, they conduct the CbAs (DF3). In the student view, all CbAs assigned to the LOs of a LU are presented to the user collectively. In other words, FFTs show all the separately created CbAs within a LU in a single page to the students. The students then can choose their own way through the CbAs. The CbAs consist either of the assessment type (DF4) FD or MCQ. When creating a MCQ, lecturers enter a question and offer multiple response options for the students to choose from.

Depending on the form of the question selected by the lecturers, either one or multiple response options can be correct. A FD basically consists of an image with incorporated errors. Students need to find these errors within a limited number of attempts and answer associated questions. In order to enable this, lecturers upload and crop flawed images and then mark the areas containing the flaws. To each of those flaws, the lecturer can add an MCQ or free text question.

#### **3.3.4 Step 5: Evaluation of Alpha Version**

The first evaluation of the instance LOOM was carried out ex ante using an artificial evaluation setup (Venable et al. 2016). The purpose of this evaluation was to measure the completeness of the DPs and DFs. To reach this goal, I conducted a focus group workshop with four lecturers and thirteen students. In the workshop, the participants tested the user stories and created feedback reports based on the criteria a) logic, b) usability, c) usefulness and d) efficiency.

The evaluation of the workshop results led to revisions of LOOM's requirements. I investigated the click paths of the students. The results showed that the students often switched between multiple possible answers. Hence, a further requirement was to offer students the possibility to provide feedback in terms of how confident they were while answering the assessment questions. Further, the results showed that the usability needs to be improved.

In the alpha version, the student performance was presented as a line diagram with a consistent line between the first and the last LU. This visualization was irritating. Additionally, the students and lecturers only received the results at LO level. The requirement emerged to revise the visualization of the results and show the student performance at different aggregation levels. During the workshop, I observed that students shared their results amongst each other in order to evaluate their performance. Based on the discussion with the students, a further requirement was requested. The students should be able to compare their performance with the performance of their fellow students. In the following section, the DFs derived from these requirements will be discussed in more detail.

#### **3.3.5 Step 6: Deriving Design Features for the Beta Version**

Based on the results of step five, the DFs of FFTs and consequently the instance LOOM were revised and extended. In Figure 3.2, the DFs added for the beta version are highlighted in grey. When answering the items of a CbA, LOOM requests the students to select their confidence level (DF5). The levels consist of "I'm sure", "I'm not sure", "I guessed" and "I don't know". The monitoring of the students' confidence levels over time, may help students to improve their performance (Nicol and Macfarlane-Dick 2006). Neither answering the questions within the CbAs nor stating the confidence level is mandatory for students. However, if students answer a question, a confidence level needs to be selected as well to continue the assessment process.

The results of the SA and CbAs are presented to students on an individual basis and to lecturers in aggregated form within student performance charts (DF6). LOOM uses individual student results of the SA and CbAs aggregated either at LO or LU level to generate a bar chart in the students' overview. One bar

represents the SA results and the other represents the aggregated CbA results expressed as a percentage of the maximum points. Consequently, the y-axis of the chart has a range from 0 to 100 percent and the x-axis displays, depending on the aggregation level, all LOs of a LU or all LUs of a course. While LOOM visualizes the individual results for students, it summarizes the results of all students of a course for lecturers. Lecturers may observe a chart showing the average SA and CbA results of their class expressed in percentage and aggregated at LO or LU level. However, lecturers do not have access to the results of individual students.

In the context of social comparison (DF7), students can compare their results with fellow students who resemble themselves in terms of competitiveness. More specifically, they can activate a third bar within the chart that reveals the average CbA score of students with the same performance goal. Festinger (1954) claimed, that people possess a drive to evaluate their opinions and abilities. There are two types of comparisons: upward and downward comparison. Upward comparison can be defined as the unidirectional drive upward, meaning that people prefer to compare themselves to others whose performance or abilities are slightly better. The definition of downward comparison, according to Wills (1981), is that people enhance their own subjective well-being by comparing themselves with less fortunate others. However, past research in the academic sector shows that upward comparison leads to better results than downward comparison. Blanton et al. (1999) observed in their experiment that students who compared themselves with well performing students increased their own academic performance the most during semester.

## 3.4 Evaluation of Beta Version

This section describes the naturalistic ex post evaluation of the beta version (Venable et al. 2016). I tested the beta version of LOOM which is an instance of FFTs in a one-group field test to answer the second research question: Does the usage of LOOM increase student performance. Therefore, I wanted to investigate whether the students who used LOOM more often performed better in the final exam. Before discussing the evaluation of beta version in detail, I

will briefly present how the artifact is embedded in the didactic concept of the teaching-learning scenario used in this dissertation.

The lecture was a mandatory course for all Management Information Systems students. Each week the students had a traditional lecture called LU. Each LU consisted of a pre-, during- and post-phase. The artifact could have been used in each of these phases. In the pre phase, the artifact could have been used in a flipped classroom concept as pre entry assessment. Hence, the students would have to carry out the assessment before the class is taking place. The aim would have been to ensure that each student is well prepared for the discussion. In the during phase, the artifact could have been used in the last 15 minutes of the lecture. The aim would have been to deepen the newly learned materials. In the post phase, the artifact could have been used after the lecture took place. The aim would have been to deepen the learned materials to transfer the knowledge to the long-term memory. In the lecture, the artifact was used four times in the during phase and four times in the post phase. The during phase was 15 minutes at the end of the LU and the post phase was the time between the end of the LU and the begin of the next LU, usually 7 days.

### **3.4.1 Data Collection**

A total of 189 students participated in the one-group field test, used LOOM at least one time, and participated in the final exam (see Table 3.1 Dataset 1). To control for mastery approach and self-efficacy and measure perceived student performance the N decreased to 92 students (see Table 3.1 Dataset 2). The N in Dataset 2 decreased due to not all students filled out the pre and post questionnaire. For the participation in the study I raffled coupons for the merchandising shop of my university.

The design of the field test was as follows: Before LOOM was used the first time, the students carried out a pre-questionnaire to measure their mastery approach and self-efficacy. Each student had a personalized token to connect the data in FFT, exam and the survey questions. In each LU the students followed a given four-step process. In the first step, the students carried out the SA, hence rating their current state of knowledge for each LO of the current LU. In the second step, the students started the CbA and answered the MCQs and flaw diagrams. In the

**Table 3.1:** Descriptive Statistics of Evaluating the IT artifact (Every Student in Dataset 2 is Part of Dataset 1)

|             | <b>Dataset 1: Objective Student Performance</b> |            | <b>Dataset 2: Objective and Perceived Student Performance</b> |            |
|-------------|-------------------------------------------------|------------|---------------------------------------------------------------|------------|
| Sample Size | N=189                                           |            | N=92                                                          |            |
| Male        | 69.31% (n=131)                                  |            | 73.91% (n=68)                                                 |            |
| Female      | 30.69% (n=58)                                   |            | 26.09% (n=24)                                                 |            |
|             | Mean                                            | SD         | Mean                                                          | SD         |
| Age         | 25.90 years                                     | 1.92 years | 25.92 years                                                   | 1.65 years |

third step the students received feedback about which questions were answered correctly and which were not, as well as how many points they received for their answer. In the fourth step, the feedback was presented to the students in the form of a bar chart. The students saw on the one hand the percentage of the SA and on the other hand the percentage of the CbA. After completing the last LU, the students carried out a post-questionnaire to measure their perceived student performance. In the final step, the students wrote the final exam of the lecture. The final exam consisted of retention (MCQs) and transfer tasks (e.g. applying learned methods) (Shute 2008).

### 3.4.2 Data Preparation

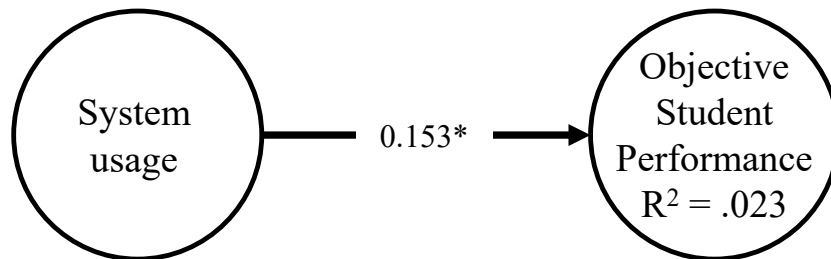
Before analyzing the datasets, I brought the survey responses into the right data structure. In R, I cleaned the data originating from the surveys by removing entries of students who did not complete the questionnaire. To prevent straight lining from affecting the survey results, I excluded students who selected the same response to different statements more than 20 times in a row (Millsap and Maydeu-Olivares 2009). Also, columns involving no student data at all were excluded from the data sets. Regarding the data originating from LOOM, I applied the boxplot method to identify and exclude outliers in terms of the usage duration at LU level. Thereby, every duration that exceeds 1.5 times the interquartile range (the middle 50% of the data) above the upper and below the lower quartile is considered an outlier (Backhaus et al. 2015). Outliers could occur, when students left the computer for hours while answering the questions



or did not finish the assessment. After removing the outliers, I enriched FFT's data with the total usage duration and the average usage duration of each student.

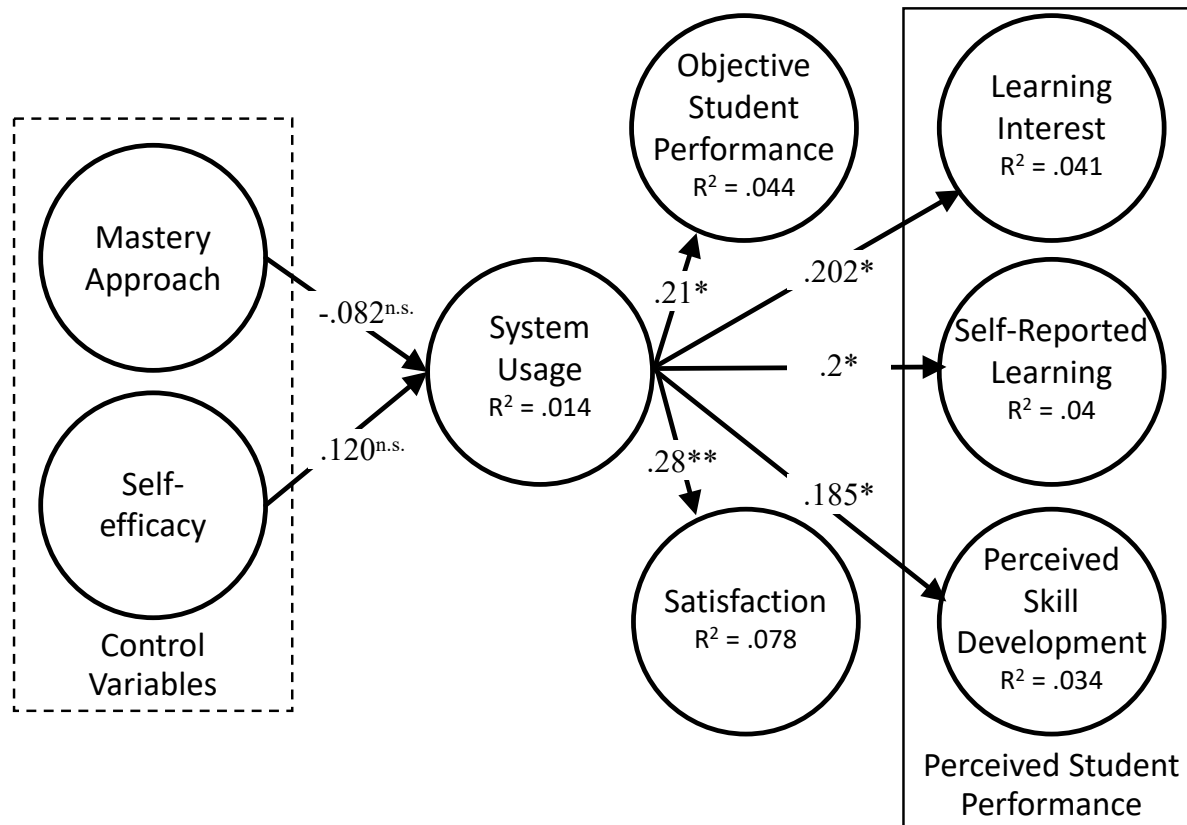
### 3.4.3 Measurement of Constructs

To measure whether the usage of LOOM increases student performance, I describe the operationalization of the measurement constructs (see research model in Figure 3.5 and 3.6). The independent variable which is the construct *FFT usage* can be measured according Straub et al. (1995) in different ways. I decided to measure this latent construct with two items. Firstly, frequency of use, which in my model is measured in terms of how many assessments the student has carried out (Ginzberg 1981). The second item is duration, which is in my case measured in terms of the timespan between starting with the first assessment task and finishing the last task (Srinivasan 1985). I decided to use these items based on the fact, that both are objective measurements and can be calculated based on database queries. The dependent variable is the construct *objective student performance*, which is measured as a manifest variable by evaluating how many points the students achieved in the final exam. The range is between 0 and 40 points. I decided to use points rather than grades, since points represent the student performance in higher precision.



**Figure 3.5:** Results Research Model A Using Dataset 1

(\*\*\* =  $p < .001$ , \*\* =  $p < .01$ , \* =  $p < .05$ , *n.s.* = *not significant*)



**Figure 3.6:** Results Research Model B Using Dataset 2

(\*\*\* =  $p < .001$ , \*\* =  $p < .01$ , \* =  $p < .05$ , *n.s.* = *not significant*)

One of the goals in education is the development of critical thinking and problem solving skills (Trilling and Fadel 2009). Hence, I argue that using LOOM should increase the students perceived skill development. In addition to the skill development, it is important that the students perceived that new content was learned. Hence, the learning of new factual knowledge and identifying concepts. Empirical studies found out, that there is a positive connection between student interest and student performance (Schiefele 1991). If students think about the topics, read further information or talk with fellow students, they strengthen the learned knowledge.

Therefore, I argue that student performance not only consists of objective student performance, but rather of the combination of objective and perceived student performance (Söllner et al. 2017). To measure the dependent latent variable, which is the construct *perceived student performance*, I used the widely known and often used measurements by Alavi (1994): Perceived Skill Development (PSD), Self-Reported Learning (SRL) and Learning Interest (LI). An example

item for PSD is: I did additional reading beyond the required assignments; for SRL: I increased my ability to critically analyze issues; and for LI: I gained a good understanding of the basic concepts of the material.

According to Arbaugh (2001), *satisfaction* plays additionally an important role for student performance, such as if the students are satisfied using LOOM and would recommend it to fellow students (example item: I am very satisfied with the course). Additionally, I checked for the students' Mastery Approach (MA) and Self-Efficacy (SE). An example item for MA is: I enjoy challenging and difficult tasks at work where I'll learn new skills; and for SE: I am sure that I can solve the exercises and tasks of the course very well. To measure the students' MA I used the construct developed by Baranik et al. (2007) and for SE the construct of Hardin et al. (2014).

I measured all latent construct variables with reflective indicators. For this purpose, I evaluated the measurement instrument with regards to its suitability to measure the constructs in a reflective manner. This was done by checking the reflective constructs according to the guidelines of Jarvis et al. (2003). I used a 7-point Likert response format that ranged from 1 ("strongly disagree") on the left to 7 ("strongly agree") on the right with 4 as a neutral point to assess the indicators.

#### 3.4.4 Data Analysis

To analyze if LOOM usage increases student performance, I used a quantitative statistical method named Structure Equation Model (SEM). This approach suits my research for three reasons (Backhaus et al. 2015). First, the nature of my research is rather confirmatory than exploratory. Second, the research model involves independent, dependent and control constructs and, thus, is rather complex. Third, the constructs of the model are latent, meaning they are not tangible or directly measurable.

As software for running the SEMs I used SmartPLS version 3.2.7 developed by Ringle et al. (2015). I used Partial Least Squares (PLS) modeling instead of covariance-based modeling, because the objective of my research is not to evaluate a rigorous theoretical model, but rather to assess the effects of FFTs usage on the objective and perceived student performance (Mertens et al. 2017).

Missing values in SmartPLS were handled by mean replacement. I ran the PLS algorithm with weighting scheme path and 300 iterations (anything else left default). The settings for the bootstrapping algorithm were 5000 subsamples and individual sign changes (anything else left default).

I measured all common quality criteria for the research model A and B (Backhaus et al. 2015). The factor loadings of all constructs exceeded 0.6. Cronbach's alpha for all constructs is above 0.8, and the Average Variance Extracted (AVE) is for all constructs is higher than 0.62. The composite reliability is for all constructs exceeds 0.83, and the Variance Inflation Factor (VIF) values are for all constructs is below 4.2. Hence, my research model fulfills all necessary quality criteria (Hair et al. 2014).

Figure 3.5 and Figure 3.6 show the results achieved by running the PLS and Bootstrapping algorithms. I ran each algorithm for both datasets. Research model A with dataset 1 (Figure 3.5) includes 189 students of the 194 students who wrote the final exam. I included this dataset to show that students who used LOOM performed in general significantly better in the final exam. As stated in the theoretical background, to measure student performance holistically, objective and perceived student performance should be combined.

Research model B (Figure 3.6) shows that FFT usage has a significant effect on both, objective and perceived student performance consisting of learning interest, self-reported learning, and perceived skill development and satisfaction. I controlled for mastery approach and self-efficacy of the students. Both constructs had no significant influence on LOOM usage.

I am aware that the  $R^2$  values of the dependent variables are not too high. This indicates that FFT usage alone does not explain a high magnitude of variation in the dependent variables. This is often the case for complex measurements. However, I want to stress that the goal of my evaluation was to assess whether FFT usage has a significant effect on important outcomes. In my case different measures of student performance as well as satisfaction.

### 3.5 Discussion and Contributions

To communicate the theoretical contribution, I document the design knowledge according to the eight components of a design theory by Gregor and Jones (2007) which are shown in Table 3.2. This systematic approach helps to structure the design knowledge and learning of my project to make a first step towards developing a design theory for FFTs. Following Gregory and Muntermann (2014), I suggest that systematically developed design knowledge satisfies the criteria for scientific knowledge and provides a legitimate academic contribution.

The type of theory I contribute to, is a “design and action” theory (Gregor and Jones 2007). The purpose of the theory is to provide prescriptions of principles of form and function for constructing artifacts. Hence, the description implies that following this recipe, the outcome is an instantiation of the class of FFTs (Gregor and Jones 2007). Furthermore, according to Walls et al. (1992), a design theory can never be purely explanatory. Hence, my design theory should also predict. With my field test I was able to show that following the recipe for the design of the artifact, through the usage of the artifact the a) objective student performance and b) perceived student performance significantly ( $p < 0.05$ ) increase.

The *purpose* of a design theory (Point 1 in Table 3.2) is to give explicit prescriptions about the requirements for the development of FFTs. As outlined above, formative feedback building SA and CbA is one of the most influential factors for student performance. Through formative feedback, the students are able to adapt their learning process before the final exam. However, to my best knowledge, there is no study that has formalized design knowledge of FFTs that combines SA and CbA. To close this gap, the goal of the first step is to contribute to a design theory, by explicitly formulating prescriptions for a class of FFTs that provide formative feedback using SA and CbA (see DSR process step one in Figure 3.1).

*Constructs* are representations of the entities of interest in a design theory (Point 2 in Table 3.2). In my project the constructs are represented by design features (DF). I derived four meta-requirements from scientific literature and nine requirements from expert interviews with lecturers and students. Based

on the derived requirements I built seven DFs, which an instantiation of a FFT should have (Walls et al. 1992). The key DFs are, students receive feedback which is based on SA and CbA. Through the performance chart the students can actively monitor their performance at any time and any place. The performance of oneself can additionally be compared to fellow students for social comparison (see step two and three).

The *principles of form and function* are the essence of the design theory (Point 3 in Table 3.2). Since, I am designing a socio-technical system (Mumford 2006), one of the main principles is learner-centricity. To reach this goal, three DPs were implemented with the purpose of enhancing usability and lowering cognitive load while using FFTs. FFTs are web-based application with responsive UI to enhance usefulness (DP1). FFTs help lecturers and students to interpret the assessment results by presenting them at different aggregation levels (DP2). That way, FFTs enable lecturers and students to decide which content was understood and which needs to be repeated. Moreover, wizards guide lecturers and students through the front-end of FFTs in order to allow for an effective usage of the tool (DP3, see Figure 3.3).

*Artifact mutability* defines the possible variability, which may occur during the theory's instantiation (Point 4 in Table 3.2). While a design theory provides a general blueprint of a FFT design, the appearance of a corresponding concrete IT artifact is dependent on the context in which it will be instantiated. Especially, the instantiation of the theory's core design features (e.g. assessment types and social comparisons) are highly depended on the context, e.g. content of the lecture or student culture. Specific content can only be assessed with a suitable assessment type. Social comparison might be lowering student performance in cultures with high power distances (Buholzer et al. 2018).

There are different ways to create *Testable Propositions (TP)* (Point 5 in 3.2). On the hand testing individual design features and on the other hand testing the effects of the tool usage in general (Gregor and Jones 2007). I decided in favor of the latter since I first want to find out if design of FFTs work in general. Therefore, I created five TPs, FFTs usage increases: objective student performance (TP1), learning interest (perceived student performance) (TP2), self-reported learning (perceived student performance) (TP3), perceived skill

development (perceived student performance) (TP4) and satisfaction (TP5) (see step seven of Figure 3.1).

*Justificatory knowledge* represents insights from literature that inform, explain and validate the design decisions (Point 6 in Table 3.2). Feedback is an important aspect to improve student performance (Price et al. 2010; Shute 2008; Hattie and Timperley 2007; Kluger and DeNisi 1996). In the meta-analysis of Hattie (2015) the effect sizes of feedback and SA in higher education belong to the top 15 effect sizes. Moreover, previous research shows that SA is able to enhance student performance (Nicol and Macfarlane-Dick 2006). McDonald and Boud (2003) have shown that training students how to carry out SA can improve their performance in final exams. Furthermore, the CDT developed by Festinger (1962) provides a theoretical lens on the psychological coherence when a cognitive dissonance occurs by showing SA, CbA and optionally social comparison results to students (see Chapter 2).

I suggest three *Principles of Implementation (PI)* (Point 7 in Table 3.2). First, LOs need to be closely related to the course content. LOs are on a continuum between very generic to very specific. Hence, the lecturer needs to define LOs which are balanced between generic and specific LOs (PI1). Second, the assessment items need to reflect the content of the lecture and need to be linked to the LOs. The assessment items are on a continuum between very easy and very difficult. The lecturer needs to find a mix between easy and difficult items (PI2). Third, the number of items per assessment should allow students to carry out the assessment in a reasonable amount of time (PI3).

*Expository instantiation* (Point 8 in Table 3.2): I implemented the artifact over the last 1.5 years and used it in the real world setting of my university (see step four and six, Figure 3.3).

**Table 3.2:** Documentation of the Design Knowledge based on Gregor and Jones (2007)

| Component                          | Description                                                                                                                                                                                                                                                                                                                                      |
|------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1) Purpose and scope               | The purpose of formative feedback tools is to provide students with formative feedback during the semester concerning their current state of knowledge. The focus lies on higher education at campus university or distance learning providers such as MOOCs                                                                                     |
| 2) Constructs                      | DF1) Taxonomy table, DF2) SA, DF3) CbA, DF4) Assessment type, DF5) Confidence level, DF6) Student performance charts, DF7) Social comparison                                                                                                                                                                                                     |
| 3) Principles of form and function | DP1) Usefulness through responsive UI, DP2) Reduce cognitive load through aggregation levels, DP3) Reduce cognitive load through walkthrough wizards                                                                                                                                                                                             |
| 4) Artifact mutability             | Core design features might be adapted to special contexts, e.g. a) content of the class and assessment types or b) social comparison based on the student culture                                                                                                                                                                                |
| 5) Testable propositions           | TP1) FFT usage increases objective student performance, TP2) FFT usage increases learning interest (perceived student performance), TP3) FFT usage increases self-reported learning (perceived student performance), TP4) FFT usage increases perceived skill development (perceived student performance), TP5) FFT usage increases satisfaction |
| 6) Justificatory knowledge         | Scientific Literature on formative feedback (Hattie and Timperley 2007), SA (Boud 2013), CbA (Boud and Falchikov 2007), CDT (Festinger 1962)                                                                                                                                                                                                     |
| 7) Principles of implementation    | PI1) LOs need to be linked to course content, PI2) Assessment items need to be linked to course content and LOs, PI3) Assessment items should allow students to carry out the assessment in a reasonable amount of time                                                                                                                          |
| 8) Expository instantiation        | I created an instance of FFTs called LOOM over the last 1.5 years and used it in large-scale classes at my university                                                                                                                                                                                                                            |



### 3.5.1 Limitations and Outlook

A limitation of this study is that I tested the requirements for one specific context: a large-scale Management Information Systems class at Masters' level. Furthermore, this study was limited by the absence of a real experiment setting with a control group.

For future research, I suggest two points. First, a further experiment with e.g. eye tracking to investigate each design principle and feature individually. Hence, which of the principles and features have the strongest effect to the students' performance. Second, based on the CDT, the next step is to investigate how the SA and CbA results change over time when students using FFTs. Furthermore, I want to investigate the long-term effects of a high or low cognitive dissonance on student performance.

## 3.6 Conclusion of this Chapter

The aim of the research presented in this chapter has been a) to examine what requirements should be considered when designing an IT tool that aims at providing formative feedback in large-scale classes and b) to evaluate the extent to which usage of the IT tool increases student performance. To design the artifact, I used the DSR approach. I derived three design principles and seven design features from the meta-requirements of the scientific literature and requirements from expert interviews with lectures and students. The results of the one-group field test showed that, for the whole class, the usage of LOOM (instance of FFTs) had a significant effect on objective-measured student performance.

Furthermore, for the subset of the student class that filled out the pre-/post-test, the results showed that the usage of LOOM significantly increased their objective and perceived performance. Specifically, students using the tool received more points in the final exam and perceived themselves to have higher learning interest, skill development and understanding of the concepts and content discussed in lecture. By documenting the design knowledge, I was able to make the first important contribution to creating a design theory for formative feedback tools.



# 4

## **Helpfulness of Online Reviews**

### 4.1 Introduction

In this chapter<sup>1</sup> I will present the current state of research on how to measure online review helpfulness. The number of online reviews is growing significantly. In only three years, from 2014 to 2017, the number of reviews on TripAdvisor grew by 300%, to 600 million reviews in total (Statista 2018a). In the same period the number of reviews on Yelp grew by 208% to 148 million in total (Statista 2018b), and the statistics are similar for other online platforms.

However, not all online reviews are equally valuable. Some reviews are perceived to be more helpful or trustworthy than others. For example, it is not clear whether longer or shorter reviews (i.e., text length) are more helpful. Filieri (2016) argues that shortness indicates a low quality review while Kwok and Xie (2016) argues that shorter reviews are more helpful because they are more readable. Hence, the vast amount of available data and mixed findings has attracted a plethora of scholars to conduct research into the role of online reviews and the factors and characteristics determining their helpfulness (Lauterbach et al. 2009; Pavlou and Gefen 2004; Ba and Pavlou 2002).

Overall, this research shows that online reviews address similar challenges to peer feedback. This is why it has been useful to apply the existing knowledge from this research field to research into peer feedbacks. Online reviews and feedback are similar in terms of the assembly of the *system*, *structure* of the review/feedback and the evaluation of the *perceived quality*. The assembly of the *system* of online reviews is similar to that for peer feedback. There is a review provider, a review receiver and an IT platform which orchestrates the process. One difference is that the review providers freely choose which product or service they review. This differs to peer feedback in which the system specifies to whom feedback should be given and how many to give. The *structure* of online reviews is similar to the structure of peer feedback. Both consist of written text and have the aim of evaluating and judging the advantages and disadvantages of, in the case of online review case, a product or service, and, in the case of peer feedback, an assignment submission. One difference is

---

<sup>1</sup> A previous version of this chapter (Sections 4.2 to 4.4) has been published in Rietsche et al. (2019). The content has been reformatted and updated to the latest state of scientific discussion at the time of submission of this dissertation.

that usually online reviews focus more on the experience of using a product or service whereas peer feedback tries to provide suggestions on how to improve the submission. The evaluation of the feedback/review *perceived quality* is similar in both cases. In the case of online reviews, often stars from 1-5 are used to indicate the helpfulness while, for peer feedback, question items with a 5 point Likert scale are often used.

The research literature on online reviews provides valuable insights into measuring online review helpfulness and thus the aim of this chapter is to answer the third research question of this dissertation:

*RQ3: Which characteristics determine the quality of technology-mediated peer feedback?*

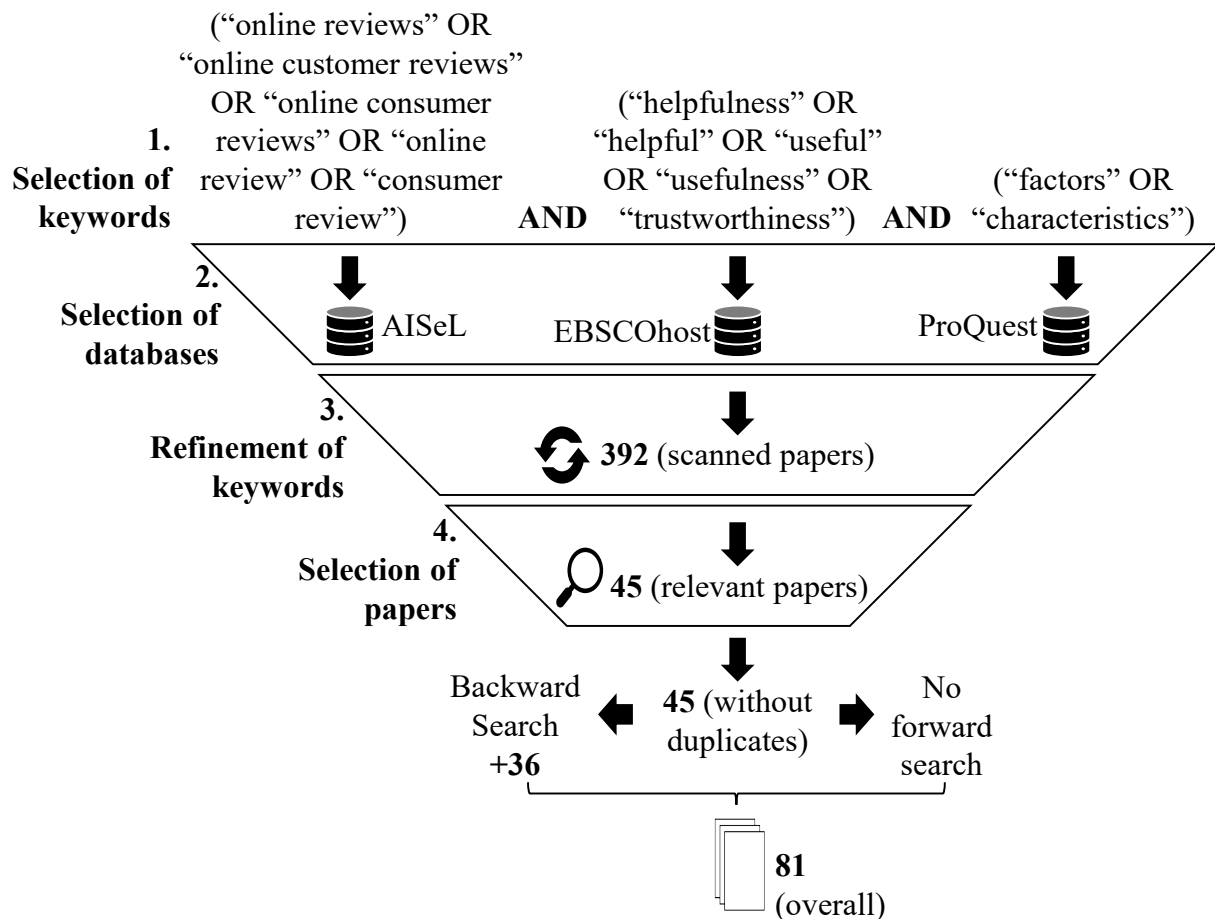
To answer that third research question, I conducted a systematic literature review of research into online reviews. The outcome of this literature review is a framework in which the literature is integrated. Overall, 81 journal articles were analyzed in depth.

The remainder of this chapter is structured as follows. In the next section I describe the research methodology and how the research studies were selected for the literature review. The following section then presents the findings of the review process on both a descriptive and thematic level. In the next section, the findings, limitations and future research are discussed. The final section provides a conclusion to this chapter.

## 4.2 Research Methodology

I conducted a structured literature review using the approach of Webster and Watson (2002) and vom Brocke et al. (2015). The first step of the literature search process included the development of the search string (see Figure 4.1), which was used for the search in the three databases: a) AISEL, b) ProQuest and c) EBSCOhost. To create a holistic search string, synonyms for all key words were included. The search was restricted to journal articles published in the years 2011 to 2018 (July 21st) and to "abstract only" because all relevant journal articles use these key words in the abstract. The period was chosen because in the years before, journal articles were only published occasionally. In total, 392

journal articles were found in the three databases, which were chosen because they focus on literature about Information Systems. The greatest part of journals conducted an analysis of publicly available data collected from various review websites. Amazon is the most prominent data source for analysis (35%). Second most used review platform is Yelp (15%) which is followed by TripAdvisor (10%).



**Figure 4.1:** Literature Search Process of Online Review Helpfulness based on the Approach of (Webster and Watson 2002) and (vom Brocke et al. 2015)

The second step included the definition of selection criteria. Only journal articles were selected for the final analysis, which were published in a journal with a 2017 Journal Impact Factor (JIF)/2017 CiteScore over 1.0. JIF values are provided by Thomson Reuters (Reuters 2017) and the CiteScores (Scopus 2017). The JIF score was used because it is a well-established measure for the relevance of journals. The CiteScore was added because not all journals were found on Thomson Reuters and in this case, the CiteScore was used as a substitute. Additionally, only journal articles which included qualitative or

quantitative data collection were used for the final analysis. The screening and selecting process left 45 of the initial 392 journal articles for further analysis. These 45 journal articles were then used for a backward search as proposed by Webster and Watson (2002). In total 871 sources were screened and selected during the backward search. The same criteria were used as for the initial search. Regarding the publication year, one exception was made for Mudambi and Schuff (2010) because this paper is fundamental for this research field and many of the following researchers cited this paper. Finally, 36 new journal articles were added to the initial 45 resulting in a total of 81 journal articles. For the final analysis, a framework with four dimensions was inductively created based on the findings in the literature.

For presenting the results, I developed a new method called co-topic analysis, which is based on the concept matrix proposed by Webster and Watson (2002). The idea of this method is to provide a holistic and compact overview of topics researched in scientific articles and more importantly their intra- and interrelationship. Co-topic analysis has the advantage over the approach of tables, of visualizing the most important information using a small amount of space. A visualization makes it easier for the reader to absorb the information. The method uses a hierarchical approach including dimensions which are at the first level an aggregation of 1:n topics, at the second level the topics and at the third level the subtopics. The co-topic analysis provides the reader insights into four important questions: 1) How many topics are researched in total. Hence, which of the factors are researched in which paper, which is shown by the small reference number for each sub-factor. 2) What are sub-topics and how often are they researched. Thus, the reader is easily able to identify the importance of dimensions and topics by considering the bubble sizes including the number of papers researching this topic. 3) How many papers researched a topic/sub-topic and 4) How many papers are researching the same topic. The reader is able to identify the factors, which are most frequently researched together, which is showed by the arrows between the sub-factors including the number of papers investigating this factor together.

### 4.3 Findings

To classify the literature, I built on the framework by Liu (2015) and Yang et al. (2017b) and inductively altered it based on the findings of our literature analysis. Whereby the components of our framework reflect the general understanding of information exchange, which includes the actor A and B, the information to be exchanged and the environment. Yang et al. (2017b) explicitly used the terms “reviewer-related attributes” and “review-related attributes”. We extended this framework with the reader-related factors and environment-related factors. Hence, the framework consist of four dimensions: 1) review-related factors, 2) reviewer-related factors, 3) reader-related factors and 4) environment-related factors. During the literature search process, the literature was clustered according to this framework. The individual factors are introduced subsequently.

The first dimension contains review-related factors. This includes all aspects which are related to the review itself, hence the content, structure and value (what information is exchanged). Consistency measures to what extent a review is similar or different to other reviews for the same product (Cheung et al. 2012). The factor images represents all visual information added to a review (Yang et al. 2017b). Persuasiveness includes different measures such as emotions in the text or the quality of the arguments (Fang et al. 2016). Posting time includes all time measurements (Fang et al. 2016). Reference measures to what extent a review is referring to other reviews or products (Weathers et al. 2015). Rating subsumes all factors in relation to numerical or star rating (Cheung et al. 2012). Depth measures the amount of information in the review by analyzing the number of sentences, words or characters (Guo and Zhou 2017). Readability measures the ease of reading a review with the help of readability tests (e.g. Gunning-Fog index, Flesch-Kincaid reading ease, automated readability index, Coleman-Liau index) (Korfiatis et al. 2012).

The second, dimension contains the reviewer-related factors. This is defined as factors, which are directly related to the person providing the review, hence the characteristics of the person as well as previous actions (who is the actor sending information). Information disclosure includes all information available about the reviewer such as location, profile picture, age, gender etc. (Guo and



Zhou 2017). Reputation measures the extent to which a reviewer is perceived to be an expert and credible, hence the total number of reviews written, the social status on the platform etc. (Cheung et al. 2012).

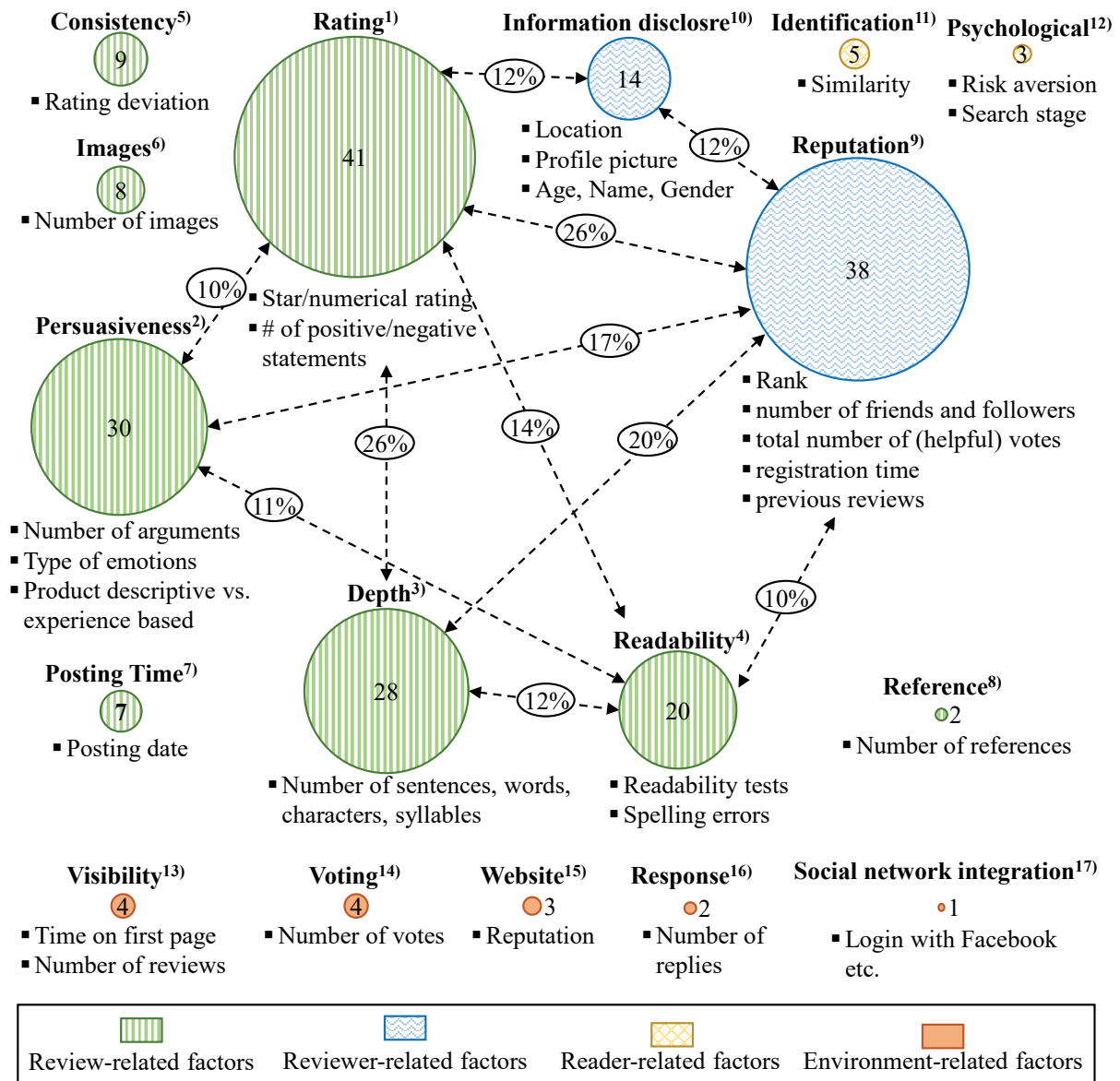
The third, dimension contains the reader-related factors. This dimension contains all factors, which are related to the person who receives the review. Hence, the psychological condition of the reader, as well as how the reader responds to the review (who is the actor receiving information). The psychological factors include all internal influences on the perception of the reviews, such as risk aversion and search strategy (Li et al. 2017). Identification measures to what extent the reader identifies him-/herself with the review environment and the reviewer (Davis and Agrawal 2018).

The fourth and last dimension contains the environment-related factors. This dimension contains all factors of how and where the reviews are provided from the reviewer to the reviewee (how and in which environment is the information exchanged). Visibility measures to what extent the review is/was visible to the readers, such as total number of available reviews (Hu and Chen 2016). Voting measures the influence of a voting system on the perception of the reviews (Kuan et al. 2015). Social network integration measures what effect the integration of social networks (e.g. Facebook) have on the online review helpfulness (Huang et al. 2017). Response deals with the effect of responses to a review on the helpfulness (Ryang et al. 2015). The website factor captures the influence of the website on which the review was posted, hence the reputation (Kim et al. 2017).

#### 4.3.1 Criteria Determining the Helpfulness of Online Reviews

Figure 4.1 shows the results of the literature review using co-topic analysis, whereby in my case a topic is a factor. The colored clusters show the different factors in each dimension: review-related, reviewer-related, reader-related and environment-related. The bubble-size represents the relative number of journal articles investigating a factor. Below the bubble, the bullet points show the most important characteristics of this factor. The superscript number at the factor name shows, which article researched this factor. The lines between the bubbles indicate the top ten frequent links of how many times the two factors were researched together.

Regarding the total number of journal articles published per year, the importance of the individual factors remained constant over the years. The size of the bubbles in Figure 4.1 below indicates how many journal articles investigated each factor, which makes them comparable and shows the current research state. Among the most important factors (at least 10 journal articles), some were found to be more likely to be investigated together. 26% of all journal articles investigated both, the influence of review depth on helpfulness and the effect of review ratings. Other note-worthy interdependencies were found between reputation and rating (26% of all journal articles investigated both factors), depth and reputation (20%), readability and rating (14%), information disclosure and reputation (12%), information disclosure and rating (12%) and readability and depth (12%).



**Figure 4.2:** Part1: Co-Topic Analysis of 81 Journal Articles. Colored clusters: Different dimensions with factors. Bubble-size: Relative number of journal articles investigating a factor. Bullet points: Most important characteristics of this factor. Superscript number: Articles researched this factor. Lines: Frequency of how many times the two factors were investigated together (in percent of the total amount of journal articles = 81).

#### 4. Helpfulness of Online Reviews

- <sup>1)</sup> Baek et al. (2013); Baek et al. (2015); Cao et al. (2011); Casaló et al. (2015); Chen (2016); Cheung et al. (2012); Filieri (2016); Gao et al. (2017); Huang et al. (2015); Jensen et al. (2013); Krishnamoorthy (2015); Kuan et al. (2015); Kusumasondjaja et al. (2012); Kwok and Xie (2016); Lee (2013); Lee (2018); Lee and Choeh (2017); Lee and Choeh (2018); Lee et al. (2011); Li et al. (2017); Liu and Park (2015); Mudambi and Schuff (2010); Pan and Zhang (2011); Park and Nicolau (2015); Pentina et al. (2018); Purnawirawan et al. (2012); Quaschnig et al. (2015); Racherla and Friske (2012); Salehan and Kim (2016); Schindler and Bickart (2012); Schlosser (2011); Singh et al. (2017); Weathers et al. (2015); Willemsen et al. (2011); Wu (2013); Yang et al. (2017a); Yang et al. (2017b); Yin et al. (2013); Yin et al. (2016); Zhao et al. (2015); Zhou and Guo (2017)
- <sup>2)</sup> Agnihotri and Bhattacharya (2016); Ahmad and Laroche (2015); Banerjee and Chua (2014); Chen and Tseng (2011); Cheung et al. (2012); Chong Alain et al. (2018); Chua and Banerjee (2016); Fang et al. (2016); Ghose and Ipeirotis (2011); González-Rodríguez et al. (2016); Huang et al. (2013); Hong et al. (2016); Kim et al. (2017); Krishnamoorthy (2015); Li et al. (2013); Luan et al. (2016); Luo et al. (2015); Malik and Hussain (2017); Racherla et al. (2012); Robinson et al. (2016); Salehan and Kim (2016); Schindler and Bickart (2012); Shan (2016); Shen et al. (2016); Teng et al. (2014); Weathers et al. (2015); Willemsen et al. (2011); Yin et al. (2013); Zhao et al. (2015); Zheng et al. (2013)
- <sup>3)</sup> Baek et al. (2013); Banerjee and Chua (2014); Cheng and Ho (2015); Chua and Banerjee (2014); Filieri (2016); González-Rodríguez et al. (2016); Huang et al. (2015); Korfiatis et al. (2012); Kuan et al. (2015); Kwok and Xie (2016); Lee (2018); Lee and Choeh (2017); Lee and Choeh (2018); Liu and Park (2015); Mudambi and Schuff (2010); Pan and Zhang (2011); Park and Nicolau (2015); Qazi et al. (2016); Quaschnig et al. (2015); Racherla and Friske (2012); Robinson et al. (2016); Schindler and Bickart (2012); Singh et al. (2017); Yang et al. (2017a); Yang et al. (2017b); Yin et al. (2013); Yin et al. (2016); Zhou and Guo (2017)
- <sup>4)</sup> Agnihotri and Bhattacharya (2016); Banerjee and Chua (2014); Cao et al. (2011); Fang et al. (2016); Filieri (2016); Gao et al. (2017); Ghose and Ipeirotis (2011); Jensen et al. (2013); Korfiatis et al. (2012); Kuan et al. (2015); Lee and Choeh (2014); Liu and Park (2015); Ngo-Ye and Sinha (2014); Robinson et al. (2016); Schindler and Bickart (2012); Singh et al. (2017); Yang et al. (2017a); Yin et al. (2013); Zhao et al. (2015); Zheng et al. (2013)
- <sup>5)</sup> Baek et al. (2013); Baek et al. (2015); Cheung et al. (2012); Gao et al. (2017); Hong et al. (2016); Luo et al. (2015); Pan and Zhang (2011); Quaschnig et al. (2015); Yin et al. (2016)
- <sup>6)</sup> Casaló et al. (2015); Cheng and Ho (2015); Filieri (2016); Lee (2018); Teng et al. (2014); Yang et al. (2017a); Yang et al. (2017b); Zhou and Guo (2017)
- <sup>7)</sup> Cao et al. (2011); Hu and Chen (2016); Kuan et al. (2015); Lee (2013); Yin et al. (2016); Zhao et al. (2015); Zhou and Guo (2017)
- <sup>8)</sup> Robinson et al. (2016); Weathers et al. (2015)
- <sup>9)</sup> Agnihotri and Bhattacharya (2016); Baek et al. (2013); Banerjee et al. (2017); Chen (2016); Cheng and Ho (2015); Cheung et al. (2012); Chong Alain et al. (2018); Chua and Banerjee (2014); Fang et al. (2016); Filieri (2016); González-Rodríguez et al. (2016); Gao et al. (2017); Hu and Chen (2016); Huang et al. (2015); Kim et al. (2017); Kuan et al. (2015); Kwok and Xie (2016); Lee et al. (2011); Lee (2018); Lee and Choeh (2017); Lee and Choeh (2018); Li et al. (2013); Li et al. (2017); Liu and Park (2015); Min and Park (2012); Park and Nicolau (2015); Racherla and Friske (2012); Shan (2016); Shen et al. (2016); Teng et al. (2014); Weathers et al. (2015); Willemsen et al. (2011); Xu (2014); Yang et al. (2017b); Zhao et al. (2015); Zheng et al. (2013); Zhou and Guo (2017); Zhu et al. (2014)
- <sup>10)</sup> Baek et al. (2013); Gao et al. (2017); Ghose and Ipeirotis (2011); Hong et al. (2016); Kusumasondjaja et al. (2012); Kwok and Xie (2016); Lee and Choeh (2017); Lee et al. (2011); Liu and Park (2015); Otterbacher (2013); Park and Nicolau (2015); Racherla and Friske (2012); Xu (2014); Zhou and Guo (2017)
- <sup>11)</sup> Davis and Agrawal (2017); Guo and Zhou (2017); Pentina et al. (2018); Shan (2016); Teng et al. (2014)
- <sup>12)</sup> Casaló et al. (2015); Memarzadeh et al. (2016); Robinson et al. (2016)
- <sup>13)</sup> Hu and Chen (2016); Kuan et al. (2015); Lee and Choeh (2017); Pan and Zhang (2011)
- <sup>14)</sup> Lee (2013); Liu and Park (2015); Yang et al. (2017a); Zhao et al. (2015)
- <sup>15)</sup> Bae and Lee (2011); Chong Alain et al. (2018); Kim et al. (2017)
- <sup>16)</sup> Kwok and Xie (2016); Ryang et al. (2015)
- <sup>17)</sup> Huang et al. (2017)

**Figure 4.3: Part 2/2: Journal Authors of Co-topic Analysis**

### 4.3.2 Review-Related Factors

#### 4.3.2.1 Rating

Rating is - next to reviewer reputation - the most important factor for determining the helpfulness of a review (Yang et al. 2017b). Most researchers found that negative reviews are more helpful than positive reviews (Lee et al. 2018; Li et al. 2017; Yang et al. 2017a; Zhou and Guo 2017; Chua and Banerjee 2016; Kwok and Xie 2016; Casaló et al. 2015; Kuan et al. 2015; Zhao et al. 2015; Baek et al. 2013; Lee 2013; Yin et al. 2013; Cao et al. 2011; Lee et al. 2011). One possible explanation for this effect is the negativity bias as explained in the theoretical background section. However, if a review is negative and at the same time promotes another product, the review is perceived to be untrustworthy and not helpful at all (Filieri 2016). Filieri (2016) argues that negative reviews tend to stand-out and are supposedly written by competitors. Few researchers however found that positive reviews are more helpful than negative reviews (Casaló et al. 2015; Huang et al. 2015; Liu 2015; Pan and Zhang 2011). Pan and Zhang (2011) explain this finding with the confirmatory bias.

The confirmatory bias describes a situation in which the customers have a pre-decisional preference and if an online review provides validation for this preference, it is perceived to be more helpful (Liu 2015). However, Pan and Zhang (2011) add that whether positive or negative reviews are more helpful strongly depends on other factors like personality traits and consumption goals and therefore, a generalization of either finding could be dangerous. Regarding product types, Willemsen et al. (2011) concluded that negative reviews are more helpful for experience goods and positive reviews are more helpful for search goods.

In contrast, Racherla and Friske (2012) discovered that the negativity effect is only present for search goods and not for experience goods. Some other researchers argue that it depends on the average rating e.g. positive reviews are more helpful for positive average rating (Yin et al. 2016; Baek et al. 2015; Quaschnig et al. 2015). However, Lee (2013) argues that reviews which differ from the average rating deliver more value because they provide new information and therefore, they are more helpful. These findings are strongly contradicting and should therefore be further investigated. Together with rating,

many researchers investigated the effect of one- and two-sidedness on the review helpfulness. One-sided reviews contain either positive or negative statements whereas two-sided reviews include positive and negative statements (Cheung et al. 2012). Most researchers agree that two-sided reviews are more helpful compared to one-sided reviews (Filieri 2016; Salehan and Kim 2016; Weathers et al. 2015; Jensen et al. 2013; Cheung et al. 2012; Schindler and Bickart 2012; Schlosser 2011; Willemsen et al. 2011).

Salehan and Kim (2016) argue that one-sided reviews are perceived to be biased. Schlosser (2011) adds that – with the exception of reviews with extreme ratings – two-sided reviews increase the trustworthiness. However, Lee et al. (2018); Pentina et al. (2018); Purnawirawan et al. (2012), also found that one-sided reviews are more helpful. Purnawirawan et al. (2012) argue that one-sided reviews provide a clear direction and therefore increase the perception that the information in the review is true. Furthermore, Lee et al. (2018) add that two-sided reviews in general fail to offer a clear statement leaving the reader with uncertainty. The explanation of Pentina et al. (2018) is directed to the confirmation bias. Pentina et al. (2018) argue that users of digital review sites are not looking for objective unbiased information but rather want to confirm their pre-existing intentions. Third and in connection to extremity, Kuan et al. (2015) found that reviews with extreme ratings receive more voting but are perceived to be less helpful. They say that on the one hand moderate rated reviews are more likely to be disregarded and therefore, extreme reviews get more voting but on the other hand extreme rated reviews have a greater potential of being biased (Kuan et al. 2015). Kuan et al. (2015) conclude, that a rating in between the average and the absolute extreme provides the greatest value. Although Mudambi and Schuff (2010) agree with Kuan et al. (2015), some other researchers found that extreme ratings increase the review helpfulness (Gao et al. 2017; Lee and Choeh 2014; Yin et al. 2013; Wu 2013). In numbers, reviews with either 1- or 5-stars are the most helpful, 2-star reviews are second most helpful and 4- and 3-star reviews are least helpful (Park and Nicolau 2015). In contrast, Filieri (2016) argues that reviews with extreme ratings were found to be potentially manipulated and therefore rating extremity should always be measured together with other factors which allow the reader to determine whether a review is fake or real.

#### 4.3.2.2 Persuasiveness

Persuasiveness is determined by the number and strength of the arguments, comprehensiveness, timeliness and sidedness (Chong et al. 2018). Most researchers agree that argument quality increases the helpfulness of a review (Shen et al. 2016; Shan 2016; Zhao et al. 2015; Cheung et al. 2012). Shan (2016) sees the reason thereof in the greater trustworthiness that is delivered by strong arguments. Zhao et al. (2015) argue that in the anonymity of the internet, people seek for more cues to judge information and that strong arguments make this process of judgment easier. Shen et al. (2016) conclude that readers are more likely to adopt the information of a review if the argument quality is high. Furthermore, the effect of argument quality is stronger if either the reader perceives a strong sense of membership Luo et al. (2015) or if the similarity of reader and reviewer is high (Racherla and Friske 2012). In general, argument quality was found to be one of the strongest determinants of online review helpfulness (Teng et al. 2014; Cheung et al. 2012).

According to Willemsen et al. (2011) a higher number of arguments increases the review helpfulness independently of the argument quality. This is in line with Robinson et al. (2012) who say that information in the review is a strong predictor of its helpfulness. Emotions also influence the persuasiveness and therefore review helpfulness. However, researchers do not agree on how they affect the review helpfulness (Agnihotri and Bhattacharya 2016; Fang et al. 2016; González-Rodríguez et al. 2016; Hong et al. 2016; Salehan and Kim 2016). Ahmad and Laroche (2015) found in their analysis of emotion features that happiness and disgust in a review positively influence helpfulness and that anxiety has a negative effect on helpfulness. Yin et al. (2013) proclaim that reviews containing anxiety are more helpful than reviews containing anger.

Although these findings are contradictory, two interesting observations were made by Malik and Hussain (2017) and by Banerjee and Chua (2014). First, Malik and Hussain (2017) stated that positive emotion features (e.g. trust, anticipation and joy) are better predictors of helpfulness than negative emotion features (e.g. anxiety, sadness and anger). Second, Banerjee and Chua (2014) found fewer negative emotion words in negative reviews that were voted as

helpful and fewer positive emotion words in positive reviews that were voted as helpful.

Regarding persuasiveness connected to product types, most researchers agree that reviews with more product descriptive statements are more helpful for search goods (Malik and Hussain 2017; Chua and Banerjee 2016; Luan et al. 2016; Krishnamoorthy 2015; Weathers et al. 2015; Huang et al. 2013). Some argue that product descriptive statements in general increase the helpfulness (Zheng et al. 2013; Schindler and Bickart 2012; Chen and Tseng 2011). Weathers et al. (2015) add that information about the usage of the product increases the helpfulness for both product types. However, experience-based reviews seem to be more helpful for experience goods (Malik and Hussain 2017; Krishnamoorthy 2015; Huang et al. 2013). Luan et al. (2016) found that for experience goods in general it does not matter whether a review is attribute- or experience-based because the reader simply wants to get as much information as possible. This statement is in accordance with Chen and Tseng (2011) and Zheng et al. (2013) respectively who say that personal experience and opinions increase the helpfulness.

Persuasiveness is further linked to subjectivity/objectivity. In general, it is easier to determine the helpfulness of objective reviews (Zheng et al. 2013). Therefore, objective reviews are often considered to be more helpful (Kim et al. 2017; Li et al. 2013). Nevertheless, a mixture of objective sentences mixed with extreme subjective content was also found to be helpful (Ghose and Ipeirotis 2011).

##### 4.3.2.3 Depth

Regarding review depth, most researchers found that longer reviews are more helpful (Lee et al. 2018; Lee and Choeh 2017; Yang et al. 2017a; Zhou and Guo 2017; González-Rodríguez et al. 2016; Yin et al. 2016; Cheng and Ho 2015; Liu 2015; Park and Nicolau 2015; Quaschnig et al. 2015; Yin et al. 2013). However, reviews can also be too long and therefore, review helpfulness increases proportional to length but only up to a certain point (Baek et al. 2013; Schindler and Bickart 2012). One simple explanation for this finding is that longer reviews have a higher chance of providing more information to the reader. Nevertheless, at the same time, in the case of long average reviews, an even



longer review fails to attract the readers' attention, because the reader must put in more effort into reading the longer review for a perceived small amount of extra information (Qazi et al. 2016; Chua and Banerjee 2014). For DVDs, e.g., reviews with more than 430 words were found to be too long, for books a review should contain around 230 words to be exceptionally helpful (Kuan et al. 2015). Findings regarding length in connection with product type show that length is more important for search goods than experience goods but it is an important factor for both product types (Baek et al. 2013; Pan and Zhang 2011; Mudambi and Schuff 2010). In contrast, Racherla et al. (2012) found that longer reviews for experience goods can be less helpful. A possible explanation therefore is that due to the many available reviews and the connected information overload, consumers are overwhelmed and do not pay attention to longer reviews (Racherla et al. 2012). Additionally, the positive effect of review length is weaker for negative reviews (Banerjee and Chua 2014) and among top reviewers, the effect of review length is insignificant (Huang et al. 2015).

Length compared to other factors only has a small impact on helpfulness (Singh et al. 2017). Factors like readability, rating or reputation of the reviewer are more important (Yang et al. 2017b; Korfiatis et al. 2012). An explanation for this is provided by Robinson et al. (2012) who say that length preferences strongly depend on the search stage of the recipient. For example in an early search stage, the reader tries to get an overview of the product and prefers shorter reviews, however, in a late search stage, the reader prefers longer reviews to collect detailed information about the product.

In contrast to most other researchers, Kwok and Xie (2016) found shorter reviews to be more helpful which they affiliated with greater readability and Qazi et al. (2016) discovered that shorter reviews are more helpful if they are suggestive. However, Filieri (2016) disagrees because his research revealed that short reviews are perceived to be fake and untrustworthy. Filieri (2016) argues that short reviews often do not provide detailed information about product or usage experiences. Especially in connection with sensational titles, emotional and gushy language and the use of superlatives shorter reviews are frequently perceived to be fake (Filieri 2016).

##### 4.3.2.4 Readability

Regarding readability, Singh et al. (2017) found that it is a strong indicator of review helpfulness for both, search and experience goods. However, researchers found mixed findings for the direction of the effect. Some researchers concluded that less readable reviews with longer sentences and more complex words are more helpful, because they are considered as more professional (Yang et al. 2017a; Kuan et al. 2015). Agnihotri and Bhattacharya (2016) agree that too simplistic reviews are less helpful but nevertheless found that readability increases the review helpfulness although only up to a certain point.

Most researchers, however, represent the opinion that greater readability in general increases review helpfulness (Fang et al. 2016; Liu 2015; Zhao et al. 2015; Banerjee and Chua 2014; Yin et al. 2013; Ghose and Ipeirotis 2011). According to Ghose and Ipeirotis (2011) readability improves the comprehensibility of reviews and therefore a larger number of users can potentially read it and give voting. Korfiatis et al. (2012) compared the effect strength of different factors and found that readability has a greater impact on helpfulness than review length. Regarding language in general, promotional language decreases the helpfulness because customer do not trust these kind of reviews (Filieri 2016).

However, humor and slang can increase the helpfulness but only up to a certain point (Schindler and Bickart 2012). Schindler and Bickart (2012) argue that the informal style can increase the readers feeling of similarity with the reviewer and therefore makes the review more helpful. Nevertheless, too much informality can make the reader feel uncertain regarding the reviewers competence (Schindler and Bickart 2012). According to Robinson et al. (2012) and Singh et al. (2017), spelling mistakes only have a minor impact on helpfulness. Nevertheless, various researchers found that few spelling mistakes and short sentences with familiar words increase the helpfulness of a review (Lee 2014; Zheng et al. 2013; Cao et al. 2011; Ghose and Ipeirotis 2011). Summarized, language influences the perception of reviews in different ways and therefore, text mining methods can be useful for the prediction of review helpfulness (Ngo-Ye and Sinha 2014).

#### 4.3.2.5 Consistency

Most researchers found that consistency with previous reviews increases the helpfulness of a review (Yin et al. 2016; Baek et al. 2015; Luo et al. 2015; Quaschnig et al. 2015; Baek et al. 2013; Cheung et al. 2012). However, Baek et al. (2013) found that this factor is more important for experience goods and for products with a low price. They argue that it depends on the purpose of reading online reviews and that in the case of experience goods and low-priced products most people are narrowing down possible choice options rather than evaluating alternatives (Baek et al. 2013). Strong disagreement among reviews leads to a decrease in the helpfulness of all reviews because of the resulting uncertainty for the reader (Pan and Zhang 2011).

However, some researchers found reviews with a strong deviation from the average rating to be more helpful for the reader (Gao et al. 2017; Hong et al. 2016). This might be explained by the greater amount of information delivered by an inconsistent review. The explanation is similar to the argumentation of the negativity bias, which concludes that negative reviews are more helpful than positive reviews because of providing newer information.

#### 4.3.2.6 Images

Occasionally, reviewers add images to their reviews. Filieri (2016) regards images as proof that the reviewers used the reviewed product and therefore argues, that images increase the helpfulness. Other researchers support this finding (Zhou and Guo 2017; Cheng and Ho 2015; Teng et al. 2014). Additionally, Yang et al. (2017a) discovered that images of the food and beverages increase the helpfulness of restaurant reviews. However, Casaló et al. (2015) and Lee et al. (2018) disagree because they found that images have an insignificant effect on the review helpfulness.

#### 4.3.2.7 Posting time

Posting time directly influences review helpfulness because more recently posted reviews get more helpfulness votes (Zhou and Guo 2017; Yin et al. 2016; Zhao et al. 2015; Cao et al. 2011). Lee (2013) found the exact opposite, namely

that early posted reviews are read more often and therefore receive more voting (Kuan et al. 2015). It is possible that the most helpful review of a product is the most helpful review among early posted reviews only (Lee 2013). Therefore, posting time is generally important for the prediction of review helpfulness (Hu and Chen 2016).

##### 4.3.2.8 References

Either references in reviews can be directed to other reviews or they are used to compare the product to products from other brands. Weathers et al. (2015) found that references to other brands significantly increase the helpfulness for experience goods but have almost no effect for search goods. References to other reviews increase helpfulness for both product types (Weathers et al. 2015). According to Robinson et al. (2012) references increase the amount of information in a review and therefore generally rise its helpfulness.

##### 4.3.3 Reviewer-Related Factors

###### 4.3.3.1 Reputation

Reputation has a strong impact on review helpfulness (Chua and Banerjee 2016). High ranked credible reviewers write more helpful reviews (Chong et al. 2018; Lee et al. 2018; Banerjee et al. 2017; Li et al. 2017; Kwok and Xie 2016; Shen et al. 2016; Cheng and Ho 2015; Kuan et al. 2015; Park and Nicolau 2015; Teng et al. 2014; Zhu et al. 2014; Baek et al. 2013; Cheung et al. 2012; Racherla and Friske 2012). This factor is more important for experience goods and products with a lower price (Zhu et al. 2014; Baek et al. 2013). For example for higher priced hotels, the effect of credibility was found to be weaker than for lower priced hotels (Zhu et al. 2014). Zhu et al. (2014) argue that the cognitive trust is stronger for lower priced hotels. However, other sub-factors of reputation like the review rating and the total number of helpfulness votes a reviewer received in the past have a greater effect on review helpfulness (Shan 2016).

The helpfulness of future reviews is greatly determined by the reviewers' capability of writing helpful reviews (Lee and Choeh 2018). Gao et al. (2017) found that the rating behavior of reviewers is consistent over time and that therefore reviewers who write helpful reviews in the past are more likely to write helpful

reviews in the future (Banerjee et al. 2017; Fang et al. 2016). In general, the total number of helpfulness votes or the average helpfulness of a reviewer (total number of helpfulness votes / total number of votes) is a good predictor of the helpfulness of future reviews (Lee and Choeh 2018; Banerjee et al. 2017; Yang et al. 2017b; Kwok and Xie 2016; Huang et al. 2015; Zheng et al. 2013). Therefore, most researchers agree that expertise - based on credentials of others (e.g. votes) - of the reviewer increases the helpfulness of his/her reviews (Kim et al. 2017; Zhou and Guo 2017; Agnihotri and Bhattacharya 2016; González-Rodríguez et al. 2016; Park and Nicolau 2015; Weathers et al. 2015; Zhao et al. 2015; Zhu et al. 2014; Lee et al. 2011; Willemsen et al. 2011). However, Min and Park (2012) disagree and warn of a possible early bird bias. They found that the number of experiences mentioned in the review is much more important.

Various researchers discovered that the total number of reviews written by a reviewer increases the trustworthiness and makes future reviews more helpful (Lee and Choeh 2018; Banerjee et al. 2017; Filieri 2016; Lee et al. 2011). However, Huang et al. (2015) argue that this is not a significant factor because quantity does not correlate with quality.

Sociability or the total number of friends and followers of a reviewer can be used as an indicator of helpfulness of his/her reviews (Lee and Choeh 2018; Banerjee et al. 2017; Li et al. 2017; Zhou and Guo 2017; Cheng and Ho 2015; Liu 2015; Zheng et al. 2013). Cheng and Ho (2015) argue that the number of followers is an indicator that the reviewer is an opinion leader and that he is trusted by many other readers. Liu (2015) add that social information can be used to assess the source of a review and to judge the credibility of a reviewer. This has an effect on the uncertainty reduction regarding the service quality (Liu 2015). However, Xu (2014) differentiated between positive and negative reviews and found that sociability only in-creases the perceived helpfulness of negative reviews.

#### 4.3.3.2 Information disclosure

Information disclosure leads to more helpfulness votes (Kusumasondjaja et al. 2012; Ghose and Ipeiritis 2011). Gao et al. (2017) found that intrinsic characteristics (e.g. culture, gender) capture the largest fraction of future rating deviation. The location of a reviewer can give a hint on the cultural background,

which allows conclusions about the degree of conformity of his/her reviews with other reviews. Hong et al. (2016) discovered that on one hand, reviewer from a collectivist culture are less likely to deviate from the average rating and that they express less emotions. However, reviewers from an individualistic culture are more likely to deviate from the prior rating and they express more emotions (Hong et al. 2016). The disclosure of the real name of the reviewer increases his/her credibility and the helpfulness of the reviews (Lee and Choeh 2018). Various researchers found that the disclosure of a profile picture increases the review helpfulness (Zhou and Guo 2017; Park and Nicolau 2015; Liu 2015).

Xu (2014) observed the same effect but only for negative reviews. Regarding the gender, Kwok and Xie (2016) found that reviews written by male reviewers are more helpful than reviews written by female reviewers (reviews written by female reviewers are 13.6% less helpful). A reason for this could be that reviews written by male reviewers have the tendency to be more fact based (Kwok and Xie 2016). Additionally, Otterbacher (2013) investigated the differences in male and female writing and found that helpful reviews written by female writers often show male style characteristics (e.g. less vocabulary richness, decreased use of 1st and 2nd person pronouns).

In contrast to these findings, Lee et al. (2011) revealed that men and women create similarly helpful reviews and that not disclosing the gender increases the helpfulness. Lee et al. (2011) argue that people who do not disclose their gender probably use the anonymity to share their experiences and personal feelings more openly and therefore the reader perceives these reviews to be more honest and sincere. Finally, the age of the reviewer has no influence on review helpfulness (Kwok and Xie 2016; Lee et al. 2011).

#### **4.3.4 Reader-Related factors**

##### **4.3.4.1 Similarity**

Perceived similarity between the reader and the reviewer influences the information adoption (Davis and Agrawal 2018). The greater the similarity, the higher is the helpfulness of a review (Pentina et al. 2018; Shan 2016; Teng et al. 2014). The helpfulness of positive reviews is increased, when the reviewer uses the same linguistic style as the reader would use (Guo and Zhou 2017). If the reader

perceives the reviewer to be on an equal level of expertise as himself, negative reviews are more helpful (Guo and Zhou 2017).

#### 4.3.4.2 Psychology

Psychological factors can influence the way a reader perceives a review. First, the extent to which the reader believes that a certain review can change his/her behavior positively influences the review helpfulness (Memarzadeh et al. 2016). Second, the risk aversion of the reader changes the readers' preference regarding positive and negative reviews. High risk aversion readers prefer negative reviews, low risk aversion readers have no preference for either positive or negative reviews (Casaló et al. 2015). Third, the search stage has a great influence on the effect of the factor depth (Robinson et al. 2012). Reader in an early search stage prefer shorter reviews because they try to get a general overview on different products and later on, they prefer longer reviews to get detailed information about specific products (Robinson et al. 2012).

### 4.3.5 Environment-Related Factors

#### 4.3.5.1 Visibility

The total number of reviews available for a product has a significant impact on the review helpfulness (Lee and Choeh 2017). This has to do with the visibility of the reviews, if there are fewer reviews, each review has the potential to offer new information and therefore, the helpfulness of each single review increases (Pan and Zhang 2011). In general, when conducting research about review helpfulness, it is very important to include the visibility factor because the total number of reviews is negatively correlated with voting and therefore can distort the results (Hu and Chen 2016; Kuan et al. 2015).

#### 4.3.5.2 Voting systems

Voting systems generally increase the helpfulness of review websites because they allow the reader to quickly assess whether a review should be/should not be considered in the purchase decision (Zhao et al. 2015). Thereby, not only voting systems for helpfulness are relevant, but also voting systems for enjoyment (Liu

2015). Enjoyment of a review is positively related to the helpfulness of the review (Yang et al. 2017a). Liu (2015) argue that in the computer-mediated environment, individuals are seeking entertainment and therefore enjoyable reviews are more helpful. The effect of voting is greater for search goods than for experience goods because people who are looking for reviews about search goods are more interested in opinions of others (Lee 2013).

##### 4.3.5.3 Review Website

Various review websites (e.g. Yelp, TripAdvisor) can be found online. Readers assign them with different credibility levels depending on their awareness level. Kim et al. (2017) found that the website credibility influences the helpfulness of a review. So does the perceived ease of use of the website (Chong et al. 2018). Regarding the provider of the website, reviews about experience products are more helpful if posted on a consumer developed review site (Bae and Lee 2011). In contrast to review websites, consumer developed review sites are defined as online communities or blogs (Bae and Lee 2011). Bae and Lee (2011) argue that customers who want to evaluate the dominant attributes of experience goods expect to benefit more from these kind of websites.

##### 4.3.5.4 Responses

The helpfulness of a review is increased whenever the management of the rated service or product responds to the review (Kwok and Xie 2016). Kwok and Xie (2016) argue that reviews with a response of the management often include detailed information and provide more reference value to the readers. Furthermore, the manager response is regarded as a cross-validation of the information in the review (Kwok and Xie 2016). Additionally, discussions between users also increase the review helpfulness, especially when the reviewer himself is actively involved in the discussion (Ryang et al. 2015). Ryang et al. (2015) argue that the discussion provides the opportunity to further elaborate on the shared opinions.



#### 4.3.5.5 Social Network Integration

Social network integration was found to increase the review volume. At the same time, less negations were used in the reviews while more positive emotions were expressed (Huang et al. 2017).

## 4.4 Discussion and Contributions

The previous research primarily focused on factors related to the reviews itself and to the reviewer who wrote the reviews. Reader-related factors and environment-related factors got less attention in general. Additionally, within the four dimensions, some factors were analyzed by significantly more researchers than other factors. Regarding the individual factors, sometimes it is possible to make a clear statement whether the factors increases or decreases the helpfulness while for others it is very hard to tell because of mixed findings.

The most important findings are summarized subsequently. Extremely rated reviews are of great value for the readers while negative reviews are more helpful than positive ones. Nevertheless, the combination of positive and negative statements increases the review helpfulness. A higher number of arguments as well as the argument quality positively affects the helpfulness of a review. Regarding emotions, the findings were mixed and need further research. Reviews with many product descriptive statements are more helpful for reviews about search goods while experience-based reviews are more helpful for reviews about experience goods.

Longer reviews provide greater value to the reader than short reviews, but only up to a certain point. Readability, consistency with previous reviews and references to other brands and reviews increase the review helpfulness. Posting time and the inclusion of images seem to be important factors but their effect is not yet fully discovered. Regarding the reviewer, online reputation and information disclosure both increase the helpfulness of his/her reviews. The similarity between the reader and the reviewer has a positive effect on review helpfulness. Finally, the website credibility and replies to the review both make a review more helpful to the reader. To get new perspectives, further researchers should consider doing a qualitative research instead of quantitative. Finally, all

the theoretical knowledge should be condensed into a model or program and then be tested.

This chapter contributes to both theory and practice alike. I contribute to theory, by inductively creating a theoretical framework trying to understand the context of online reviews, their factors and characteristics. Furthermore, through the application of 1) review-related, 2) reviewer-related, 3) reader-related and 4) environment-related factors, I offer a conceptualization of what and in which dimension the factor contributes to the perceived helpfulness of an online review. Moreover, I developed a new visual representation form to present analyzed topics in a structured literature reviews. I contribute to practice by providing insights in how to design a review system, guiding reviewers to write helpful reviews and new ways of how to display the most helpful reviews.

##### **4.4.1 Limitations and Outlook**

One limitation of this literature review is the selection process of the journal articles, the scope and the extraction of information. For every step, strict rules were predefined. Additionally, all steps were precisely documented. However, all decisions that were made when setting the rules have an influence on the outcome of the literature review. First, the limitation that journal articles published between 2011 to mid-2018 were included automatically excluded some journal articles. Second, journal articles, which were not published in a top journal, were excluded. Additionally, the chosen databases have an impact on the origin of the journal articles. This also explains why most journal articles that were used in the final analysis are from the research field of information systems. We used the new invented visualization method called co-topic analysis in our literature review. Further research needs to be carried out showing the generalizability of the used method. Our next step includes training a machine-learning algorithm for automatically extracting topics from research paper and creating the visualization.

Future research regarding the helpfulness of online reviews should focus on reader-related factors and environment-related factors because the co-topic analysis showed that these two dimensions got less attention in the previous research. E.g. a topic that was greatly disregarded so far is the influence of

social media. In the digital age, sharing of information through social media is of great importance (Aggarwal 2016). People share their opinions in these like-minded communities and it is hard to separate fake from real. Therefore, researchers should investigate factors, which are related to digital sharing in social networks and find ways to predict the helpfulness of reviews that are shared on social media. Additionally, more in-depth knowledge regarding the effects of review-related factors would be helpful to solve the problem of mixed findings. To complete the ‘big picture’, the research needs more connections between the individual factors in general.

## 4.5 Conclusion of this Chapter

The aim of the research presented in this chapter has been to evaluate related research on how to measure peer feedback quality. I chose to review the literature on online reviews because of their strong similarity to peer feedback. Both are similar in terms of several features: the assembly of the *system*, the *structure* of the review/feedback and the evaluation of their *perceived quality*. I carried out a systematic literature review in which I analysed 81 journal articles in depth to determine which were the most important characteristics determining online review helpfulness.

The analysis revealed that review- and reviewer-related factors, especially ratings and reputation, were most researched. Furthermore, both factors were also mostly researched together. With regard to review-related factors, ratings and depth were often investigated within the same research study. Using these findings and integrating them to the research literature on peer feedback provide interesting suggestions for future research into which characteristics are important when measuring peer feedback quality. For example, online reviews and peer feedback are highly similar with regard to review-related-factors such as rating of the review, number of positive/negative statements, length, readability, number of arguments etc. Posting time is rather unique to online reviews and do not play a role in peer feedback as there is usually a set time-frame for writing the feedback.

However, review-related factors are factors that usually not often researched with regard peer feedback. In peer feedback usually the student receives an

anonymous feedback and does not know any characteristics of the reviewer. However, research into online reviews has revealed that this is an important factor. Thus, this provides valuable insights for future research into peer feedback. The feedback receiver could receive more information about the feedback provider such as how much feedback this person has already written, what the quality of their previous feedback was and what his or her perceived state of knowledge or skill level is with regard to providing high quality feedback.

Overall, the results of the systematic literature review have shown that in online reviews there is much more transparency in order to establish trust in the review system and the reviews. However, this system still maintains anonymity through pseudonymization. Hence, increasing transparency and providing additional information about the feedback provider to the feedback receiver could be of high value for peer feedback systems.

# 5

## **Measuring Peer Feedback Quality**

### 5.1 Introduction

In this chapter<sup>1</sup> I will present an in-depth evaluation of how to measure peer feedback quality. Future graduates need so called higher-order thinking skills - critical thinking, collaboration and communication - to be successful in the job market (OECD 2019). According to Hattie and Timperley (2007), there are two essential factors for training a higher-order thinking skill: a) a realistic teaching-learning scenario that allows learners to apply the skills that need to be trained and b) the provision of rich and meaningful feedback about the learners' progress. A major challenge in this regard is that most universities struggle in providing these two boundary conditions necessary for training higher-order thinking skills due to increasing numbers of students paired with financial resource constraints.

Consequently, we need to find solutions to help learners develop higher-order thinking skills in increasingly large classes that are likely characterized by an increasing lecturer-learner ratio. Prior research has highlighted the potential of peer learning approaches for creating a teaching-learning environment that is characterized by opportunities to apply different skills as well as the provision of feedback by fellow learners (Boud et al. 2014). Furthermore, previous research has shown that IT can help to create peer learning settings that can scale to hundreds or thousands of learners (Brinton et al. 2014; Piech et al. 2013).

Therefore, peer feedback in educational settings serves two important goals. First, it gives students a platform to train the skill of providing feedback in the learning environment and, thus, better prepare them for their future career. Second, it helps to address the problem of growing numbers of students. However, one challenge remains when aiming to create a teaching-learning scenario that is in line with the requirements posed by Hattie and Timperley (2007). Even though feedback is present in peer learning approaches, the quality of the feedback remains questionable, since it is provided by fellow students rather than an expert such as the lecturer. This may question whether this feedback is really “rich and meaningful”.

---

<sup>1</sup> A previous version of this chapter (Sections 5.2 to 5.7) has been submitted as Rietsche et al. to the Journal Information Systems Research. The manuscript is under review at the time of submission of this dissertation. The content has been reformatted and is presented in more detail.

Formative feedback in this dissertation is formed of four different types. Figure 2.2 in Section 2.3.2 shows these four types: *L1*, *L2*, *H1* and *H2*. *L* and *H* are related to the cognitive dimensions. *L* represents the lower-order thinking skills and *H* the higher-order thinking skills. The numbers *1* and *2* are related to the knowledge dimensions. Number *1* represents the factual, the conceptual and the procedural knowledge dimensions and number *2* represents the meta-cognitive knowledge dimension. The formative feedback used in this chapter is based on *H1* and *H2*.

Consequently, the aim of this chapter is to answer the third research question of this dissertation:

*RQ3: Which characteristics determine the quality of technology-mediated peer feedback?*

To answer that third research question, I sought to develop an approach that relied on the capabilities of modern technology to automatically estimate the perceived quality of a written piece of peer feedback. To achieve this goal, I gathered 7,036 examples of peer feedback from over the last four years in one of our large-scale classes. A typical peer feedback was approximately 250 words long, and the recipient of each peer feedback provided an assessment of the perceived feedback quality. Based on theoretical insights on feedback quality from the literature, I developed corresponding text features allowing us to automatically estimate different facets of quality of a given peer feedback.

The remainder of this chapter is structured along the *CRoss-Industry Standard Process for Data Mining* (CRISP-DM, Wirth and Hipp 2000). CRISP-DM provides a systematic approach for analysing large text data-sets. This process is iterative and consists of five main stages, which are presented in Figure 5.1 along with the corresponding sections of the chapter that describe them in detail.

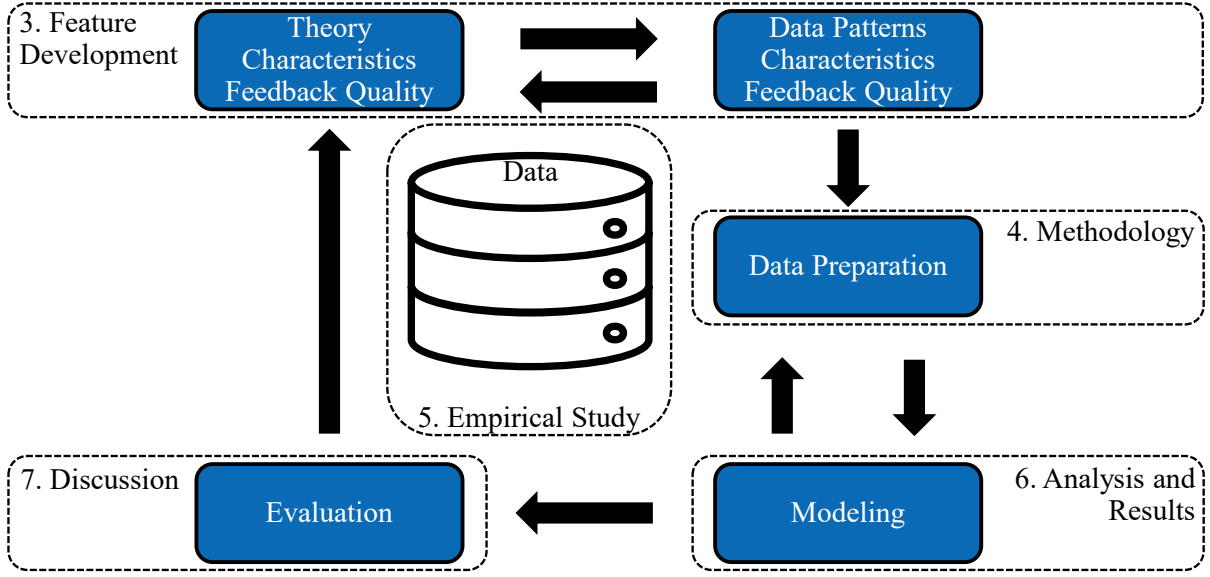
The first stage focuses on deductive derivation of the key feedback characteristics determining its quality, based on theoretical knowledge from prior research literature. In the second stage, I use the real-world data to inductively identify additional feedback characteristics affecting its quality. Both of these feature development stages are presented in Section 5.3. In my study, I use 7,036 cases of feedback accompanied with their quality evaluation ratings. The third stage operationalizes the set of created characteristics as features and prepares the

data for the consequent analysis. Section 5.4 describes the methodology I use for the operationalization of the features. In the fourth stage, I fit the regression models to identify the set of important features with a statistically significant impact on the feedback quality. The fifth stage evaluates the obtained results. In this chapter, Section 5.6 describes the analysis and results, while Section 5.7 provides the discussion of my findings. Iterating through the well-structured CRISP-DM process helps me to develop, analyze and refine the set of key features significantly affecting the quality of feedback.

## 5.2 CRoss-Industry Standard Process for Data Mining

There is no common agreement on applying text mining methods in the scientific community. However, one of the most popular processes used in research publications and applied in industry is CRISP-DM (Wirth and Hipp 2000). It offers a systematic approach to analyzing large text data sets. This process is iterative and consists of five main parts. We adapted it for our task and those parts are presented in Figure 5.1 along with the corresponding sections of the paper where we describe them in details. The first part is focusing on deriving deductively the key characteristics of feedback determining its quality based on theoretical knowledge of prior literature. In the second part, we use the real-world data to inductively identify additional feedback characteristics affecting its quality. Both feature development parts are presented in Section 5.3. In our study, we used 7,036 feedback instances accompanied with their quality evaluation ratings. The third part operationalizes the set of created characteristics as features and prepares the data for the consequent analysis. Section 5.4 describe the methodology we used for operationalization of the features. In the fourth part, we fit the regression models to identify the set of important features with a statistically significant impact on the feedback quality. The fifth part evaluates the obtained results. In our paper, Section 5.6 describes the analysis and results, while Section 7 provides the discussion of our findings. Iterating through the well-structured CRISP-DM process helped us to develop, analyze, and refine the set of the key features significantly affecting the quality of the feedback.



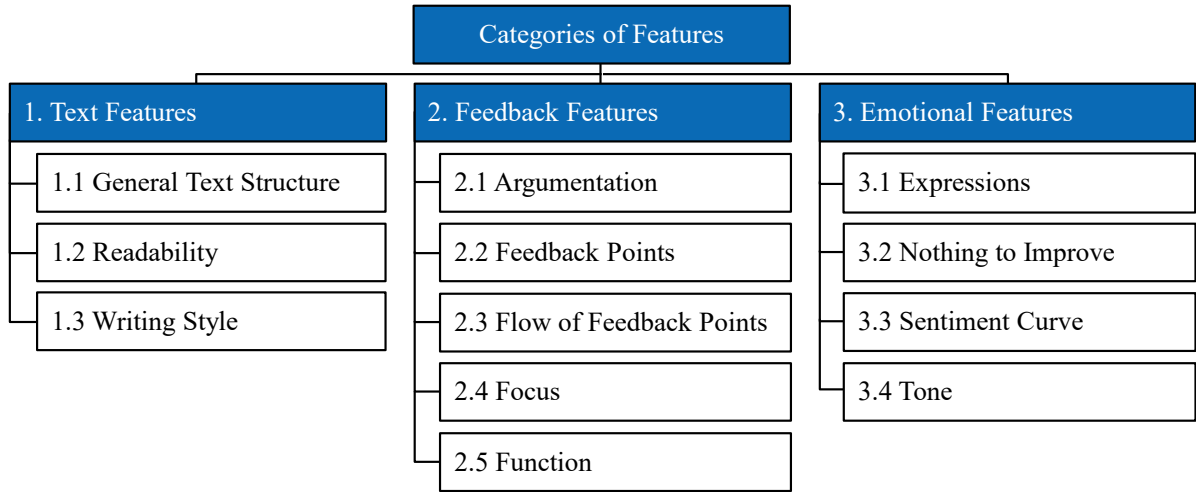


**Figure 5.1:** CRISP-DM process for Data Mining (CRISP-DM) adapted from Wirth and Hipp (2000)

### 5.3 Feature Development

Following the CRISP-DM process, in this section, we derive key feedback characteristics deductively based on the scientific literature and inductively by analyzing real world data. To organize the large set of all created features  $\mathcal{F}$ , we developed a three-level hierarchical structure  $\Omega$  that has categories, sub-categories, and groups of features. This hierarchy is adapted from the “Formative feedback guidelines” framework developed by Shute (2008). The first two levels of the proposed hierarchy are presented in Figure 5.2. In particular, we have three main categories: 1) Text Features (TF), 2) Feedback Features (FF) and 3) Emotional Features (EF). The first category contains features measuring the standard text characteristics, such as general structure and readability. The second category contains features focusing on specifics. The third category contains features measuring the emotional characteristics of the feedback that can affect the perception of its quality. We used the words “text” and “emotional” as collective terms representing the related research streams in text mining and NLP research.

In the following sections, we provide a detailed description of the considered features  $\mathcal{F}$ . We also explain the underlying theoretical basis behind each of them.



**Figure 5.2:** First Two Levels of the Proposed Feature Hierarchy  $\Omega$  Presenting Categories and Subcategories

A complete list of all the proposed features  $\mathcal{F}$ , their categories, sub-categories, and groups can be found in Table 5.1 provided at the end of the paper. The numbers of the features represent their location in the proposed hierarchy. For example, in the feature number 2.4.1.3, 2. represents its category (Feedback Features), 2.4 its subcategory (Focus), and 2.4.1 the feature’s group. Column 3 of Table 5.1 shows the type of *Measurement* that was used to calculate the feature, such as “Quantity”, “Average Quantity”, or “Weight”. The next column *Method* shows which method we applied for operationalizing the feature, such as “HTML”, “Dictionary”, or “TF-IDF”. In Section 5.4, we provide a description of all the measurements and methods we used in this study. Column five of Table 5.1 shows the *Effect* of the feature on the quality of the feedback. In case if it is significant, we added “positive” or “negative” indicating the direction of that effect, otherwise we added “-”.

### 5.3.1 Text Features

*Text features (1.)* contains standard features developed for general written text, such as text structure, sentences’ structure, and readability.

*General Text Structure (1.1)* features measure the various ways to structure the text, such as using sections, subsections, paragraphs, bullet points, lists, etc. (Alred et al. 2015). The main purpose of these elements is to provide easy access to the content and text reading (Dickson et al. 1995).

*Paragraphs (1.1.1)* have important functions in text organization. Namely, they encapsulate a unit of thought and provide a logical and visual break in the text (Alred et al. 2015).

*Paragraph length (1.1.2)* is an important indicator showing authors' ability to organize thoughts. Although there is no recommendation for the best number of sentences in a paragraph, the general rule is that it should be long enough to adequately deal with the subject of its main unit of thought. Text organization with a series of short paragraphs would probably break a single idea into many pieces. On the other hand, long paragraphs can fail to provide the reader with manageable subdivision of thoughts (Alred et al. 2015). In order to differentiate between titles, subtitles, and paragraphs in the actual content, we define three categories of paragraphs: *short*, which are no more than three words, *medium* which are no longer than one sentence, and *long* which are longer than one sentence (Alred et al. 2015). Short and medium paragraphs are usually titles, subtitles, or pose a question that is answered in the subsequent long paragraph. These structural elements are very important because they help readers to link the text to the abstract concepts in their minds (Dickson et al. 1995). Therefore, we propose to measure the numbers of paragraphs of each category. In addition, we also measure the total and the average length (number of sentences) of the long paragraphs, because these measures are important indicators of a good organization of the text content.

*Lists (1.1.3)* can also be used to organize texts (Stein 2003). The main purpose of bullet-points lists is to reduce cognitive load of the reader, by summarizing the most important points of text into tiny chunks (Alred et al. 2015).

*Readability (1.2)* is important for measuring ease of reading. On the one hand, the cognitive load caused by reading high quality text should stay reasonable (Kliegl et al. 2004). On the other hand, a text's complexity should match the reader's expectations (Weiß and Meurers 2018). For instance, a publication in an academic journal is expected to use more advanced grammatical constructions compared to a children's book.

*General (1.2.1)* readability group features include average word length and average sentence length which are the most important factors for measuring text complexity and readability (Berendes et al. 2017). Words' length affects

the eye-fixation ratio, meaning that longer words take the eye a longer time to capture (Kliegl et al. 2004); thus, it might be more difficult to comprehend longer words (Nagy et al. 1987). Similarly, sentences' length was also shown to be meaningful and reliable feature in many different languages (Feldman and Sanger 2007). Longer sentences contain more information; thus, the cognitive load for processing shorter sentences is lower and it can be easier to understand them (Bamberger and Vanecek 1984). Therefore, in addition to the general feature of average sentence length in this category, we also examine how this interacts with some other feedback features, such as author provided suggestions or directions.

*Scores (1.2.2)* are developed to provide a single measure of text readability. Kincaid et al. (1975) combined average word and sentence lengths into a single *Flesch-Reading-Ease* score. High values of this score mean that the text would be easily understood by pupils, while lower scores refer to higher complexity and shows that the text was written for only college graduates. One issue with this score is that it contains application-specific constants. Thus, researchers from other fields adapted this approach and designed new scores for their fields. For example, *Gunning Fog* index is focused on the business related texts (Gunning 1952). The scale of this score has the opposite direction, which means that higher values refer to greater complexity.

*Information density (1.2.3)* is the amount of information an author puts in a unit of text. The main parts of that information can be captured by searching for nouns and, therefore, density can be measured by normalizing the number of nouns by the total length of the text (Feldman and Sanger 2007). According to linguistics research, good text should have relatively high information density, but writing style with an excessively frequent use of nouns and noun groups can be considered complicated, stodgy and abstract (Schumacher 2017).

*Writing style (1.3)* is a sub-category containing the *Author Style (1.3.1)* group and is based on the idea that authors have unique styles of writing (Pennebaker and King 1999). In particular, these styles can be distinguished by the use of stop-words. For example, Marsden et al. (2013) used this approach to distinguish Shakespeare and Middleton.

### 5.3.2 Feedback Features

*Feedback Features (2.)* measures how feedback relates to the task or process of learning and fills the gap between what is understood and what is aimed to be understood (Sadler 1989). For generalizability purposes, we develop application independent feedback features. Derived from the body of feedback literature.

*Argumentation (2.1)* is an important element of the feedback that supports author's critiques. It can be given in many different ways (Shute 2008). One of the most important ways is to provide *Examples (2.1.1)* which can help the feedback receiver to imagine how the discussed comments could be implemented.

*Feedback Points (2.2)* measures the number and focus of main thoughts expressed in the feedback. Hattie and Timperley (2007) emphasized that feedback should answer the questions "How am I going" (Feed-back) and "Where to next" (Feed-forward). Hence, feedback should discuss the advantages and disadvantages of the submission and provide suggestions of how to improve it. Feedback can focus either on a single point, on the main points, or discuss every aspect of the authors work. If a piece of feedback focuses only on a single point, it might not capture all critical aspects; however, if it is too comprehensive, it might be too overwhelming (Nicol and Macfarlane-Dick 2006).

*Focus (2.2.1)* on the main issues is an important characteristic of high quality feedback. A large number of feedback points can lead to a higher cognitive load and might cause the receiver to reject the feedback (Nicol and Macfarlane-Dick 2006). In our study, we measure feedback focus by calculating its similarity to the submission in terms of the overlap of their important words and topics. A high overlap would indicate that the feedback discusses the most important points of the submission.

*Strength and Weaknesses (2.2.2).* Feedback can be positive or negative, indicating praise or criticism (Nicol 2010). Prior research has shown that there should be a balance between positive and negative feedback points (Weaver 2006). If the feedback contains only praise it may not be as helpful for improvement (Pitt and Norton 2017) as critical feedback, which is often perceived to be penetrating and useful (Yang and Carless 2013). However, if the feedback has too many negative points, it might be perceived as demotivating and will be rejected by the receiver (Hattie and Gan 2011).

*Feedback Length (2.2.3)* is an important factor affecting its quality. Longer feedback often suggests the feedback provided spent more time evaluating the work (Shute 2008). However, if the feedback is too long, then the receiver may not pay attention to it because the cognitive load for assimilating the provided information would be too great (Shute 2008).

*Suggestions General (2.2.4)* is a group of features that measures the amount of new information provided in the feedback. The main goal of feedback is to improve the work (Winstone et al. 2017) and suggestion on the ways to do it is the basic element of constructive feedback (Li et al. 2010). If feedback contains only critic it might not be perceived as helpful and cause negative emotions which would lead to a defensive state and overall rejection (Fong et al. 2016). An important aspect of suggestions is the way they are provided in a feedback. Therefore, we also created *Suggestions Pronoun (2.2.5)* which indicates whether the suggestion is provided in an I-form, You-form, or a mix. Similarly, *Suggestions Length (2.2.6)* measures the length of the sentence containing suggestions, and *Suggestion Sentiment (2.2.7)* examines if the sentiment of the suggestion affects the perceived quality of the feedback.

*Flow of Feedback Points (2.3)* evaluates the sequence of the discussed thoughts, which can follow one of the two common structures. Either a feedback is mixing positive and negative thoughts or separates them and provides one group after another. Prior literature suggest that effective feedback is built around addressing three questions (Hattie 2008): 1) What are the strengths? 2) What are the weaknesses? 3) What should be considered for improvement? Starting with a positive discussion of strengths reduces receivers' anxiety when reading the feedback (Weitzel 2000) and providing actionable suggestions on how to improve the work at the end of the feedback minimizes the risk that the receiver becomes defensive (Brookhart 2008).

*Verification and Elaboration (2.4)* measures whether feedback contains these two important components (Kulhavy and Stock 1989). The verification component indicates that the feedback contains statements claiming that something is right or wrong, whereas the elaboration component provides further information and more details about the issue. Prior literature indicates that high quality feedback should contain both components (Shute 2008). Furthermore, according to the framework of reflective thinking, explanations in a feedback show better

ability to elaborate (Sparks-Langer et al. 1990). An explanation can be justified by providing a reference (van Popta et al. 2017). Similar to the *suggestions* group, we also measure the way elaborations were provided in terms of pronouns used, sentence length and the attached sentiment.

*Function (2.5)* measures how well the feedback serves special functions stated by Shute (2008). There are two main pillars of these functions. First, feedback should be trustworthy. Trustworthiness, can be achieved by being unbiased, objective, and specifically listing the points that need to be improved. Second, feedback should be directive and facilitative. It should provide suggestions and ideas causing the receiver to think (Black and Wiliam 1998). The concept of directive and facilitative feedback is also present in the ways we speak. Speech act theory posits that statements are not limited to containing information; instead, some statements can also function as an action (Searle and Searle 1969). This results in one of three moods: indicative (stating facts), imperative (issuing orders), and subjunctive (making requests).

*Unbiased and Objective (2.5.1)* is an important characteristic of feedback (Shute 2008), which means that it should not be biased, i.e. be in favor of the reviewed work or disliking it (Black and Wiliam 1998). The feedback should be objective and should not reveal provider's personal attitude to the work.

*Directive (2.5.2-5)* feedback uses language containing commands and identifying what needs to be changed in the work. It provides clear steps to improve the work and usually uses the imperative mood (Searle and Searle 1969). This approach could be very helpful because of the clear list of actions, but the imperative mood and commanding language can be perceived as dictatorial. Similar to the *suggestions* group, we also measure pronoun use, sentence length, and the attached sentiment.

*Facilitative (2.5.6-9)* feedback tries to stimulate the receivers' mind by providing comments and suggestions to guide them in their own revision and conceptualization (Cunningham and Duffy 1996). The idea of facilitative feedback is to trigger self-regulated learning processes, helping receivers to think for themselves (Nicol and Macfarlane-Dick 2006). In contrast to directive feedback, facilitative feedback usually contains indicative and subjunctive moods. Note that feedback is not purely directive or facilitative, but rather exists along the

spectrum. Similar to the *suggestions* group, we also measure pronoun use, sentence length, and the attached sentiment.

*Asking Questions (2.5.10)* can be used to stimulate the feedback receivers in addition to facilitative words (Carless et al. 2011). Questions serve multiple functions, such as triggering reflective processes in the receiver's mind, or revealing that there is something missing in the work. Suggestions could also be provided in a form of questions.

### 5.3.3 Emotional Features

*Emotional Features (3.)* are composed of feedback characteristics that can trigger receiver's emotions. Categories 1. and 2. are mainly based on the existing knowledge from the literature. In this category, we also use the knowledge engineering approach to find patterns in the data and inductively create features based on the collected dataset.

*Expressions (3.1)* represent words or phrases that can set the feedback receiver to the positive emotional state and emphasize the provided statements. Those expressions can also reflect self-efficacy of the feedback provider (Bandura 1982).

*Salutation (3.1.1)* might be provided at the beginning of the feedback, whereas *Signing off (3.1.2)* occurs at the end. These phrases can put the receiver into an open mood and enable them to perceive the feedback as more helpful. Furthermore, salutations and sign offs might indicate an aspiration to build a relationship. Notably, the peer-feedback process in our setting is completely anonymous; students did not even know the gender of the receiver. However, this habit might appear as a part of a formal protocol when interacting with humans by text or short messages (Brown and Levinson 1987). In instant messaging, users utilize various salutations, such as "Hi", "Hello", and closings, such as "Bye", "Goodbye" (Spagnolli and Gamberini 2007). Salutation at the beginning of the feedback, might be a sign of trying to build up a relationship with the feedback receiver. Saying e.g. *Hello* can set the feedback receiver in an open stated mood, hoping that the feedback is seen as helpful rather than as a threat.



*Emoticons (3.1.3)* can be defined as textual or graphical symbols designed to mimic facial expressions (Derks et al. 2008). One of the difficulties in computer-based interactions, such as providing feedback, is the absence of body language and facial expression (Wang et al. 2014). Emoticons can help to deliver traditional nonverbal cues in these settings and alleviate this problem. The analysis of our dataset revealed that students used emoticons to support their positive feedback, arguments and suggestions. Emotions can support positive feedback or suggestions. However, positive emoticons such as :) with a negative feedback can cause confusion to the feedback receiver, since the intention of the feedback it is unclear (positive or negative).

*Nothing to Improve (3.2)*, in particular *Perfect Submission (3.2.1)* captures when the feedback states that the work does not need any changes. These pieces of feedback are usually very short and get low quality ratings. There are three main reasons for writing such feedback. First, the feedback provider may have wanted to finish the task as quickly as possible; providing only praises to the author and no critique takes less time than finding valid critique. Second, the feedback provider might lack subject matter knowledge or not have the ability to provide constructive critiques. In this case, the provider tries to conceal this issue by expressing only praise. Third, the submission may be entirely correct. This third option is only the case for tasks with a clear reference solution, not for creative tasks with no clear right or wrong answer.

*Sentiment Curve (3.3)* measures various characteristics in the sequence of positive, neutral, and negative segments in feedback. Prior research has shown that the sequence of these sentiments plays an important role in addition to the overall sentiment score of the text Mao and Lebanon (2007).

*Segment General Statistics (3.3.1)* includes the total number and average lengths of positive/negative sentiment segments. These features capture whether the feedback jumps excessively between positive and negative points, which can be difficult to interpret and thus place greater cognitive load on the receiver. Clustering positive and negative points together can thus help to encode the feedback. Further, these features show whether the feedback elaborates positive or negative points.

*First and Last segments' sentiments (3.3.2)* describes the positivity or negativity of feedback's opening or closing segments. A positive sentiment expressed at the beginning of the feedback can have a priming effect for the overall perceived sentiment (Myers 2009), triggering positive emotions and associations to later feedback (Lohse and Overgaard 2019). Similarly, a negative segment at the end of the feedback could color the feedback and set the receiver into a negative mood. This can reduce the perceived helpfulness of the feedback.

*First segments of each sentiment polarity (3.3.3)* measures the length and the position of the first positive and negative segments in the feedback. Similarly, the next group evaluates the *Last segments of each sentiment polarity (3.3.4)*.

*Tone (3.4)* sub-category captures the words used to express critique or suggestions that can have an enormous influence, on how the feedback receiver emotionally absorbs feedback. The proverb "It is not *what* you say, but *how* you say it" could not be more important than for the written feedback.

*Motivation (3.4.1)* provided in the feedback would encourage the feedback receiver to improve the submission and implement the provided suggestions (Dawson et al. 2016). Although, it is not always the main purpose of a feedback, but it makes the feedback receiver feel better and may affect the perceived quality of the feedback.

*Self-Esteem (3.4.2)* is seen as an important personality variable having an impact on the feedback acceptance (Ilgen et al. 1979). Positive and negative sentiment within the "I" or "you" perspective may influence the receiver's self-esteem, which is the integrated sum of self-confidence and self-respect, and describe the belief that one is competent and worthy (Mruk 2013). Positive self-esteem can cause positive outcomes, such as academic achievement or well-being (Baumeister et al. 2003). Therefore, negative feedback on the self level of the receiver can threaten their self-esteem and lead to worse performance (Kluger and DeNisi 1996). This group of features measures the various combinations of pronouns and sentiments used in the feedback.

**Table 5.1:** List of all Proposed Features  $\mathcal{F}$  Organized into Categories and Subcategories  $\Omega$

| Category, Subcategory, Group | Feature Name                                 | Measurement      | Method            | Effect   |
|------------------------------|----------------------------------------------|------------------|-------------------|----------|
| 1. Text Features             |                                              |                  |                   |          |
| 1.1 General Text Structure   |                                              |                  |                   |          |
| 1.1.1 Paragraphs             | 1.1.1.1 Paragraph tag                        | Quantity         | HTML              | Positive |
|                              | 1.1.1.2 Line break tag                       | Quantity         | HTML              | Positive |
| 1.1.2 Paragraph Length       | 1.1.2.1 Short (up to 3 words)                | Quantity         | HTML              | -        |
|                              | 1.1.2.2 Medium (1 sentence)                  | Quantity         | HTML              | Positive |
|                              | 1.1.2.3 Long (more than one sentence)        | Quantity         | HTML              | Positive |
|                              | 1.1.2.4 Sentences in Long                    | Quantity         | HTML & Tokenizer  | Positive |
|                              | 1.1.2.5 Sentences in Long                    | Average Value    | HTML & Tokenizer  | Positive |
| 1.1.3 Lists                  | 1.1.3.1 Bullet point list tag                | Quantity         | HTML              | Negative |
|                              | 1.1.3.2 Bullet point list e.g. hyphen symbol | Quantity         | Dictionary (int.) | Positive |
|                              | 1.1.3.3 Entries per bullet point list        | Quantity         | HTML              | Positive |
| 1.2 Readability              |                                              |                  |                   |          |
| 1.2.1 General                | 1.2.1.1 Word length                          | Average Value    | Tokenizer         | -        |
|                              | 1.2.1.2 Sentence length                      | Average Value    | Tokenizer         | Positive |
| 1.2.2 Scores                 | 1.2.2.1 Flesch-Reading-Ease German           | Level            | Formula           | Negative |
|                              | 1.2.2.2 Gunning Fog                          | Level            | Formula           | Positive |
| 1.2.3 Information Density    | 1.2.3.1 Noun chunks                          | Quantity (norm.) | POS               | Positive |
|                              | 1.2.3.2 Nouns and proper nouns               | Quantity (norm.) | POS               | Positive |
| 1.3 Writing Style            |                                              |                  |                   |          |
| 1.3.1 Author Style           | 1.3.1.1 Stop words                           | Quantity (norm.) | Dictionary (ext.) | Positive |
| 2. Feedback Features         |                                              |                  |                   |          |
| 2.1 Argumentation            |                                              |                  |                   |          |
| 2.1.1 Examples               | 2.1.1.1 Phrases indicating examples          | Quantity (norm.) | Dictionary (int.) | Positive |

| Category, Subcategory, Group | Feature Name                                                     | Measurement             | Method            | Effect   |
|------------------------------|------------------------------------------------------------------|-------------------------|-------------------|----------|
| 2.2 Feedback Points          | 2.1.1.2 Topic “examples“                                         | Weight                  | LDA               | Positive |
|                              | 2.2.1 Focus                                                      |                         |                   |          |
|                              | 2.2.1.1 Similarity submission and feedback                       | Weight                  | TF-IDF            | Positive |
|                              | 2.2.1.2 Similarity top 5% topics submission and feedback         | Weight                  | LDA               | Positive |
|                              | 2.2.2 Strength and Weaknesses                                    |                         |                   |          |
|                              | 2.2.2.1 <i>Positive</i> Sentiment                                | Sentiment score         | SentiWS           | Positive |
|                              | 2.2.2.2 <i>Negative</i> Sentiment                                | Sentiment score         | SentiWS           | Negative |
|                              | 2.2.2.3 Ratio between <i>positive</i> and <i>negative</i> words  | Percentage              | Dictionary (ext.) | Negative |
|                              | 2.2.3 Feedback Length                                            |                         |                   |          |
|                              | 2.2.3.1 Total number of sentences                                | Quantity                | Tokenizer         | Positive |
|                              | 2.2.3.2 Ratio feedback length and submission length in sentences | Quantity                | Tokenizer         | -        |
|                              | 2.2.4 Suggestions General                                        |                         |                   |          |
|                              | 2.2.4.1 Important words in feedback, not in submission           | Quantity                | TF-IDF            | Negative |
|                              | 2.2.4.2 New topics in feedback, which are not in submission      | Quantity                | LDA               | Positive |
|                              | 2.2.4.3 Sentence containing suggestion phrases                   | Sentence Binary (norm.) | Dictionary (int.) | Positive |
|                              | 2.2.5 Suggestions Pronoun                                        |                         |                   |          |
|                              | 2.2.5.1 Sentence suggestion and exclusive <i>I</i>               | Sentence Binary (norm.) | Dictionary (int.) | -        |
|                              | 2.2.5.2 Sentence suggestion and exclusive <i>You</i>             | Sentence Binary (norm.) | Dictionary (int.) | Positive |
|                              | 2.2.5.3 Sentence suggestion and <i>I &amp; You</i>               | Sentence Binary (norm.) | Dictionary (int.) | -        |
|                              | 2.2.6 Suggestions Length                                         |                         |                   |          |
|                              | 2.2.6.1 <i>Short</i> sentence with suggestion                    | Sentence Binary (norm.) | Dictionary (int.) | -        |
|                              | 2.2.6.2 <i>Long</i> sentence with suggestion                     | Sentence Binary (norm.) | Dictionary (int.) | Positive |

| Category, Subcategory, Group                 | Feature Name                                               | Measurement             | Method               | Effect   |
|----------------------------------------------|------------------------------------------------------------|-------------------------|----------------------|----------|
| 2.2.7 Suggestion Sentiment                   | 2.2.7.1 <i>Positive</i> sentence with suggestion           | Sentence Binary (norm.) | Dict(int.) & SentiWS | -        |
|                                              | 2.2.7.2 <i>Negative</i> sentence with suggestion           | Sentence Binary (norm.) | Dict(int.) & SentiWS | -        |
| <b>2.3 Flow of Feedback Points</b>           |                                                            |                         |                      |          |
| 2.3.1 Three Questions                        | 2.3.1.1 Organized in positive, negative and improvements   | Binary                  | Dictionary (int.)    | Positive |
| <b>2.4 Verification and Elaboration</b>      |                                                            |                         |                      |          |
| 2.4.1 Verification and Elaboration General   | 2.4.1.1 Sentence containing verification (positive)        | Sentence Binary (norm.) | Dictionary (int.)    | -        |
|                                              | 2.4.1.2 Sentence containing verification (negative)        | Sentence Binary (norm.) | Dictionary (int.)    | -        |
|                                              | 2.4.1.3 Sentence containing elaboration phrases            | Sentence Binary (norm.) | Dictionary (int.)    | Positive |
| 2.4.2 Verification and Elaboration Pronoun   | 2.4.2.1 Sentence elaboration and exclusive <i>I</i>        | Sentence Binary (norm.) | Dictionary (int.)    | -        |
|                                              | 2.4.2.2 Sentence elaboration and exclusive <i>You</i>      | Sentence Binary (norm.) | Dictionary (int.)    | Positive |
|                                              | 2.4.2.3 Sentence elaboration and <i>I &amp; You</i>        | Sentence Binary (norm.) | Dictionary (int.)    | Positive |
| 2.4.3 Verification and Elaboration Length    | 2.4.3.1 Sentence elaboration and is <i>short</i>           | Sentence Binary (norm.) | Dictionary (int.)    | -        |
|                                              | 2.4.3.2 Sentence elaboration and is <i>long</i>            | Sentence Binary (norm.) | Dictionary (int.)    | Positive |
| 2.4.4 Verification and Elaboration Sentiment | 2.4.4.1 Sentence elaboration and <i>positive</i> sentiment | Quantity (norm.)        | Dict(int.) & SentiWS | Negative |
|                                              | 2.4.4.2 Sentence elaboration and <i>negative</i> sentiment | Quantity (norm.)        | Dict(int.) & SentiWS | Negative |

| Category, Subcategory, Group | Feature Name                                             | Measurement             | Method                | Effect   |
|------------------------------|----------------------------------------------------------|-------------------------|-----------------------|----------|
| <b>2.5 Function</b>          |                                                          |                         |                       |          |
| 2.5.1 Unbiased & Objective   | 2.5.1.1 Feedback provider likes the submission           | Binary                  | Dictionary (int.)     | Positive |
| 2.5.2 Directive General      | 2.5.2.1 Sentence containing directive phrases            | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
| 2.5.3 Directive Pronoun      | 2.5.3.1 Sentence directive and exclusive <i>I</i>        | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
|                              | 2.5.3.2 Sentence directive and exclusive <i>You</i>      | Sentence Binary (norm.) | Dictionary (int.)     | -        |
|                              | 2.5.3.3 Sentence directive and <i>I &amp; You</i>        | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
| 2.5.4 Directive Length       | 2.5.4.1 Sentence directive and is <i>short</i>           | Sentence Binary (norm.) | Dictionary (int.)     | -        |
|                              | 2.5.4.2 Sentence directive and is <i>long</i>            | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
| 2.5.5 Directive Sentiment    | 2.5.5.1 Sentence directive and <i>positive</i> sentiment | Quantity (norm.)        | Dict.(int.) & SentiWS | Positive |
|                              | 2.5.5.2 Sentence directive and <i>negative</i> sentiment | Quantity (norm.)        | Dict.(int.) & SentiWS | -        |
| 2.5.6 Facilitative General   | 2.5.6.1 Sentences containing facilitative phrases        | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
| 2.5.7 Facilitative Pronoun   | 2.5.7.1 Sentence facilitative and exclusive <i>I</i>     | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
|                              | 2.5.7.2 Sentence facilitative and exclusive <i>You</i>   | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
|                              | 2.5.7.3 Sentence facilitative and <i>I &amp; You</i>     | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
| 2.5.8 Facilitative Length    | 2.5.8.1 Sentence facilitative and is <i>short</i>        | Sentence Binary (norm.) | Dictionary (int.)     | -        |

| Category, Subcategory, Group                    | Feature Name                                                | Measurement             | Method                | Effect   |
|-------------------------------------------------|-------------------------------------------------------------|-------------------------|-----------------------|----------|
|                                                 | 2.5.8.2 Sentence facilitative and is <i>long</i>            | Sentence Binary (norm.) | Dictionary (int.)     | Positive |
| 2.5.9 Facilitative Sentiment                    | 2.5.9.1 Sentence facilitative and <i>positive</i> sentiment | Quantity (norm.)        | Dict.(int.) & SentiWS | Positive |
|                                                 | 2.5.9.2 Sentence facilitative and <i>negative</i> sentiment | Quantity (norm.)        | Dict.(int.) & SentiWS | -        |
| 2.5.10 Asking questions                         | 2.5.10.1 At least one question mark                         | Binary                  | Dictionary (int.)     | Positive |
|                                                 | 2.5.10.2 Question mark                                      | Quantity (norm.)        | Dictionary (int.)     | Negative |
| <b>3. Emotional Features</b>                    |                                                             |                         |                       |          |
| <b>3.1 Expressions</b>                          |                                                             |                         |                       |          |
| 3.1.1 Salutation                                | 3.1.1.1 Salutation at the beginning                         | Binary                  | Dictionary (int.)     | Positive |
| 3.1.2 Signing-Off                               | 3.1.2.1 Signing-Off at the end                              | Binary                  | Dictionary (int.)     | Positive |
| 3.1.3 Emoticons                                 | 3.1.3.1 Emoticons                                           | Binary                  | Dictionary (int.)     | Positive |
| <b>3.2 Nothing to Improve</b>                   |                                                             |                         |                       |          |
| 3.2.1 Perfect submission                        | 3.2.1.1 Submission does not need to be improved             | Binary                  | Dictionary (int.)     | Negative |
| <b>3.3 Sentiment Curve</b>                      |                                                             |                         |                       |          |
| 3.3.1 Segment General Statistics                | 3.3.1.1 Number of segments                                  | Quantity                | Sentiment Curve       | Positive |
|                                                 | 3.3.1.2 Segment length                                      | Average Value           | Sentiment Curve       | Positive |
|                                                 | 3.3.1.3 <i>Positive</i> segment length                      | Average Value           | Sentiment Curve       | Positive |
|                                                 | 3.3.1.4 <i>Negative</i> segment length                      | Average Value           | Sentiment Curve       | -        |
| 3.3.2 First and Last segments' sentiments       | 3.3.2.1 Starts with <i>negative</i> sentiment               | Binary                  | Sentiment Curve       | Negative |
|                                                 | 3.3.2.2 Ends with <i>negative</i> sentiment                 | Binary                  | Sentiment Curve       | Negative |
| 3.3.3 First segments of each sentiment polarity | 3.3.3.1 Position of first positive segment                  | Sentence position       | Sentiment Curve       | Negative |
|                                                 | 3.3.3.2 Length of first positive segment                    | Quantity                | Sentiment Curve       | Positive |
|                                                 | 3.3.3.3 Position of first negative segment                  | Sentence position       | Sentiment Curve       | Negative |
|                                                 | 3.3.3.4 Length of first negative segment                    | Quantity                | Sentiment Curve       | -        |
| 3.3.4 Last segments of each sentiment polarity  | 3.3.4.1 Position of last positive segment                   | Sentence position       | Sentiment Curve       | Positive |
|                                                 | 3.3.4.2 Length of last positive segment                     | Quantity                | Sentiment Curve       | Positive |

| Category, Subcategory, Group                           | Feature Name                                                | Measurement             | Method            | Effect   |
|--------------------------------------------------------|-------------------------------------------------------------|-------------------------|-------------------|----------|
| <b>3.4 Tone</b><br>3.4.1 Motivate<br>3.4.2 Self-esteem | 3.3.4.3 Position of last negative segment                   | Sentence position       | Sentiment Curve   | -        |
|                                                        | 3.3.4.4 Length of last negative segment                     | Quantity                | Sentiment Curve   | Positive |
|                                                        | 3.4.1.1 Motivational words                                  | Quantity (norm.)        | Dictionary (int.) | Negative |
|                                                        | 3.4.2.1 Pronoun exclusive I and positive sentiment          | Sentence Binary (norm.) | Dictionary (int.) | -        |
|                                                        | 3.4.2.2 Pronoun exclusive you and positive sentiment        | Sentence Binary (norm.) | Dictionary (int.) | Positive |
|                                                        | 3.4.2.3 Pronoun I & you and positive sentiment              | Sentence Binary (norm.) | Dictionary (int.) | -        |
|                                                        | 3.4.2.4 Pronoun exclusive 3rd-person and positive sentiment | Sentence Binary (norm.) | Dictionary (int.) | -        |
|                                                        | 3.4.2.5 Pronoun others and positive sentiment               | Sentence Binary (norm.) | Dictionary (int.) | -        |



## 5.4 Research Methodology

In this section, we describe measurements and methods that were used to implement the features presented in Section 5.3. In particular, columns 3 and 4 of Table 5.1 show the measures used for each of the features.

### 5.4.1 Measurements

The first measurement, *Quantity*, calculates the number of occurrences of the entity in the text. *Quantity normalized* means that this number was normalized by the length of the text. *Average Value* calculates the mean value of the specified measure, such as words' length. *Binary* measurement shows if the text contains an instance of the concept, such as salutation. For the readability scores, we measure the *Level* of text readability. *Sentence Binary (norm.)* means that we calculate the number of sentences containing an instance of the concept (e.g., suggestions) divided by the total number of sentences. *Weight* indicates the calculated similarity weight or a weight of an LDA topic in the text. Finally, *Sentiment score* measures the sentiment of the text.

### 5.4.2 Methods

We designed the methods for calculating the proposed features  $\Omega$  aiming to make them easy to implement by other researchers and practitioners.

*Formulas* were used to calculate readability scores by *readability 0.3.1* python package<sup>2</sup>.

*Dictionary (ext.)* means that we use existing dictionaries from external sources. In particular, we took the list of stop-words from Stollenwerk (2013), and dictionary of positive and negative words from Linguistic Inquiry and Word Count (LIWC) developed by Tausczik and Pennebaker (2010).

*Dictionary (int.)* is our internally created dictionary that captures instances of the feature. We used the knowledge engineering approach (Feldman and Sanger 2007) for selecting relevant phrases and applied a strict procedure involving multiple experts to control the quality of the obtained results. In this process, we also follow CRISP-DM (Section 5.2) and implement five stages. At the

<sup>2</sup> <https://pypi.org/project/readability/>

first stage, experts use their domain knowledge to specify the initial list of the key phrases capturing the analyzed feature. Second, each of them reads a random sample of feedbacks and adds new phrases to the list based on their appearance in texts. At the third stage, we automatically analyze feedbacks to extend those manually created dictionaries. In our approach, we assume that phrases capturing the same feature would appear in a similar word contexts, such as “for instance” and “for example” both referring to the feature of “providing examples.” This is closely related to popular word2vec method (Mikolov et al. 2013). Therefore, we train the word2vec model to find similar phrases for the lists created in the first two stages. In particular, we trained one model based on our feedback dataset and the second model based on the corpus of political speeches (Barbaresi 2018). Next, for each of the phrases from the initial list we identify ten most similar phrases based on each of the models, thus creating an extended list. At the fourth stage of the process, each of the experts goes through the list of phrases and evaluates their relevance for the analyzed feature. In particular, for each of the phrases expert reads twenty examples of their appearances in the feedback and identifies if the analyzed feature is actually discussed in those sentences. Our guidelines suggest experts to select phrases that capture feature in at least 90% of the cases. At the fifth final stage of the process, we combine lists of phrases selected by the experts and select only those phrases that were labeled as relevant by a majority of votes. The described process helped us to create comprehensive dictionary for each of the features relying on this method.

*HTML* method means that we analyzed the HTML structure of the feedback to calculate the feature, such as the number of paragraphs  $\langle p \rangle$  or list items  $\langle li \rangle$ .

*LDA* (Latent Dirichlet Allocation, Blei et al. 2003) method is based on the LDA model trained on the texts of feedbacks, submissions and assignments. LDA is a generative probabilistic model of a corpus to calculate topics. The basic idea relies on, that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. In our study, we prepared those text by removing stop-words and applying lemmatization. We used MALLET 0.42 package (McCallum 2019) implemented by Gensim (2019) because it showed higher coherence score than the original implementation of

LDA in Gensim. We also experimented with a number of LDA topics and selected to use 200 based on the coherence score measure.

*SentiWS* refers to the sentiment analysis (Liu 2012) of the feedback produced based on a large corpus of words' sentiments for German language (Remus et al. 2010). A plethora of research, mostly on social media, has focused on analyzing tone which is in written text called sentiment (Hutto and Gilbert 2014). Some researchers use sentiment analysis and opinion mining interchangeably, however for simplicity we will use the term sentiment analysis (Liu 2012). Sentiment analysis is the process of capturing and analyzing peoples opinions, sentiment or emotions on a particular topic (Pang and Lee 2008).

*Tokenizer and POS* (Part-of-Speech) tagging was done by one of the world's fastest natural language processing tool spaCy 2.2 (Choi et al. 2015; Spacy 2019).

*TF-IDF* (Term Frequency-Inverse Document Frequency, Spärck Jones 1972) method was used identify the most important words of the feedback and measure its relevance to the submission. In TF-IDF, documents are represented as feature vectors. While a feature is defined as an entity without internal structure, located along a dimension in a feature space. One document in this feature-space consists thus of a sequence of features and their weights. To create a feature, most commonly the technique bag-of-words is used. This technique uses all words of a document as the features. Hence, the dimension of the feature space is equal to the number of different words in all of the documents.

There are multiple approaches to assign weights to the features. The most simple one is the binary approach, in which the feature weight is either one – if the corresponding word is present in the document – or zero otherwise. For the further details and the mathematical explanation, we refer to Feldman and Sanger (2007). In our study, we calculated IDF based on the Wikipedia dump (Wikimedia 2019) from April 2019. We selected this corpus because it contains all the frequent general words that are not important for our specific educational context.

*Sentiment Curve* method follows the ideas of local vs. global sentiment analysis by Mao and Lebanon (2007) and emotional arcs by Reagan et al. (2016). Both papers proposed to analyze the sequence of sentiments provided in the text for

various purposes. In our paper, we propose to study how various characteristics of the feedback sentiment curve affect its quality as perceived by the feedback receiver. In particular, we developed a method for constructing the sentiment curve of the feedback and use it to calculate the proposed list of features. Our method consists of the five following steps:

1. Calculate sentiment for each sentence and create the raw sentiment curve.
2. Remove noise from the raw sentiment curve by applying smoothing techniques.
3. Set a certain threshold  $\alpha$  for the absolute value of the sentiment and consider everything below or equal  $\alpha$  as being *neutral* sentiment.
4. Generate segments of consequent sentences with the same sentiment polarity.
5. Analyze the structure of the obtained sentiment segments and calculate various features.

*Step 1.* First, for each sentence in the text, we calculate its sentiment polarity score as a sum of its words' sentiments taken from the dictionary. There are various available sentiment dictionaries that could be used for these purposes, such as Stanford CoreNLP (Manning et al. 2014) or SentiWS for the German language (Remus et al. 2010). The sequence of obtained sentence sentiment scores constitutes the raw sentiment curve of the text.

*Step 2.* Next, we apply a smoothing technique to the raw sentiment curve in order to reduce the noise in it. The smoothing technique is based on a method used in stock market analysis (Lo et al. 2000) and sentiment flow analysis (Mao and Lebanon 2007). In particular, Lo et al. (2000) developed an automatic method for technical pattern recognition using non-parametric kernel regression. This method was originally developed with the goal to identify regularities in the time series of prices by extracting nonlinear patterns from noisy data. Therefore, they use kernel regression to smooth the curve and distinguish between the actual signal and the noise. Smoothing minimizes observational errors by averaging the data in sophisticated ways (Lo et al. 2000). Further, Mao and Lebanon (2007) used a similar idea of kernel regression to developed an algorithm for

predicting the global sentiment based on local sentiments. They created a model to represent the sentiment flow of a text.

We use the technique from these papers and apply kernel regression to get the smoothed sentiment curve of a text. Note that the bandwidth  $h$  is an important factor in kernel regression which defines the degree of smoothing. The low values of the bandwidth would lead to a choppy curve, while the high values would make the curve very smooth and lots of important information could be lost (Wand and Jones 1994). In our study, we tested different values of the bandwidth and got the best results with  $h = 1.4$ .

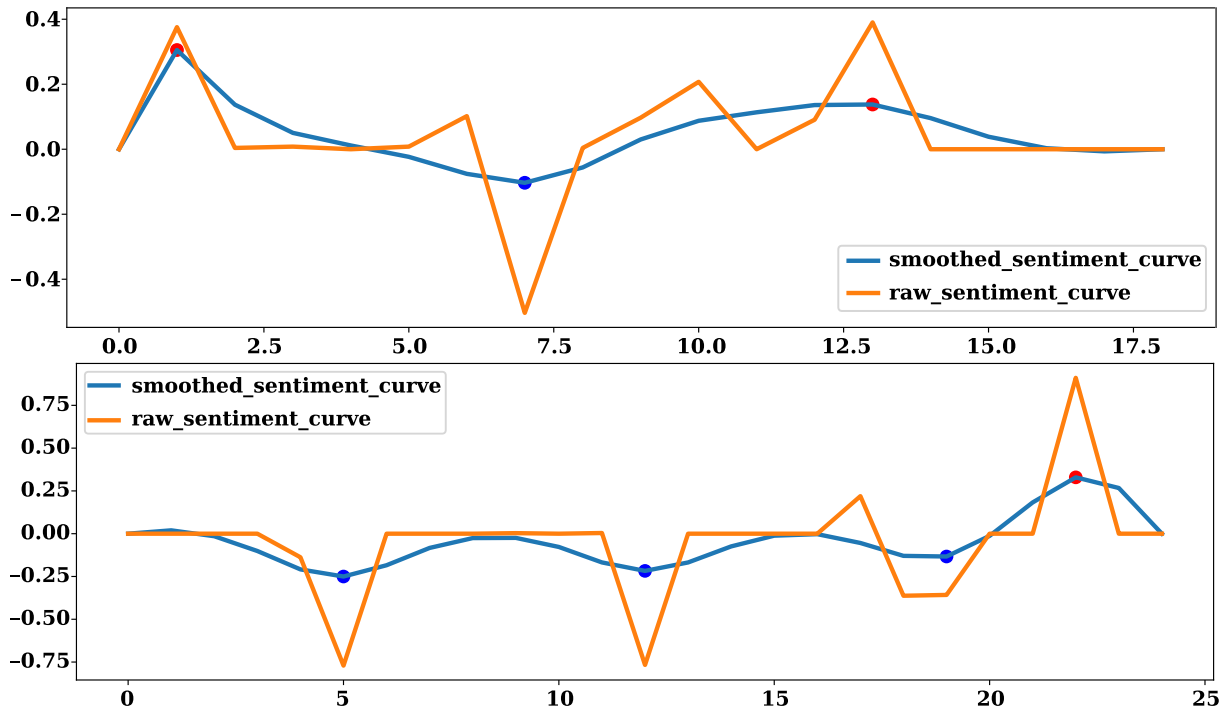
*Step 3.* We set threshold  $\alpha$  for the absolute value of the sentiment  $s$ . Next, we consider the polarity of all the sentiments  $s > \alpha$  to be *positive*, polarity of sentiments  $s < -\alpha$  to be *negative*, and everything else ( $-\alpha \leq s \leq \alpha$ ) to be *neutral*. This threshold  $\alpha$  controls the sensitivity of the sentiment curve analysis, because it defines which levels of sentiment would be ignored and considered as neutral.

*Step 4.* In this step, we analyze the sequence of sentiment polarities of text sentences and combine them into segments. Each of those segments contain all the consequent sentences with the same sentiment polarity. Thus, two consequent segments always have different polarities (positive, neutral, or negative).

*Step 5.* Finally, we analyze the sequence and structure of the obtained sentiment segments and calculate numerous features  $\mathcal{F}$  based on this analysis. In this paper, we propose the four groups of such features presented in Table 5.1.

The first group includes the features measuring the general statistics about the segments, such as the *Number of Segments*, *Average Segments Length*, *Average Length of Positive Segments*, *Average Length of Negative Segments*. Although, the length of a segment could be calculated in different ways, in our study, we use the number of sentences in a segment for these measures. The second group of features contains information about the polarity of the first and last non-neutral segments. It includes two binary features: *If Text Starts with Negative Segment*, and *If Text Ends with Negative Segment*. Further, the third group of features describes first segments of each sentiment polarity and contains the features, such as the *Position of the First Positive/Negative Segment*, and the *Length of the First Positive/Negative Segment*. The position of a segment is measured by

## 5. Measuring Peer Feedback Quality



**Figure 5.3:** Examples of Sentiment Curves: (top) *Positive-Negative-Positive-Neutral*; (bottom) three *negative* statements followed by one *positive*.

the ordinal number of its first sentence divided by the number of sentences in the text. Note that position ranges between 0 and 1. Similarly, the fourth group of features describes last segments of each sentiment polarity and includes features, such as the *Position of the Last Positive/Negative Segment*, and *Length of the Last Positive/Negative Segment*. In this case, the position is calculated using the ordinal number of the last sentence in the segment.

Figure 5.3 shows the result of running the first two steps of the proposed method based on the two pieces of feedback. The orange curve represents the initial raw sentiment curve (Step 1), and the blue curve shows the smoothed sentiment curve of a text (Step 2). The top part of the figure shows that the author of the text started with a positive statement, then mentioned something negative, then positive, and completed the text with the neutral statements. The second example shows the sentiment curve of a text that starts with three negative statements and has only one positive segment at the end. These sentiment segments are further used to calculate various features, such as number and length of segments, and “if feedback starts with negative.”

## 5.5 Empirical Study

In order to test how our features work in practice, we collected a unique data set containing textual feedbacks accompanied by their quality assessment ratings provided by the feedback receivers. In particular, we collected the data at a public university in Europe between 2015-2018. The analyzed graduate course was taught in German and focused on the development and design of business models. In total, we collected 7,036 feedbacks posted by 582 students (see Table 5.2).

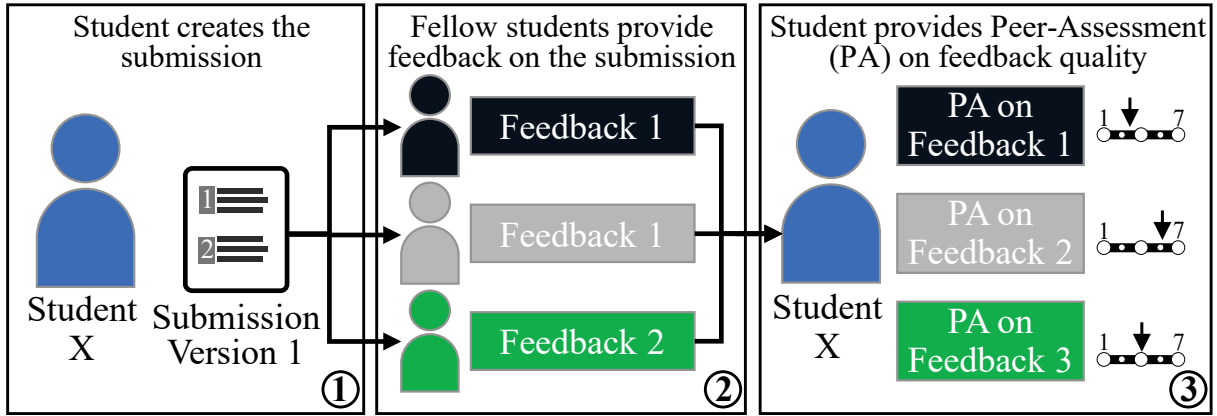
**Table 5.2:** Number of Reviews Gathered per Year

| <b>Time point</b> | <b>2015</b> | <b>2016</b> | <b>2107</b> | <b>2018</b> | <b>Overall</b> |
|-------------------|-------------|-------------|-------------|-------------|----------------|
| t1                | 152         | 246         | -           | 599         | <b>997</b>     |
| t2                | 145         | 255         | 448         | -           | <b>848</b>     |
| t3                | 170         | 297         | 539         | -           | <b>1006</b>    |
| t4                | 182         | 296         | -           | 608         | <b>1086</b>    |
| t5                | 170         | 265         | 546         | -           | <b>981</b>     |
| t6                | 155         | 265         | 557         | -           | <b>977</b>     |
| t7                | 129         | 240         | -           | 603         | <b>972</b>     |
| t8                | 169         | -           | -           | -           | <b>169</b>     |
| <b>Overall</b>    | <b>1272</b> | <b>1864</b> | <b>2090</b> | <b>1810</b> | <b>7036</b>    |

### 5.5.1 Peer Feedback Process

The implemented IT-based peer feedback (ITPF) process follows the standard peer-review process used in academic journals (Ziman 1974), where each of the papers assigned to the three anonymous reviewers. These reviewers evaluate the quality of the submission and provide feedback about strength and weaknesses of the paper and specify ways to improve it (Meadows 1998). The structure of our ITPF process is presented in Figure 5.4.

Each of the students participates in ITPF in two roles: first, as an author of the submission and the feedback receiver and, second, as a feedback provider. Students write and receive three feedbacks from their peers. The ITPF process



**Figure 5.4:** Structure of the IT-based Peer Feedback (ITPF) Process.

consists of the three main stages (Figure 5.4). At the first stage, student  $X$  creates the submission and uploads it to the system. At the second stage, our system anonymously distributes the submission to three peers and each fellow student provides feedback. In our study, we did not set any specific rules on the length or the content of the feedback. We only provided the three following guiding questions as assistance: 1. What are the strengths of the submission? 2. What are the weaknesses of the submission? 3. How could the submission be improved? At the third stage, student  $X$  receives the three feedbacks to the submission and provides peer assessment (PA) of their quality by answering four questions (items see Table 5.3).

The four questions used in PA are designed to evaluate the latent construct of the feedback quality  $q$ . In our study, high value (0.942) of Cronbach’s alpha verified high internal reliability of these questions. Next, we ran factor analysis and used the obtained scores as weights to calculate the aggregate value of quality  $q$ .

## 5.6 Analysis and Results

In this section, we analyze the effect of the proposed text features  $\mathcal{F}$  on the quality of the feedback  $q$  based on the unique data collected as described in Section 5.5. First, we present the evaluation methodology that we used in our analysis, and then we report the results in Section 5.6.2.



**Table 5.3:** Peer Assessment Items Measuring the Latent Construct of Feedback Quality  $q$  (Cronbach’s alpha = .942).

| Item | Question                                                                                    | Scale          | Source                 |
|------|---------------------------------------------------------------------------------------------|----------------|------------------------|
| 1    | The feedback I got from Reviewer X was helpful?                                             | Likert:<br>1-7 | (Strijbos et al. 2010) |
| 2    | The feedback I got from Reviewer X was high quality?                                        | Likert:<br>1-7 | (Li et al. 2010)       |
| 3    | The Reviewer X was able to identify critical aspects in my assignment?                      | Likert:<br>1-7 | (Li et al. 2010)       |
| 4    | The Reviewer X was able to provide constructive suggestions on his stated critical aspects? | Likert:<br>1-7 | (Li et al. 2010)       |

### 5.6.1 Evaluation Methodology

In our analysis, we used two approaches to evaluate the effect of the proposed textual features  $\mathcal{F}$  on the quality of the feedback  $q$ . First, we tested the significance of that effect for each feature  $f \in \mathcal{F}_\omega$  within its category of features  $\omega \in \Omega$ . In order to do that, we ran the regression analysis for each category  $\omega \in \Omega$ . We used OLS model with student-level fixed effects in order to take into account individual preferences of the students. As a dependent variable, we used the aggregated quality of the feedback  $q_{st}$  calculated based on the peer-assessment provided by the student  $s$  at the time  $t$  as described in Section 5.5. We used all the text features  $\mathcal{F}_\omega$  as independent variables. In particular, for each category  $\omega \in \Omega$  we trained the following model:

$$q_{st} = \sum_{f_i \in \mathcal{F}_\omega} a_i f_{i,st} + \alpha_s + u_{st} \quad (5.1)$$

where  $s = 1, \dots, n$ ,  $t = 1, \dots, T$ , and  $\alpha_s$  are the student-specific intercepts that capture heterogeneity across students, and  $u_{st}$  is the error term. We also ran separate models for various subgroups of the features from the category  $\omega \in \Omega$  in order to take into account their potential interrelationships. Further, we use the results of this regression analysis to identify the group of *important* features,

i.e. those that have statistically significant coefficients in the regression models (5.1).

To evaluate the effect of the proposed features  $\mathcal{F}$ , in the second approach we aim to identify the relative importance for the problem of predicting feedback quality. In particular, we train *Random Forest* regression model based on the set of all features  $\mathcal{F}$  to predict the quality of feedback  $q$ . Breiman (2001) shows that this model produces feature weights that can be used to measure their relative importance. In our analysis, we used 1000 trees and ran cross-validation to identify the best values of other parameters. As a result, we report the top of the most valuable features for the feedback quality prediction problem.

### 5.6.2 Results

The results of our comprehensive analysis are reported in Table 5.1 provided at the end of the paper. This table includes 90 text features  $\mathcal{F}$  and their categories  $\Omega$ . In particular, the last column “Effect” of Table 5.1 shows the direction of the feature’s effect on the quality of the feedback in case if its coefficient is significant ( $p - value < 0.05$ ). Overall, 66 of the developed features significantly affect the quality of the feedback as perceived by the receivers. This list includes various structural features applied on a general text level as well as on the sentence level. We examined the general text structure, readability, authors writing style, provided argumentation, feedback points, the focus of the feedback, function, emotional expressions, sentiment patterns, and tone. Out of all the studied features  $\mathcal{F}$ , we selected four representative groups from each of the categories to discuss the regression analysis and the obtained results in more details. We also present individual regression models for all features considered in Tables A1-A12 provided in Appendix A: All Tables of the OLS Regression.

**Table 5.4:** OLS Regression with Student-Level Fixed Effects for *1.2 Readability* Group of Features from *1. Text Features* Category (\* $p - value \leq 0.05$ , \*\* $p - value \leq 0.01$ , \*\*\* $p - value \leq 0.001$ ).

| <i>Feature Name</i>                    | <i>M1</i> | <i>M2</i> | <i>M3</i> | <i>M4</i> | <i>M5</i> |
|----------------------------------------|-----------|-----------|-----------|-----------|-----------|
| <b>1.2.1 General</b>                   |           |           |           |           |           |
| 1.2.1.1 Average Word length            | 0.079     |           |           |           | 0.024     |
| 1.2.1.2 Average Sentence length        | 0.023***  |           |           |           | 0.030***  |
| <b>1.2.2 Scores</b>                    |           |           |           |           |           |
| 1.2.2.1 Flesch-Reading-Ease German     |           | -0.006**  |           |           | 0.001     |
| 1.2.2.2 Gunning Fog                    |           |           | 0.021***  |           | 0.001     |
| <b>1.2.3 Information Density</b>       |           |           |           |           |           |
| 1.2.3.1 Noun chunks (norm.)            |           |           |           | 1.429*    | 2.059***  |
| 1.2.3.2 Nouns and proper nouns (norm.) |           |           |           | 0.884*    | 1.284**   |
| $R^2$                                  | 0.006     | 0.002     | 0.002     | 0.004     | 0.014     |

The first subcategory of features -“Readability”- contains three groups and six text features in total. The results of the regression analysis for this subcategory are reported in Table 5.4. Column 1 shows the names of the features and the rest of the columns contain coefficients for the corresponding models. The analysis of the first group shows that the average length of the sentence has strong positive effect on the perceived quality of the feedback, while the effect of the average length of words is not significant. The second group includes two standard readability scores. The negative effect of the Flesch-Reading-Ease score shows that receiver of the feedback would value the complexity of the text and would not give a high-quality rating for the feedback written in simple and basic terms. The effect of Gunning Fog is positive, because the scale of this score is opposite to the Flesch-Reading-Ease and higher values of this measure refer to higher complexity of the text. Thus, the significant effect of this feature confirms the need for a more eloquent language in the provided feedback. The third group of readability features measures information density by calculating the number of noun chunks, nouns and proper nouns normalized by the length of the feedback. Both features have a strong positive effect on the perceived quality of the feedback.

**Table 5.5:** OLS Regression with Student-Level Fixed Effects for the Features Related to *Suggestions* from 2. *Feedback Features* Category (\* $p - value \leq 0.05$ , \*\*  $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Feature Name</i>                                                             | <i>M1</i> | <i>M2</i> | <i>M3</i> | <i>M4</i> | <i>M5</i> |
|---------------------------------------------------------------------------------|-----------|-----------|-----------|-----------|-----------|
| <b>2.2.4 Suggestions General</b>                                                |           |           |           |           |           |
| 2.2.4.1 New words with high TF-IDF in feedback, which are not in the submission | -0.115*** |           |           |           | -0.115*** |
| 2.2.4.2 New LDA topics in feedback, which are not in the submission             | 0.161***  |           |           |           | 0.162***  |
| 2.2.4.3 Percent of sentences with suggestions                                   | 0.436***  |           |           |           | 0.402     |
| <b>2.2.5 Suggestions Pronoun</b>                                                |           |           |           |           |           |
| 2.2.5.1 Suggestion sentence with exclusive <i>I</i>                             |           | 0.278     |           |           | 0.270     |
| 2.2.5.2 Suggestion sentence with exclusive <i>You</i>                           |           | 0.709**   |           |           | 0.621     |
| 2.2.5.3 Suggestion sentence with <i>I</i> & <i>You</i>                          |           | 0.711     |           |           | 0.846*    |
| <b>2.2.6 Suggestions Length</b>                                                 |           |           |           |           |           |
| 2.2.6.1 <i>Short</i> sentence with suggestion                                   |           |           | 0.173     |           | 0.310     |
| 2.2.6.2 <i>Long</i> sentence with suggestion                                    |           |           | 0.669***  |           | 0.153     |
| <b>2.2.7 Suggestion Sentiment</b>                                               |           |           |           |           |           |
| 2.2.7.1 <i>Positive</i> sentence with suggestion                                |           |           |           | 0.120     | -0.381    |
| 2.2.7.2 <i>Negative</i> sentence with suggestion                                |           |           |           | 0.464     | -0.084    |
| $R^2$                                                                           | 0.091     | 0.003     | 0.002     | 0.001     | 0.093     |

The second example is a set of features focusing on *suggestions* provided in the feedback. This set consists of four groups of features belonging to the 2.2 *Feedback Points* subcategory of the 2. *Feedback Features* category. The results of the regression analysis for these features are reported in Table 5.5. The analysis of the first group (2.2.4) revealed some unforeseen results. Feature

2.2.4.1 counts the number of important words based on TF-IDF that appear in the feedback but not in the submission. This feature has a strong negative effect, which means that providing many of new terms that do not appear in the submission would decrease the perceived quality of the feedback. The next feature 2.2.4.2 calculates the number of new LDA topics appearing in the feedback which were not mentioned in the submission. In contrast to the previous one, this feature has a strong positive effect, which means that discussing a new topic in the feedback increases its quality. In the context of providing the feedback for the submission, the new topic is considered as a new suggestion. Although, “TF-IDF new words” and “LDA new topics” are related, they have the opposite effect on the quality of the feedback. We tested them in separate models and they preserved the opposite and significant coefficients. This result suggests that providing suggestions by discussing and elaborating new topics is of importance, while just mentioning many new terms may decrease the quality of the feedback. Further, the percentage of sentences containing suggestions as calculated by the words from the dictionary (feature 2.2.4.3) has a positive effect on feedback quality. Thus, feedback receivers generally appreciate suggestions on how to improve.

The analysis of the *Suggestions Pronoun* group shows that the most effective way to formulate your suggestion would be using “you” pronoun, e.g. “You can improve your submission by ...”. Further, *Suggestions Length* group shows that suggestions are more effective if they appear in longer sentences, which confirms the idea that they have to be elaborated. The sentiment of the sentences containing suggestions does not have strong effect on the quality.

The third example of the feature analysis we discuss in this paper is related to subgroup 3.2 *Nothing to Improve* of 3. *Emotional Features* category. This group contains one feature 3.2.1 which captures if the feedback contains that there is nothing that could be improved. The pattern of students writing that this submission is perfect and nothing needs to be improved caught our attention and the results are surprising. The significant negative coefficient  $-0.233^{***}$  suggests that feedback receivers demand for critiques and would not be satisfied with only the praises.

**Table 5.6:** OLS Regression with Student-Level Fixed Effects for *Sentiment Curve* Subcategory of Features from 3. *Emotional Features* Category (\* $p - value \leq 0.05$ , \*\*  $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Feature Name</i>                                    | <i>M1</i> | <i>M2</i> | <i>M3</i> | <i>M4</i> | <i>M5</i> |
|--------------------------------------------------------|-----------|-----------|-----------|-----------|-----------|
| <b>3.3.1 Segment General Statistics</b>                |           |           |           |           |           |
| 3.3.1.1 Number of segments                             | 0.121***  |           |           |           | 0.097***  |
| 3.3.1.2 Average segment length                         | 0.190***  |           |           |           | 0.207***  |
| 3.3.1.3 Average <i>positive</i> segment length         | 0.040**   |           |           |           | 0.037     |
| 3.3.1.4 Average <i>negative</i> segment length         | 0.013     |           |           |           | −0.078    |
| <b>3.3.2 First and Last segments</b>                   |           |           |           |           |           |
| 3.3.2.1 Starts with negative sentiment                 |           | −0.173**  |           |           | −0.032    |
| 3.3.2.2 Ends with negative sentiment                   |           | −0.230*** |           |           | −0.180    |
| <b>3.3.3 First segments of each sentiment polarity</b> |           |           |           |           |           |
| 3.3.3.1 Position of first positive segment             |           |           | −0.281*** |           | −0.220    |
| 3.3.3.2 Length of first positive segment               |           |           | 0.065***  |           | −0.005    |
| 3.3.3.3 Position of first negative segment             |           |           | −0.469*** |           | −0.200    |
| 3.3.3.4 Length of first negative segment               |           |           | 0.010     |           | 0.011     |
| <b>3.3.4 Last segments of each sentiment polarity</b>  |           |           |           |           |           |
| 3.3.4.1 Position of last positive segment              |           |           |           | 0.053***  | 0.006     |
| 3.3.4.1 Length of last positive segment                |           |           |           | 0.383***  | −0.019    |
| 3.3.4.3 Position of last negative segment              |           |           |           | 0.159     | 0.169     |
| 3.3.4.4 Length of last negative segment                |           |           |           | 0.030**   | 0.067     |
| $R^2$                                                  | 0.107     | 0.011     | 0.023     | 0.023     | 0.113     |

We report the results of the regression analysis for the *Sentiment Curve* subcategory presented in Table 5.6. These results show that the number of sentiment segments, their average length, and the average length of the positive segments have significant positive effects on the perceived quality of feedback. Furthermore, starting or ending the feedback with a negative segment significantly decreases that quality. The negative coefficients for the position features of the first positive and negative segments suggest that the author should decrease the number of neutral statements at the beginning of the feedback in order to appear higher in quality. Additionally, the last positive segment should be closer to the

end of the feedback; the position of the last negative segment does not have any effect.

Finally, Table 5.7 reports the top 20 features by their weights in the *Random Forest* regression model. The highest weight belongs to 2.2.3.1 *Total number of sentences* feature, showing that the length of the feedback is the best feature for differentiating between the good and bad feedback. This list also contains other length-related features, that are more specific than total length, such as number of sentences in long paragraphs (1.1.2.4) and average number of sentences in long paragraphs (1.1.2.5). The second important category in this top 20 list is *readability*, represented by features 1.2.1.2, 1.2.2.2, 1.2.3.2, 1.2.3.1, 1.2.2.1, and 1.2.1.1 discussed above. Besides the length of feedback and its readability, the Random Forest model identified other helpful features for the quality prediction task. The list also contains suggestion features (2.2.4.1 and 2.2.4.3), the match between feedback and submission (2.2.1.1 and 2.2.3.2), facilitative group of features (2.5.8.2 and 2.5.6.1), questions (2.5.10.2), sentiment (2.2.2.3), author style (1.3.1.1), provided examples (2.1.1.2), and sentiment curve feature 3.3.3.1. With these results, we were able to identify and operationalize lots of other concepts affecting the quality of the feedback as perceived by its receiver.

**Table 5.7:** Feature Weights in the Random Forest Regression Model

| #  | <i>Feature Name</i>                                                  | <i>Weight</i> |
|----|----------------------------------------------------------------------|---------------|
| 1  | 2.2.3.1 Total number of sentences                                    | 0.3418        |
| 2  | 1.2.1.2 Average Sentence length                                      | 0.0440        |
| 3  | 1.1.2.4 Number of Sentences in Long parts                            | 0.0322        |
| 4  | 2.2.1.1 Similarity submission and feedback                           | 0.0268        |
| 5  | 1.2.2.2 Gunning Fog                                                  | 0.0262        |
| 6  | 1.2.3.2 Nouns and proper nouns                                       | 0.0259        |
| 7  | 2.2.4.1 Important words in feedback, which are not in the submission | 0.0257        |
| 8  | 2.2.3.2 Ratio feedback length and submission length in sentences     | 0.0254        |
| 9  | 2.5.8.2 <i>Long</i> facilitative sentence                            | 0.0247        |
| 10 | 2.2.4.3 Percent of sentences with suggestions                        | 0.0179        |
| 11 | 2.5.10.2 Question mark                                               | 0.0151        |
| 12 | 2.5.6.1 Sentences containing facilitative phrases                    | 0.0137        |
| 13 | 2.2.2.3 Ratio between positive and negative words                    | 0.0136        |
| 14 | 1.3.1.1 Stop words                                                   | 0.0134        |
| 15 | 2.1.1.2 LDA Topic “examples”                                         | 0.0132        |
| 16 | 1.2.3.1 Noun chunks                                                  | 0.0132        |
| 17 | 1.2.2.1 Flesch-reading-ease German                                   | 0.0124        |
| 18 | 1.1.2.5 Average number of Sentences in Long parts                    | 0.0124        |
| 19 | 1.2.1.1 Average Word length                                          | 0.0119        |
| 20 | 3.3.3.1 Position of first positive segment                           | 0.0116        |

## 5.7 Discussion and Contributions

### 5.7.1 Contribution and Theoretical Implications

To cope with a world that is increasingly characterized by VUCA, companies started to adapt an agile mindset and work in small agile teams. This new



way of working leads to increasingly flat hierarchies in organizations and often require a more intense communication between employees at similar levels of seniority (Highsmith and Cockburn 2001). Hence, graduates working later in those agile teams, need to have so called higher-order thinking skills-critical thinking, collaboration and communication - to be successful on the job market (OECD 2019). To train these higher order thinking skills, two essential factors are needed a) a realistic teaching-learning scenario that allows learners to apply the skills that shall be trained and b) the provision of rich and meaningful feedback to the learners' progress.

Most universities struggle providing such boundary conditions, due the increasing number of students paired with financial resource constraints. One solution is IT based peer learning, where students can train higher order thinking skills such as communication and receive qualitative feedbacks from their peers. However, since the feedback is provided by fellow students and not by an expert, such as instructor, it might be questionable if the feedback is really "rich and meaningful". Thus, our paper studies the key characteristics of written feedback that can help to automatically assess its quality. We derived those text features deductively based on the literature and inductively from the data collected in real educational settings.

Researchers over the last decades have focused on conceptualizing feedback. Developing frameworks to capture its underlying circumstances, scholars attempted to answer questions such as: *who provides feedback*, *when is feedback provided*, *what is the feedback for* or *how effective is feedback* (Price et al. 2010). However, the literature addressing those questions often contains generic definitions of high and low quality feedback and the operationalization and automated measurement of feedback quality has been largely unexplored up to now. For example in one of the most cited works about feedback, Hattie and Timperley (2007) developed a model of feedback to enhance learning. The model contains abstract requirements, such as that a feedback should answer three questions, *Where am I going?*; *How am I going?*; *Where to next?*. Whereby each of the question can be answered on four levels the task, process, self-regulation and self level. However, certain measurements, for example whether the feedback is on the task or process level have not been developed. Nicol and Macfarlane-Dick (2006) defined seven principles of good feedback practice to develop self-

regulated learning. Those principles offer generic suggestions such as “deliver high quality information to students about their learning”.

However, scholars have yet to specifically address what constitutes high quality as perceived by the feedback receiver and how it can be measured? Shute (2008) conducted a comprehensive literature review on formative feedback. Outlining feedback guidelines to enhance learning such as “feedback to the learner should address specific features of his or her work in relation to the task, with suggestions on how to improve” (Shute 2008, p. 177). Moreover, neither of the cited studies in this literature review provides recommendation as to how to measure feedback quality. Similarly, although Brookhart (2008) provides comprehensive guidelines on how to give effective feedback to students, this paper does not offer empirical insight on how to operationalize those features and measure them.

Related subject areas, such as online reviews, offer analogous initial research on how to measure text quality. Most studies in this field investigated simple features, such as review length, readability and sentiment (Yin et al. 2013; Mudambi and Schuff 2010), but those features alone cannot comprehensively describe feedback quality as proposed by Shute (2008). Thus, in the domain of feedback as well as in related domains such as online reviews, comprehensive studies providing empirical evidence on how to operationalize features of high and low quality feedback are missing.

Addressing this gap, we contribute to theory by providing prescriptive knowledge on how to operationalize important features measuring peer feedback quality and testing them based on a unique real-world data set containing 7,036 peer feedbacks. These features were derived deductively from the prior literature and inductively from real-world data. Furthermore, we provide a framework including the hierarchical relationship of the features including their statistically significant effects on the perceived quality by the feedback receiver. Overall, our analysis identified 66 important features (out of our 90) that have a significant impact on feedback quality.

Our results confirmed that feedback length (2.2.3.1) is one of the most influential features for the feedback quality. Many other features depend on the feedback length, since only longer feedback can contain a comprehensive discussion of

strength and weaknesses, and can offer suggestions on how to proceed with revisions. Additionally, longer sentences containing suggestions and elaborations significantly impact feedback quality. Our analysis revealed further important features. For example, the negative effect of *nothing to improve* (3.2.1.1) feature shows that learners want to receive critical comments to improve their work, not only praise. A possible explanation for this might be rooted in the context of our study: a competitive business school. Students learn that competing is important and that they should strive to understand how to improve and become more successful. Another reason might be social investment: If learners invest a high amount of energy in creating their submission, they may expect the feedback provider to invest time in creating high-quality feedback.

Furthermore, our study revealed interesting results concerning the form in which feedback should be written: *I*, *you*, *I and you*, or *neutral-form*. The literature proposes that feedback should be written in the *I*-form to avoid being perceived as personal attack (Brookhart 2008). However, our results show that a more finely-tuned use of different forms can have great effect. For example, the *you-form* and *I and you-form* have a positive effect on perceived feedback quality when elaborating or providing suggestions; for directive words, only the *I-form* and *I and you-form* have an impact; and for facilitative words, these forms do not have any impact. This highlights the importance of showing that elaborations and suggestions are focused on the author and their submission. It is also vital to communicate that the feedback provider actually engaged with the submission in depth. However, directive language may push the feedback receiver to a defensive state of mind. Thus, it is important to use *I-form* or *I and you* to signal that it is not a personal attack but rather the feedback provider's opinion. Finally, for the sentiment analysis, we found a significant positive effect of only exclusive *you-form* and positive sentiment. This result shows the effect of personal praises for the feedback receiver.

### 5.7.2 Practical Implications

Ultimately, well-structured feedback can improve peers' work. Although feedback authors undoubtedly have several motivations, chief among them is likely the desire to write high quality feedbacks in order to support their peers. At a broader level, universities, companies, and MOOC providers can

utilize our findings when using peer learning concepts to train students higher-order thinking skills. Our research findings will be of special interest to those institutions when developing guidelines for students to provide feedback of a higher quality.

The automated assessment of feedback quality based on our proposed features serves at least three purposes. First, it can be displayed to the feedback provider during feedback creation, thereby immediately helping to improve the feedback before it is submitted. Second, it can be provided to the feedback receiver as additional information; if the feedback is rated low, additional information can be provided, such as “please read this feedback with caution.” Third, automated feedback quality evaluation can be provided to instructors to help identify students that need individual assistance on their feedback providing skills. Moreover, if the learner consistently receives feedback of lower quality than her peers, then the instructor can step in to provide his or her own feedback to this student. Thus, automated feedback quality evaluation can help the instructor optimize their time when targeting students for help.

### 5.7.3 Limitations and Future Research

Although we examined many different features of low and high quality feedback, using an in-depth and comprehensively approach based on a unique large real-world data set with over 7,036 feedbacks, our study does not come without limitations, which at the same time provide opportunities for future research. For example, additional factors such as language and cultural specifics may also have an influence on the perceived quality of the feedback. Although students from different countries attended our courses, cultures of those countries are quite consistent. Thus, in our study, we cannot provide insights on whether cultural differences impact features of low- or high-quality feedback. However, we expect that perceived peer feedback quality is different for individualistic and collectivistic learners. Learners from individualistic cultures, such as the United States, prefer more direct feedback that is particularly related to effort, whereas direct critique is more uncommon in collectivistic cultures (Hattie and Timperley 2007). Furthermore, readability and sentence complexity might differ across languages, but these generally defined features should still have a significant impact. Hence, future studies should include data from different

cultures. This would be particularly valuable for multi-national organizations, which increasingly rely on teams consisting of members from different countries. The same holds true for MOOC providers, which typically have a very international group of learners taking courses together.

## 5.8 Conclusion of this Chapter

The aim of the research presented in this chapter was to make the first step towards automatically evaluating the quality of peer feedback. I followed the CRISP-DM approach whereby I, first, deductively derived the key feedback characteristics determining its quality based on theoretical knowledge coming from the prior literature and, second, used real-world data to inductively identify additional feedback characteristics affecting its quality. Overall, this resulted in 90 features by which to measure the characteristics of peer feedback quality.

I used 7,036 examples of peer feedback that I collected over the last four years in one of our large-scale classes. I ran regression analyses in order to identify the most powerful text features for evaluating peer feedback quality. My results showed that 66 features had a significant effect on feedback quality.

I was able to confirm that feedback length is one of the top influential features determining feedback quality. This is similar to the results of the literature review on online reviews discussed in Chapter 4. However, my research found many more important features beyond length that determined peer feedback quality. Two examples are *nothing to improve* and certain *pronouns* that should be used for providing suggestions, elaborations, etc.



# 6

## **Cultural Influences in Peer Feedback**

### 6.1 Introduction

In this chapter<sup>1</sup> I present a study on developing design principles to support cultural sensitivity peer feedback in MOOCs. In contrast to formal education, there are courses in MOOCs with over 150,000 registered students (Allen and Seaman 2015). This leads to immense scalability challenges, especially as MOOCs do not only impart factual knowledge but also encourage learners to apply higher-order thinking skills. Thus MOOCs are a very good example of a large-scale class trying to support students in developing their higher-order thinking skills.

MOOCs use mostly peer evaluation and, consequently, peer feedback to support skill development. Peer evaluation is a prerequisite for peer feedback and describes the process when students evaluate the quality of their peer's submission. Peer feedback describes the process when students try to communicate the evaluation to their peers in the form of written text. Thus, the use of both approaches, peer evaluation and peer feedback, are promising as they not only overcome the scalability challenges in MOOCs, through using peers instead of often only one lecturer, but also enable the students to apply their higher-order thinking skills (Kulkarni et al. 2013).

However, research has long acknowledged that cultural differences can inhibit the successful use of IT and its acceptance (Leidner and Kayworth 2006). The case of using peer feedback in MOOCs, which directly connects students from various cultural backgrounds, is especially challenging as students have different culturally embedded learning preferences and educational beliefs, which might affect the acceptance of peer feedback.

For example, in China it is common to receive feedback, and in particular criticism, only from higher ranked people. In Switzerland there is more interdependence between less and more powerful people. To provide a better quality learning experience for all students, researchers and instructional designers of MOOCs are required to carefully study the learning and design preferences of

---

<sup>1</sup> A previous version of this chapter (Sections 6.2 to 6.5) has been published in Buholzer et al. (2018). The content has been reformatted and updated to the latest state of scientific discussion at the time of submission of this dissertation.



different cultures to deliver culturally sensitive instructions for the peer feedback process.

Thus, a culturally sensitive design is a very important foundation to the peer feedback in my research, since it addresses the problem of how students can provide high quality feedback to their peers. Therefore, this research provides valuable insights into how to design culturally sensitive peer feedback and thus the aim of this chapter is to answer the third research question of this dissertation:

*RQ3: Which characteristics determine the quality of technology-mediated peer feedback?*

To answer research question three, I use Design Science Research (DSR) and, in particular, follow the six-step DSR approach by Peffers et al. (2007). I derive requirements from the scientific literature and develop design principles for culturally sensitive peer feedback. I empirically evaluate these design principles through a qualitative and comparative study in the form of semi-structured interviews with students who had Swiss and Chinese cultural backgrounds.

The remainder of this chapter is structured as follows. The next section describes the research methodology and how I follow the DSR process to derive requirements and develop design principles for culturally sensitive peer feedback. The following section explains the problem identification process and design principles in more detail. The succeeding section presents the results of the semi-structured interviews. Followed by the next section that contains the discussion, limitations and ideas for future research. The final section provides a conclusion to this chapter.

## 6.2 Research Methodology

The main goal of this study is to develop and define design principles that support cultural sensitivity of the peer feedback in MOOCs. To do so, I followed the six-step DSR approach by Peffers et al. (2007). After the problem identification, the challenges in the current design by analyzing the MOOC platform Coursera<sup>2</sup> were identified and objectives of a solution derived. To do that three MOOC courses were analyzed in detail: Modern Art & Ideas<sup>3</sup>,

<sup>2</sup> [www.coursera.org](http://www.coursera.org)

<sup>3</sup> The Museum of Modern Art

Leading Innovations in Arts and Culture<sup>4</sup> and Coursera Mentor Community and Training Course<sup>5</sup>. These courses were chosen, based on its high popularity and number of registered students.

Five areas of peer feedback were defined, which eligible for cultural sensitive adaption and different design solutions for those areas that are subject to cultural preferences were created. These solutions are inferred in close reference to the literature, in particular Hofstede's cultural dimension theory (Hofstede and Hofstede 2005). For evaluating the design principles, 16 semi-structured interviews with 8 randomly selected participants from Switzerland and 8 randomly selected participants from China were conducted. To aim was to evaluate whether the developed design principles correspond to the learners' preferences in practice.

All participants had prior experiences with peer feedback. However, only 4 of the Chinese and 3 of the Swiss participants completed a peer feedback process in a MOOC. Therefore, prior the interviews, procedure of the peer feedback in Coursera were explained. This was important to ensure a consistent level of knowledge about the process. The age of the interviewees ranged from 19 to 27 (mean 24.19 years; mean Swiss 23.50 years; mean Chinese 24.88 years). I chose this approach for our inquiry as qualitative research designs have been shown to be particularly well suited to analyze dynamic, interactive processes (Lee et al. 1999).

Further, I decided to interview Chinese and Swiss students as the coherent cultures inhibit major differences and therefore offer a firm base for a comparative study. Prior the interviews, the participants were asked to fill out the Value Survey Model (VSM) developed by Hofstede (2013) to evaluate if the participants of both sample groups correlate with the official scores of their relating country. The VSM is a 30-item questionnaire for comparing culturally influenced values and sentiments of similar respondents from two or more countries. The Table 6.1 gives an overview of the VSM data of the participants of this study and the official scores published by Hofstede (2013) in parentheses.

The scores do not perfectly match with the official country scores published by Hofstede (2013). This might several explanations: First our sample group represent students, which is a very specific population group of these countries.

---

<sup>4</sup> Vanderbilt University

<sup>5</sup> Coursera Community Team

**Table 6.1:** Value Survey Model Results of the Study Participants and Official Scores by Hofstede (2013) in parentheses

| VSM Data                       | China      | Switzerland |
|--------------------------------|------------|-------------|
| Power Distance                 | 48,50 (80) | 16,88 (34)  |
| Individualism vs. Collectivism | 21,88 (20) | 43,75 (68)  |
| Masculinity vs. Femininity     | 52,50 (66) | 30,63 (70)  |
| Uncertainty Avoidance          | 61,88 (30) | 65,00 (58)  |
| Long Term Orientation          | 82,50 (87) | 70,00 (74)  |

Further, our sample groups are relatively small and the answers might also reflect other characteristics than culture, e.g. gender and age.

However, the results show the tendency and in particular the differences of cultural values between the Swiss and Chinese are visible. Afterward completing the VSM, all participants received: the interview guidance with explanations of the aim of this study; the aim of the interview; background information about MOOCs; and the peer feedback process in Coursera and the following 25 open-ended interview questions. The interview questions are divided into six categories. The first set of questions target to get a deeper understanding of the participant's general perception of peer feedback. The following five categories question on the participants' preferences regarding the prior defined design elements. During the interviews, I asked for both roles, feedback provider and feedback receiver. The interview ended with collecting information about the participant's current country of residence, age, gender, educational background and current nationality as well as nationality at birth. On average, the interviews took 30 minutes. For the analysis I prepared written transcripts of all interviews and analyzed the text data, seeking correlations in descriptions of the participants' preferences regarding the peer feedback process in MOOCs.

### 6.3 Towards Designing and Evaluating Peer Feedback Design Principles

Following the DSR process, this section contains the first activity which emphasizes the *problem identification*, the second activity containing the *definition*

*of objectives* for a solution and the third activity containing the *design and development of artifacts*. The Sections 6.4 and 6.5 contain the activity four which is *demonstration* and activity five which is *evaluation*.

### 6.3.1 Problem Identification

MOOCs attract students from all over the world. However, so far MOOC providers do not consider cultural differences of course participants. Research has long acknowledged that cultural differences can inhibit the successful use of information technology (Leidner and Kayworth 2006). To prevent learner rejection and fluctuation, MOOC provider need to consider cultural differences. This is especially important during peer feedback, as this process directly connects students from various cultural backgrounds.

### 6.3.2 Definition of Objectives

During typical peer feedback in Coursera, students are requested to complete an assignment and submit it online. Each assignment is then distributed to three randomly selected fellow students. Each fellow learner then rates the assignment based on pre-specified scoring rubrics and is requested to provide further feedback in the form of written comments. The mean score of all ratings as well as the written comments are then made available to the learner who submitted the assignment (Suen 2014). The multicultural nature of MOOCs leads to individual and collective enrichment. However, cultural differences among the students also provide challenges for instructional designers.

First, the *acceptance of feedback*: Referring to the dimension of power distance, it is questionable to what extent students accept the received feedback from their peers, especially if the feedback involves criticism. In cultures with high power distance, students expect the teacher to outline paths. Status is important and hierarchies are well accepted (Hofstede and Hofstede 2005). At the moment, Coursera does not offer any information to students about their peer reviewers. However, especially students from cultures with high power distance might prefer to have more information.

Second, *group dynamics*: Group dynamics appear very different in collectivistic and individualistic cultures. Whereas collectivistic cultures put a strong empha-

sis on social networks, harmony and the preference of small group belongings, individualistic cultures tend to focus on individuals (Hofstede and Hofstede 2005). During the peer feedback process in Coursera the learner's assignment is randomly distributed to three different reviewers. If the learner has to complete multiple assignments involving peer feedbacks, the reviewers will change during the course. The current allocation process neglects collectivistic values like the preference of small group belongings.

Third, *competition and comparison between students*: Whereas feminine cultures tend to avoid situations, which distinguish clear winner and losers, masculine cultures put emphasis on achievement and success (Hofstede and Hofstede 2005). The results of the peer feedback in Coursera are shown as percentage. However, there is no information about the performance scale of the course. Competitive students have no opportunity to set their performance in relation to the performance of the other MOOC participants.

Fourth, *Instruction*: In cultures with weak uncertainty avoidance students feel comfortable in unstructured learning situation whereas students from cultures with strong uncertainty avoidance tend to prefer structured learning situations (Hofstede and Hofstede 2005). Students from cultures with high uncertainty avoidance might prefer strict instructions how to assess their peers' assignments, whereas students from cultures with low uncertainty avoidance might feel more comfortable with more freedom regarding the feedback process. At the moment Coursera offers rubrics for feedback to ensure uniform and consistent reviewing and ask the reviewer to provide additional qualitative feedback.

Fifth, *interaction between students*: The need for interaction might vary between students from different cultures. Cultures with long-term orientation believe that truth depends on context and situations (Hofstede and Hofstede 2005). Thus, students from those cultures might prefer the opportunity to discuss an assignment and the peer feedback to gain a deeper understanding of the context. At the moment, Coursera does not offer the opportunity for interaction between students during or after the peer feedback process. Discussion does solely take place in the discussion forums, open and visible for everyone.

### 6.3.3 Design and Development

Addressing the discussed areas of the peer feedback process eligible for adaption, the following section deduces design principles (DP) for a cultural sensitive peer feedback process. It should be noted that the deduced DPs refer to the process of the peer feedback and are not DPs of the user interface. The DPs, their description and corresponding dimensions are shown in Figure 6.2.

First, to address the acceptance of feedback, the peer feedback process needs to include information about the reviewers. The information might include basic information such as name, age as well as the educational background and relevant work experience (DP1). Second, to address differences in group-dynamics, students should have the option to keep the same reviewer during a course to meet the preferences of small group belongings of collectivistic cultures (DP2). Third, to give competitive students the opportunity to set their performance in relation to the performance of other MOOC participants, the results of the peer feedbacks should be made comparable, for example on a grade-scale (DP3). Fourth, to meet the need for instructions, especially in high uncertainty avoiding cultures, MOOC providers should consider the inclusion of a more extensive support in the form of more information and instructions on how to give feedback during the peer feedback process (DP4). Finally, students should be offered the opportunity to discuss their assignments and received grades. It should be considered to include the opportunity of private messages with reviewers after the feedback is done (DP5).

**Table 6.2:** Design Principles and their Inherent Cultural Dimension

| <b>Design Principles</b>                                   | <b>Description</b>                                                                                                                                                                                                                                        | <b>Dimension</b>                     |
|------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------|
| <b>DP1:</b><br>Profile Information                         | Students from cultures with high power distance prefer to have information about their reviewer whereas students from cultures with low power distance prefer anonymity.                                                                                  | Power Distance                       |
| <b>DP2:</b><br>Group Dynamics                              | Students from collectivistic cultures prefer to keep the same reviewers during the duration of a MOOC whereas students from individualistic cultures prefer to change reviewers.                                                                          | Individualism<br>vs.<br>Collectivism |
| <b>DP3:</b><br>Competition and Comparison between students | Students from masculine cultures prefer to receive grades instead of “passed/failed” and see a scale of their performance in relation to the performance of their fellow students.                                                                        | Masculinity<br>vs. Femininity        |
| <b>DP4:</b><br>Instructions                                | Students from cultures with high Uncertainty Avoidance prefer to have detailed instructions for the peer feedback process in a MOOC, whereas students from cultures with low Uncertainty Avoidance prefer to have more freedom how to review their peers. | Uncertainty<br>Avoidance             |
| <b>DP5:</b><br>Interaction between students                | Students from cultures with Long Term Orientation prefer to interact with their reviewers and rate their received feedback.                                                                                                                               | Long Term<br>Orientation             |

## 6.4 Results

To evaluate the DPs derived from the literature to design cultural sensitive peer feedback, 16 students a Coursera course were interviewed. of this study about their preferences. Table 6.3 summarizes the findings.

**Table 6.3:** Comparison of Swiss and Chinese Preferences for the Developed Design Principles

| Design Principles | Description                     | China      | Switzerland |
|-------------------|---------------------------------|------------|-------------|
| <b>DP 1</b>       | Preference for Information:     | 7 (87.50%) | 1 (12.50%)  |
|                   | Preference for Anonymity:       | 1 (12.50%) | 7 (87.50%)  |
| <b>DP 2</b>       | Keep the same Reviewers:        | 3 (37.50%) | 0 (0%)      |
|                   | Change Reviewers:               | 5 (62.50%) | 8 (100%)    |
| <b>DP 3</b>       | Preference for Grades:          | 7 (87.50%) | 6 (75%)     |
|                   | Preference for Scale:           | 8 (100%)   | 7 (87,50%)  |
|                   | Preference Passed/Failed:       | 1 (12.50%) | 1 (12,50%)  |
|                   | Preference for No Scale:        | 0 (0%)     | 0 (0%)      |
| <b>DP 4</b>       | Preference for Instructions:    | 3 (37.50%) | 4 (62,50%)  |
|                   | Preference for Freedom:         | 5 (62.50%) | 4 (37,50%)  |
| <b>DP 5</b>       | Preference for Interaction:     | 6 (85.71%) | 3 (37,50%)  |
|                   | Rate the received Feedback:     | 8 (100%)   | 7 (87,5%)   |
|                   | Preference no Interaction:      | 1 (14.29%) | 1 (14,29%)  |
|                   | No Rating of received Feedback: | 0 (0%)     | 0 (0%)      |

## 6.5 Discussion and Contributions

The aim of this study is to investigate how to design a cultural sensitive peer feedback process in MOOCs to enhance the learning experience for students. For this purpose, I used cultural dimension theory developed by Hofstede (2013) to derived requirements based on the scientific literature and develop DPs for a cultural sensitive peer feedback process. The DPs were empirically tested



by conducting 16 semi-structured interviews consisting of Swiss and Chinese students. I claimed that culture has an impact on design preferences of the peer feedback procedure in MOOCs. On the one hand, the results show that students from different cultures have varying design preferences. However, some DPs could not be supported and were not in line with the requirements derived from scientific literature.

*Profile Information (DP1).* The results show differences in design preferences. Whereas Swiss students stated they prefer anonymity, Chinese students prefer to have more information about their reviewers in a MOOC. As a quote by one Chinese participant shows:

“Yeah, I prefer to get more information about the peers. I do not care about the name but I really care about his background and his major, or experience, because I want to know what kind of person [...]. Because I want to know if I can get some connection between he or she and me. If he or she can make a judgment.”

Based on the cultural dimension theory these differences can be explained through the strong power distance in Chinese culture. In this culture less powerful people should be dependent and therefore not entitled to criticize and in this case even assess the performance of more powerful people (Hofstede 2013).

*Group Dynamics (DP2).* The study does not support this DP. I hypothesized that, because of collectivistic values like the importance of group belongings, Chinese MOOC participants prefer to keep the same reviewers during the duration of a course. However, only 37,5% of the participants stated they prefer to keep the same reviewers whereas 62,5% stated they prefer change. Most of the participants indicated that they value diversity of opinions over the development of relationships. As a quote by one Chinese participant shows:

“I think this is a way to have me learn more how people evaluate my work. Different people’s opinion might be different from each other. I think it’s earning things from experience from the evaluation from different people.”

Although the majority of Chinese students stated to prefer differing reviewers during a MOOC, there are still noticeable differences to Swiss students. Whereas all of the Swiss students (100%) claimed to prefer to change reviewers frequently, some Chinese students also considered the advantages of keeping the same reviewers. As a quote by one Chinese participant shows:

“Oh yeah, I think it’s a question of consistency. So I think to have the same reviewers all through the course is better.”

Future research will need to execute a deeper analysis of this DP to evaluate its significance.

*Competition and Comparison (DP3).* I hypothesized that the participants from China and Switzerland prefer to have distinct grades and the opportunity to compare their performance. Based on our results, this holds true in practice. Both countries have similar scores on Hofstede (2013) cultural dimension “masculinity vs. femininity” which conduce as explanation for the similarity in preferences. Participants from both countries predominantly valued the opportunity to set their performance in relation to the performance of their peers as a quote from one Swiss participant shows:

“I think that would be really helpful. Sometimes just having a number does not really tell you what this number means. So if you have something to compare you with that can be very useful”

*Instructions (DP4).* I hypothesized that students from countries with high uncertainty avoidance prefer to have stricter instructions on how to give feedback during the peer feedback process in a MOOC. The data supports this assumption. The majority of Swiss participants stated they prefer to have strict instructions how to review their peers in a MOOC as this quote by a Swiss participant demonstrates:

“I think strict instructions are useful, because they imply consistent evaluations which are not subjective.”

In contrast, the majority of Chinese participants valued freedom during the peer feedback process. The preferences are congruent with the cultural dimension

theory. Strong uncertainty avoidance implies an emotional need for rules expressed by the desire for more instructional support on how to do the peer feedback process in MOOCs. Weak uncertainty avoidance manifests in dislike of rules and the Chinese participants are more comfortable with weak instructional support. When having a closer look at what instructions the participants perceived as valuable, it is noticeable that the Swiss participants predominantly value support how to assess the content of the peer's assignment leading to a consistent and fair evaluation process. The Chinese participants emphasized that there might not be one best answer and that, for example a single best-case solution is not enough. They prefer support which aspects they should consider in their evaluation for example through multiple cases and multiple sample answers for comparison. Which is shown by the quote of one Chinese participant:

“Maybe just a best practice. Oh, or maybe, yeah I know that. So maybe a best practice from like different point of views. So not just one, maybe we can have couple of it.”

From a theoretical point of view, these differences can be explicated by the tolerance of deviant ideas and ambiguity, resulting from weak uncertainty avoidance in the Chinese culture.

*Interaction Between Students (DP5).* This DP holds partially true in practice. I hypothesized that students from long-term orientation cultures prefer the opportunity to rate the received peer feedback in MOOCs. The data supports this assumption. Most participants stated they believe this opportunity will lead to more efficient peer feedbacks. As a quote from one Chinese participant shows:

“Yeah I think this can help, like to make the evaluator to evaluate others more fairly.”

This belief can be explained by the desire for accountability, honesty and self-discipline in cultures with long-term orientation (Hofstede and Hofstede 2005). However, the majority of Swiss participants stated that they do not value the opportunity to further interact with their reviewers. Despite main work values like learning and perseverance in long-term orientation cultures the Swiss participants predominantly emphasized the increased workload and the risk

of manipulation through the opportunity of direct and anonymous interaction between peers after and in particular during the feedback. As a quote by one Swiss participant shows:

“I think anonymity is good because otherwise it is an immense effort to personally discuss with the peers and as I said I strictly prefer anonymity during the whole thing and to discuss personally would be contradictory.”

Accordingly, this study has theoretical as well as practical contributions. DSR studies are expected to produce generalized outcomes of the problem instance and the solution instance and to derive DPs (Hevner 2007). The class of problems addressed by this study was the problem of varying cultural sensitive needs in the peer feedback process. While the specific context was international large-scale classes such as MOOCs. Through this study, I provided the first step of developing generalizable DPs to design a culture sensitive peer feedback process. Furthermore, I provide practitioners valuable insights and guidelines on how to design MOOCs, to adapt the system to the needs of different cultures. The class of solutions addressed by my study was the extant approaches used to solve the problems of varying cultural needs in the peer feedback process. Thus, this study contributes prescriptive knowledge, to the existing body of knowledge of cultural research by adapting an existing framework to the specific case of peer feedbacks. Additionally, through developing DPs, the study provides a contribute to theory building for the class of cultural sensitive peer feedback processes. Furthermore, through testing the DPs with participants from China and Switzerland, the study contributes empirical profound data.

Moreover, I provide a practical contribution, through offering guidelines which can be directly used by instructional designers of MOOCs, to enhance the peer feedback process. These guidelines consist of a proposal to adapt the user interface according to the cultural background of the students. For example, if a student from a high power distance does receive a feedback, it might be important to provide additional information about the feedback provider, such as how many feedbacks this person has already provided, what was the average feedback quality etc.

### 6.5.1 Limitations and Future Research

This study offers valuable contributions for theory and practice. However, the possible limitations need to be taken into account. Nations might not be the best units for studying cultural differences, as they do not produce groups of people with uniform codes of behavior. For example, a Swiss person who spent couple of years in Australia might inhibit a mixture of both cultures. Therefore, it is important that for every student participating in MOOCs, fill out the Value Survey Model.

Furthermore, this initial study contains a rather small sample size. Hence, further research is needed with a much larger sample size and with a diverse cultural backgrounds to gain a better understanding of how design preferences evolve and consequently develop general solution concepts for different cultures. Additionally, further research needs to quantitatively examine the impact of the developed DPs on the learning experience of MOOC participants.

## 6.6 Conclusion of this Chapter

The aim of the research presented in this chapter was to develop DPs to make a first step towards creating a culturally sensitive peer feedback process. Firstly, I used the cultural dimension theory by Hofstede and Hofstede (2005) to derive requirements based on the scientific literature. Secondly, I derived requirements from analyzing the peer feedback process in three Coursera courses. Based on these requirements I developed five generic DPs which were empirically tested with 16 semi-structured interviews. The participants had Chinese and Swiss cultural backgrounds.

The results of the Value Survey Model by Hofstede (2013) showed evidence of differences between the two cultures. Furthermore, the results of the 16 interviews showed that, for example, for DP one the preference for profile information was clearly different for both countries. Whereas Swiss students preferred anonymity, Chinese students desired more information about the feedback provider. DP two was not able to be supported in this study. For DP three, the hypothesis that participants from China and Switzerland preferred to have distinct grades and the opportunity to compare their performance was

supported. Furthermore, DP four was also supported in that students from countries with high uncertainty avoidance preferred to have stricter instructions on how to give feedback during the peer feedback process in a MOOC. In contrast DP five was only partially supported by this particular study. There were mixed findings for the hypothesis that students from long-term orientation cultures preferred to have the opportunity to rate the received peer feedback in MOOCs.

Overall, this study provides valuable information for developing a system to measure peer feedback quality automatically, since feedback quality and the teaching-learning scenario in which it is used may be dependent on the cultural backgrounds of the participating students.

# 7

## **Develop the Peer Feedback Providing Skill**

### 7.1 Introduction

In this chapter<sup>1</sup> I present a study on measuring the effects of technology-mediated formative feedback on higher-order thinking skills involved in providing peer feedback. Future graduates need so called higher-order thinking skills – critical thinking, self-regulated learning and communication – to be successful in the job market (OECD 2019). One example of an increasingly important higher-order thinking skill is the ability to provide feedback (Highsmith and Cockburn 2001; Williams and Cockburn 2003). The uprise of agile modes of working and increasingly flat hierarchies in organizations often require more intense communication between employees at similar levels of seniority. However, the provision of high quality feedback needs to be learned like most other skills and, thus, corresponding exercises need to be integrated in the learning process.

According to Hattie and Timperley (2007), there are two essential factors for training higher-order thinking skills: a) a realistic teaching-learning scenario that allows students to apply the skills that need to be developed and b) the provision of rich and meaningful feedback about the learners' progress, which is one of the most important factors affecting performance. A major challenge to this is that most universities struggle to provide such boundary conditions, due to increasing numbers of students paired with financial resource constraints.

As a solution, I propose an IT-based peer feedback system (ITPF), in which students a) find realistic teaching-learning scenario in which they are not frightened by failure, since the peer learning is based on anonymity, and b) develop their higher-order thinking skills such as the skill of providing high quality feedback to their peers. ITPF can be implemented using a learning management system (LMS) such as the web-based IT-tool Moodle<sup>2</sup> in a lean way that keeps the additional workload for a lecturer low, since the LMS facilitates the whole process of anonymizing and distributing the submissions and feedback.

In this dissertation, formative feedback is composed of four different types. Figure 2.2 in Section 2.3.2 shows these four types: *L1*, *L2*, *H1* and *H2*. *L* and *H* are related to the cognitive dimensions. *L* represents the lower-order thinking

---

<sup>1</sup> A previous version of this chapter (Sections 7.2 to 7.4) has been published in Rietsche and Söllner (2019). The content has been reformatted and updated to the latest state of scientific discussion at the time of submission of this dissertation.

<sup>2</sup> [www.moodle.org](http://www.moodle.org)



skills and  $H$  the higher-order thinking skills. The numbers 1 and 2 are related to the knowledge dimensions. Number 1 represents the factual, the conceptual and the procedural knowledge dimensions and number 2 represents the meta-cognitive knowledge dimension. The formative feedback used in this chapter is based on  $H1$  and  $H2$ .

In  $H1$  feedback is conceptualized as a higher-order thinking skill hence, the ability to provide feedback to another person. In  $H2$  feedback is conceptualized as the information given in a formative manner to assess the current state of the development of the skill of  $H1$ . The formative feedback of  $H2$  consists of an internal assessment and an external assessment. The internal assessment is carried out by the students; it is a Self-Assessment (SA) of their own ability to provide peer feedback. The external assessment is a Peer Assessment (PA); it is conducted by the peers of the students, revealing their perceptions of the students' ability to provide high quality feedback.

Thus, the overall the aim of this chapter is to answer research question four of this dissertation:

*What are the effects of technology-mediated formative feedback  
RQ4: on the students' perceptions of their skill at providing peer  
feedback?*

To answer this research question, I conducted a quasi-experiment with a one-group pre-/post-test design which consisted of between 73 and 101 students participating in seven ITPFs rounds. To evaluate their skill development, students had to self-assess their current ability to provide feedback to a peer prior to the first learning unit (baseline). After the lecture of each learning unit, the students completed a free-text or modeling submission and self-assessed their ability to provide feedback (SA). Afterwards, the system sent the submission anonymously to three peers for peer feedback. Hence, each student had to write three feedbacks for their peers and received three peer feedbacks on his or her own submission. To ensure that the students received well-written feedback, the students in turn judged the quality of the feedback provided by their peers (PA). In the last phase, the students rated their own ability to provide feedback to a peer (same survey questions as the baseline). To analyses whether the students'

ability to provide feedback improved over time, I used two-sided paired t-tests on both SA and PA data.

The remainder of this chapter is structured as follows. The next section describes the research methodology and how I design the experiment, the measurements and the ITPF process. The following section presents the results of the experiment, which are the foundation for the discussion in next chapter. The final section provides a conclusion to this chapter.

## 7.2 Research Methodology

### 7.2.1 Experimental Design

The ITPF tool (module implemented in Moodle) was used in a quasi-experiment with a one-group pre-test post-test design (Millsap and Maydeu-Olivares 2009) in a large-scale the master's level university class on business model innovation. The study consisted of seven ITPFs that took place in weekly over a period of one semester. Between 73 and 101 students participated in each ITPF. In case a student did not participate in a particular ITPF, which leads to a missing value, this was statistically considered (marked as NA in R). Missing values as consequence lead to different Ns per construct. The attendance in the study was optional and rewarded with points for the final grade.

To ensure that the students had enough knowledge in terms of how the Moodle LMS works and how to provide constructive feedback to their peers, an initial tutorial session and a one-week practice test was carried out before the first ITPF iteration took place. In the tutorial session the overall ITPF process and the function of Moodle was explained followed by a hands-on instruction of how to provide constructive feedback. The instructions were based on common didactical guidelines for providing feedback, such as focus on the topic not the person, be constructive, and write in "I" form etc. (Rietsche et al. 2017a).

During the training period, the learners were also invited to contact the lecturer and assistant with any questions regarding the ITPF, either in the classroom or via mail. After the end of the practice test, the first ITPF iteration was started. In each ITPF iteration, for the SA and PA, two types of scales were used to gather rigor research data: a) SA using a percentage scale and b) PA using a

Likert scale. In order to compare the SA and the PA data, the percentage scale was transformed into a Likert scale using the formula (7.1) (IBM SPSS 2018). We transformed the percentage scale to the Likert scale because the common transformation is from a higher granular scale to a low granular scale.

$$SA_{likert} = (Likert_{max} - Likert_{min}) * (SA_{percent} - Percent_{min}) / (Percent_{max} - Percent_{min}) + Likert_{min} \quad (7.1)$$

### 7.2.2 Measurements

Table 7.1 shows the items for measuring the students' ability to provide feedback. The SA1 measures the ability based on the students' own perception (SA). The items PA1-PA4 measure the students' ability based on the perception of the peers (PA). The mean value of the four items builds the latent construct ability to provide feedback (PA).

**Table 7.1:** Survey Questions Self-Assessment (SA) and Peer-Assessment (PA)

| Item | Question                                                                                    | Scale          | Source                 |
|------|---------------------------------------------------------------------------------------------|----------------|------------------------|
| SA1  | How do you rate your personal ability to give your fellow students feedback?                | 0-100%         | Own Development        |
| PA1  | The feedback I got from Reviewer X was helpful?                                             | Likert:<br>1-7 | (Strijbos et al. 2010) |
| PA2  | The feedback I got from Reviewer X was high quality?                                        | Likert:<br>1-7 | (Li et al. 2010)       |
| PA3  | Reviewer X was able to identify critical aspects in my assignment?                          | Likert:<br>1-7 | (Li et al. 2010)       |
| PA4  | The Reviewer X was able to provide constructive suggestions on his stated critical aspects? | Likert:<br>1-7 | (Li et al. 2010)       |

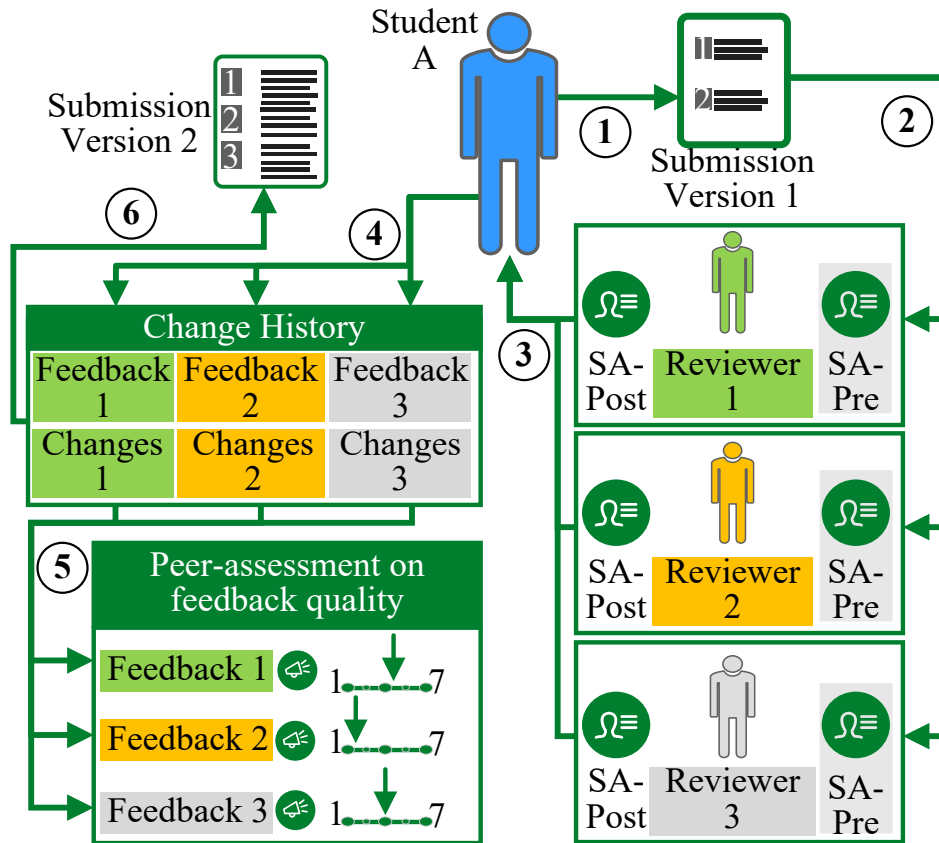
### 7.2.3 ITPF Process

Figure 7.1 shows the ITPF process used in my experimental setting, where the student is in the role of being a feedback receiver. A student usually has two roles: a) being the feedback receiver and b) being the feedback provider.

In the first step, student A uploads the submission version 1. The submission contains the solution of student A, being an entrepreneur in applying the methods for describing and modeling their business model. In the second step, the submission is anonymously distributed to the three reviewers. The three reviewers carry out an SA prior to the first ITPF iteration to rate their ability to provide feedback. This pre-SA is carried out only one time before the first ITPF iteration. Afterwards, the reviewers write the review to emphasize the strengths and weaknesses of the submission and how the submission could be improved. This is succeeded by carrying out an SA post each ITPF. The SA post is carried out in every ITPF iteration. SA pre and SA post are rated on a percentage scale from 0 to 100% with 0% meaning the student judges his ability to provide feedback to peers to be very poor and 100% meaning very good. In the third step, student A receives the three feedbacks. In the fourth step, student A creates a change history. This contains the feedbacks as well as how the comments from the reviewer were understood and applied the changes for the final submission. In the fifth step, student A provides feedback on the review quality (PA), meaning how helpful it was. The question was based on a Likert scale from 1 (poorest quality) to 7 (highest quality). In the final step, student A revises the submission based on the change history and uploads version 2.

## 7.3 Results

For the data analysis, I used R version 0.99.902 (free statistical programming software) without any additional packages. For analysis, I used the R one-sided, paired t-test. The Figure 7.2 shows the survey results for the SA (solid line) question SA1 (Table 7.1) and the PA (dashed line) average of questions PA1 to PA4 (Table 7.1). The first measurement  $t_0$  is the baseline measurement at the beginning of the class, before the first ITPF iteration took place. For the baseline measurement, there is no PA available, since the peers did not review



**Figure 7.1:** Steps of the IT-based Peer Feedback Process (Self-Assessment and Peer-Assessment)

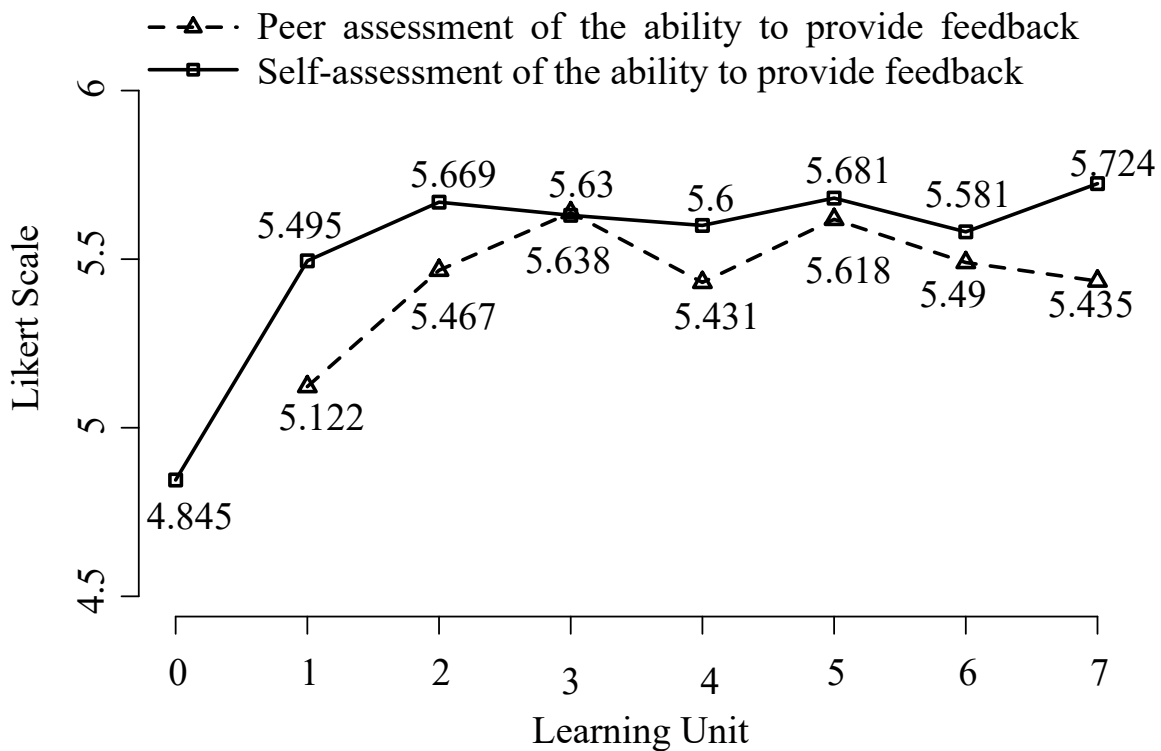
a submission yet. In  $t_0$ , the students carried out the SA pre-assessment and self-assessing their ability in providing feedback to their peers.

Time points  $t_1$  to  $t_7$  show the seven ITPF iterations. Comparing  $t_0$  to  $t_1$ , (SA) the students' perceived ability to provide feedback to peers increased by .65 points on the Likert scale, which is highly significant ( $p < .001$ ). Moreover, when considering the time points  $t_1$  to  $t_7$  (SA), the students enhanced their perceived ability to provide feedback through completing the ITPF multiple times. The difference between  $t_1$  to  $t_7$  is .229 points for the SA ( $p < .05$ ). Overall, the difference difference from  $t_0$  to  $t_1$  (SA) was the sharpest increase (.65 points) in the students' ability to provide feedback. This increase is followed by a smaller growth between  $t_1$  and  $t_2$  of .174 points. From  $t_2$  to  $t_6$ , the line settles at around 5.6 points. The line ends with an increase of .143 points in  $t_6$  to  $t_7$ .

The second line (dashed) in 7.2 represents the PA of the students' ability to provide feedback. The line starts with a sharp increase by .516 points in  $t_1$  to  $t_3$ . Followed by a drop of .207 points from  $t_3$  to  $t_4$ . In the minimal turning

point, the line increases again to 5.618 points and decreases afterwards from  $t_5$  to  $t_7$  by 183 points. The increase as shown by the dashed line PA from  $t_1$  to  $t_7$  is significant ( $p < 0.05$ ).

Table 2 shows an overview of the statistical results. I calculated the mean and standard deviation for SA and PA with two intervals: Interval A,  $t_0$  to  $t_1$ ; and Interval B,  $t_1$  to  $t_7$ . The mean for SA Interval A is 5.17 points ( $SD = .459$ ) on the Likert scale and for SA Interval B 5.626 ( $SD = .076$ ). The PA Interval A is not available, since the peers did not review a submission yet. For Interval B the mean is 5.457 points ( $SD = 0.17$ ).



**Figure 7.2:** Results of the Students' Ability to Provide Feedback to Peers (SA and PA)

## 7.4 Discussion and Contributions

Peer-feedback approach is not new to in higher education and was already investigated in the 1970s (Smith et al. 1996). However, there is still a lack of empirical research using ITPF in quasi-experiment design settings (van Zundert

et al. 2010). Since, to my best knowledge, the current available research mainly concentrates on using ITPF to develop students' cognitive knowledge, such as declarative or conceptual knowledge and not for developing higher order thinking skills. The aim of my study is to investigate whether ITPF can be used to develop students' feedback providing skills in large-scale classes with limited available resources in a management-orientated course.

The findings of my study show that using ITPF has a significant positive effect on the students' ability to provide feedback. This effect has been supported in terms of SA and PA. The results further show that students' ability to provide feedback already benefits from a single round of practice and benefits even further when ITPF is used multiple times in the learning process. This leads to the conclusion that students highly benefit from the use of ITPF, especially in terms of developing their feedback providing skill. Interestingly, comparing the ability to provide feedback on the SA and PA, there is a tendency for students to rate their personal ability to provide feedback higher than the ability assessed by their peers. For time point  $t_1$  (SA-PA) and  $t_7$  (SA-PA), this can be statistically shown with a two-sided, paired t-test, which shows that  $p < .05$ .

However, it needs be considered that the students did not view the PA of their own submissions. Thus, when the students self-assessed their ability to provide feedback, they did not know how their peers ranked their ability to provide feedback in the previous ITPFs. According to Moshavi et al. (2003), over-estimators are believed to be associated with the most negative individual and organizational outcomes. This can lead to a misdiagnosis of strengths and weaknesses of the over-estimators and results in ignorance towards how others perceive them.

Hence, future research could possibly include providing feedback on the previous PA results before each ITPF. It could be investigated if a convergence between the SA and PA takes place over time, to decrease the students' overestimation of their ability of to provide feedback. Furthermore, in the current research setting, cognitive dissonance may be a reason why students' SA increased over time.

Since each student, on the one hand, wrote three reviews and, on the other hand, received three reviews, the dissonance maybe occurred while comparing their

own feedback to the received reviews. Thus, over-time students may try to minimize the dissonance and rate their own ability similarly to the ability of their peers. However, further research is necessary to investigate this phenomenon. Especially interesting will be measuring cognitive dissonance when showing the students the SA and PA scores at the same time. Furthermore, considering the drop in  $t_4$  (7.2) may be caused by the change of the task from a descriptive to a modeling manner. This tendency is also the case regarding  $t_6$ . However, the correlations cannot be statistically shown. The tendency shows that the ability to provide feedback may be related to task difficulty and complexity. Further research is necessary to statistically explore this possible correlation.

The results of my research are of practical relevance for researchers, lecturers, and practitioners alike, since they provide insights in how ITPF can be used to develop the students' ability to provide feedback. Furthermore, the research provides a description of the steps of the ITPF process. The theoretical contribution of my research is in providing a first step in applying the concept of cognitive dissonance (SA and PA) as a theoretical lens for explaining why students' ability to provide feedback increases over time. Furthermore, I contribute to the collaborative learning theory by providing prescriptive knowledge in the form of empirical results. I used the collaborative learning theory in an innovative teaching-learning scenario in which students develop their own business model over time and simultaneously practice the skill of providing feedback to peers.

### 7.4.1 Limitations and Future Research

Any findings of an empirical study need to be considered in the light of its limitations. Since I gathered the data in a real class setting, that provided the students with the possibility to take part in a quasi-experiment with a one-group pre-test post-test design, I did not have an experimental setting with a control group. This leads to the limitation that I cannot exclusively demonstrate that the ability to provide feedback to peers results from students' participation in the ITPFs. Based on practical and ethical reasons, it is sometimes, not possible to use randomization when selecting participants for a treatment or control group (Millsap and Maydeu-Olivares 2009). At my university, this was the case, as students were free to choose their class and students of the control group would not receive the benefits like the students in the treatment group did. According to



**Table 7.2:** Statistical Results of the Students' Ability to Provide Feedback to Peers (SA and PA)

| <b>Self-assessment (SA)</b> |                                   |                            |                                    |
|-----------------------------|-----------------------------------|----------------------------|------------------------------------|
| <b>Time point<br/>(t)</b>   | <b>Statistics<br/>(in points)</b> | <b>t(df)<br/>(t-value)</b> | <b>Differences<br/>(in points)</b> |
| 0 to 1                      | Mean = 5.17,<br>SD = .459         | t(77)= 5.736***            | .65                                |
| 1 to 7                      | Mean = 5.626,<br>SD = .076        | t(87)= 2.978*              | .879                               |
| <b>Peer-assessment (PA)</b> |                                   |                            |                                    |
| <b>Time point<br/>(t)</b>   | <b>Statistics<br/>(in points)</b> | <b>t(df)<br/>(t-value)</b> | <b>Differences<br/>(in points)</b> |
| 0 to 1                      | -                                 | -                          | -                                  |
| 1 to 7                      | Mean = 5.457,<br>SD = 0.17        | t(97)= 2.472*              | .344                               |

Significance with \*  $p < .05$ , \*\*\*  $p < .001$  SD=Standard Deviation

Millsap and Maydeu-Olivares (2009) a quasi-experiment has up to seven threats for internal validity (maturation, history, seasonality, testing, instrumentation, attrition and statistical regression). Nevertheless, these threats heavily depend on the actual design setting, such as the type of treatment, the participants, the time between pre- and post-test, and other factors. According to Eckert (2000) none of these threats are likely or plausible in educational interventions teaching materials that are highly unlikely to be learned somewhere else between the pre- and post-test. If the threats to internal validity play a role, this mainly depends on the context in which the quasi-experimental design was used and is built on two conditions: a) the pre- and post-test measure focus solely on the material being taught and b) the time interval between the pre- and post-test is short. Both conditions are met in my study and, hence, this leads us to the conclusion that my research findings have valid insights.

Moreover, for further research, the research design could be revised in a way that the number of ITPFs is halved. In the first half, the first group would be the treatment group and the second group the control group, and in the second half this could be reversed. However, this would have an impact on the learning effect because the students need to get comfortable with the setting of PA and its processes. Additionally, in future research, the impact of the submission on the PA could be investigated by using a questionnaire that retrieves the difficulty and complexity of the submission. Hence, the correlation between the submission's complexity, knowledge growth, and ability to provide feedback could be investigated.

### **7.5 Conclusion of this Chapter**

The aim of the research presented in this chapter has been to measure the effects of technology-mediated formative feedback on the students' higher-order thinking skill of providing peer feedback. As a solution for developing this skill, I devised the ITPF tool which can be implemented, for example, in Moodle. I tested the ITPF in a quasi-experiment with a one-group pre-test/ post-test design. In seven rounds, between 73 and 101 students participated in each ITPF.

The study showed that the use of ITPF could enable the lecturer to support students in their learning process in a resource-saving way. Additionally, ITPF

provides an environment in which students can anonymously train their skill of providing feedback. I emphasize that providing feedback is an important skill for students graduating university and working in the future in agile teams. However, the development of this skill is usually neglected, caused by problems of large-scale classes and the scarcity of resources, which lead to a lack of feedback.

The results of the quasi-experiment with a one-group pre-test, post-test design showed that, when using ITPF, the students' ability to provide feedback increased significantly. The students' ability to provide feedback already benefited from a single round of practice but benefited even further when ITPF was used multiple times in the learning process.



# 8

## **Synthesis of the Findings**

In this chapter I provide a synthesis of my research and the inter-relationship between the research questions. This dissertation has two guiding principles. Firstly, to design systems in teaching-learning scenarios for me it is more important to understand what the characteristics and aspects of the design should be as well as its components, rather than how (process) this system is designed. Secondly, it is necessary to measure the effects of the design by applying within a real world scenario.

Today's world is changing faster than ever before. Students graduating from university find a world that can be characterized by the acronym *VUCA*, which stands for volatility, *uncertainty*, complexity and *ambiguity*. To respond to VUCA conditions, a new movement has emerged over the last two decades which is shaped by an agile mindset (Beck et al. 2001). More and more companies have adopted this mindset and have started to work in small agile teams. In these teams, members work in short cycles with rapid feedback to respond quickly to changing conditions.

However, this 'new' way of working requires employees to have a different kind of skill set, with mainly higher-order thinking skills. Two of the most important skills are (1) being able to reflect on one's current state of knowledge and skills and (2) being able to provide feedback to team members to steer them in the right direction (Williams and Cockburn 2003; Highsmith and Cockburn 2001).

According to Hattie and Timperley (2007), there are two essential factors for developing higher-order thinking skills: Factor A, a realistic teaching-learning scenario that allows students to apply the skills that need to be developed, and Factor B, the provision of rich and meaningful feedback of the students' progress, which is one of the most important aspects. However, in trying to provide those boundary conditions, most universities are struggling due to an increasing number of students and limited financial budget.

Consequently, the research problem I address in my dissertation is how to create and evaluate a challenging teaching-learning scenario that helps students develop their higher-order thinking skills through providing technology-mediated feedback. This is researched in the context of boundary conditions where lecture sizes keep growing and financial resources stay constant or are even decreasing.

To address the described problem, one solution is to use information technology (IT). IT can help to leverage scaling effects and address the problem of large-scale classes. Hence, the objective of this dissertation is to design and create an IT-enhanced teaching-learning scenario in which students are able to a) develop their reflection and feedback skills and b) receive technology-mediated formative feedback during the learning process.

However, within the teaching-learning scenarios different learning outcomes are addressed, which requires the formative feedback to be assembled accordingly. The revised Bloom learning taxonomy by Krathwohl (2002) provides a basis for matching formative feedback to the learning outcomes. The fundamental concept is that this taxonomy consists of *Knowledge* and *Cognitive Dimensions*. Thus, to develop the students skills holistically, formative feedback needs to be assembled in way that addresses each domain of the taxonomy specifically. Figure 2.2 in Section 2.3.2 shows the four types of feedback in this dissertation. The formative feedback in *L1* covers four Cognitive Dimensions (Remember, Understand, Apply and Analyze) and three Knowledge Dimensions (Factual, Conceptual and Procedural). The formative feedback in *L2* covers the same Cognitive Dimensions as L1 but the Knowledge Dimension is Meta-Cognitive. The formative feedback in *H1* covers two of the higher level Cognitive Dimensions (evaluate and create) and the same three Knowledge Dimensions as L1 (Factual, Conceptual and Procedural). The formative feedback in *H2* covers the same Cognitive Dimension as H1, but the Knowledge Dimension is Meta-Cognitive. The connection between these four types of feedback to the research questions are described in the next section.

## **8.1 Building IT Artifacts for Providing Technology-Mediated Formative Feedback**

There are many IT tools in the market to support teaching-learning scenarios for higher education. Popular ones are Blackboard<sup>1</sup>, Canvas<sup>2</sup> and Moodle<sup>3</sup>. However, most of the IT tools have in common the purpose of providing students with learning materials such as lecture slides and additional readings, offering

---

<sup>1</sup> [www.blackboard.com](http://www.blackboard.com)

<sup>2</sup> [www.canvas.net](http://www.canvas.net)

<sup>3</sup> [www.moodle.com](http://www.moodle.com)

forums to share knowledge and providing a platform to organize assignments (Dobre 2015). Some IT tools have additionally integrated an add-on module to support the creation of MCQs. However, those tools have not been specifically designed for providing formative feedback in higher education.

Furthermore, those tools have been developed by commercial providers. Hence, there is no publicly available design knowledge describing the requirements or design principles of the general class of tools providing formative feedback. Formative feedback and the tools providing it address various learning outcomes. The system I designed for this dissertation answers research questions one and two. For these research questions, formative feedback is assembled as L1 and L2. For L1, this means that the formative feedback consists only of external assessment, which is provided with the aim of, for example, testing whether students remember declarative knowledge. For L2, this means that the formative feedback consists of an internal and external assessment. If the gap between both of these assessments exceeds a certain threshold, then this leads to cognitive dissonance, triggering self-regulated learning processes.

Thus, the aim of these research questions was to design a system which provides formative feedback to students and document the design knowledge which emerged during the design process. Since I sought to design a socio-technical system, the design was based on requirements from the scientific literature (theories) and real-world data (interviews with lecturers and students). The results of this research study are design principles and design elements describing the design knowledge of formative feedback tools.

One of the challenges of CbA is that it has only limited ability to be use for open ended questions, because of its restricted accuracy when assessing creative work. For many years researchers and companies have been trying to develop Automated Essay Scoring (AES) systems (Graesser and McNamara 2012). However, up until now these systems have not been able to fully replace a human, especially in creative tasks such as essay writing (Balfour 2013). Furthermore, those systems are only able to provide grades and feedback on grammar. They are not able to provide rich qualitative feedback on how the submission could be improved. Additionally, at the moment there is still a high chance of cheating those systems (Farag et al. 2018). Thus, for scoring essays and, hence, open ended questions which involve creativity, humans



are still needed. However, within large-scale classes, those essays cannot be evaluated by lecturers or experts due to the limited available amount of resources, especially finances and time.

Hence, an alternative approach needs to be taken that overcomes the challenges of large-scale classes. One of these approaches is peer learning, whereby peers judge the quality of their peers' work and communicate their evaluations by providing written qualitative feedback. However, students are not experts in providing feedback to their peers, which means that lecturers need to provide students with guidelines on how to provide peer feedback. Yet the problem is that the scientific literature only provides rather generic principles describing what high and low feedback quality is.

Hence, the third research question addresses this challenge and aims to a) provide a systematic literature review on related work, b) provide design guidelines which define high and low quality peer feedback based on empirically grounded research and c) develop design principles for designing culturally sensitive peer feedback systems.

These three aspects build on each other. Firstly, a plethora of research has been carried out in the related field of online review helpfulness. This body of knowledge provides valuable insights into which characteristics of an online review have an impact on review helpfulness. The characteristics of online reviews are very similar to peer feedback. The results of the literature review revealed that review-related factors are the most highly researched aspects, followed by reviewer-related factors. However, hardly any research has been carried out on reader- or environment-related factors. Secondly, the insights of the first aspect of the research were used to build the guidelines for measuring peer feedback quality. I derived features (characteristics of feedback quality) deductively from the scientific literature and inductively from analysing real-world data' The results showed that out of the 90 features, 73%, had a significant impact on peer feedback quality. Similarly to online review helpfulness, feedback (review) length was one of the most important factors. Thirdly, to address the challenge of students having different cultural backgrounds, I derived design principles deductively from scientific theory and inductively based on analysing courses in MOOCs. Cultural backgrounds cause students to have distinct requirements when participating in a teaching-learning scenario

using peer feedback. The results showed that for 3 out of the 5 design principles there were clear differences between Swiss and Chinese cultures with regard to peer feedback requirements.

### **8.2 Effects of Technology-Mediated Formative Feedback on Student Performance**

Feedback is one of the top influential factors affecting student performance (Hattie and Timperley 2007). When considering formative feedback, which is a sub-type of feedback, this type belongs to the top three influential factors affecting student performance (Hattie 2008). The main aim of formative feedback is to guide students within their learning process and provide timely feedback to enable students to adjust their actions while developing their skills. As defined earlier, formative feedback is composed in four different types to address the whole range of learning outcomes in the Bloom's revised learning taxonomy (Krathwohl 2002).

The second research question aims to measure the effects of formative feedback, composed as L1 and L2, on student performance. Overall, L comprises the Remember, Understand, Apply and Analyze *Cognitive Process Dimensions*. While L1 targets the Factual, Conceptual and Procedural *Knowledge Dimensions*, L2 targets the Meta-cognitive *Knowledge Dimension*.

To measure these effects, I built an instance called LOOM based on the design knowledge that was the outcome of research question one. I evaluated the instance by following the evaluation framework proposed by Venable et al. (2016). I performed an artificial ex ante evaluation of the design features of the alpha version and a naturalistic ex post evaluation of the design features of the beta version. The results showed that students using the system, and hence receiving formative feedback assembled as L1 and L2, significantly increased their objective and perceived student performance.

The fourth research question aimed to measure the effects of formative feedback, assembled as H2 on higher-order thinking skills and in particular focused on the perceived skills of providing feedback. The effects of H1 are not part of this dissertation, since substantial research has been carried out on measuring peer evaluation (Formanek et al. 2017; Piech et al. 2013; Tahir 2012). However, to

the best of my knowledge there is no research on how formative feedback based on internal and external assessment impacts students' perceptions of their ability to provide feedback to their peers.

Future graduates need so called higher-order thinking skills - critical thinking, self-regulated learning and communication - to be successful in the job market (OECD 2019). One example of an increasingly important higher-order thinking skill is the ability to provide feedback. The emergence of agile modes of working and increasingly flat hierarchies in organizations often require stronger communication between employees at similar levels of seniority. In this research I designed the formative feedback to consist of an internal and an external assessment. The internal assessment is SA and the external assessment is PA. Hence, the former is the student's own perceived skill level and the latter is the peers' perceptions of the student's skill level. I evaluated the effects in a quasi-experiment, within a large-scale class including students at Master's level. The results showed that the students' own and by the peer perceived skill, of providing high quality feedback increases when using peer feedback only once as well as multiple times.



# 9

## **Theoretical Contributions and Implications**

A theory is an attempt to model aspects of the empirical world (Dubin 1976). There are two main drivers for building a theory. Firstly, the world is too complex to model its full nature and therefore needs to be conceptually simplified to understand it. Secondly, merely observing the empirical world does not reveal the underlying relationships among the observed phenomena. Hence, a theory tries to explain the world in an simplified way and to make sense of the observed phenomena by ordering the elements and explaining their relationship (Dubin 1976).

Research can contribute to theory in two ways, either by theory building or by theory testing (Colquitt and Zapata-Phelan 2007). Theory building is usually characterized by an inductive approach, in which observations made by the researcher are used to generate theory through inductive reasoning. The characteristics of the elements and their relationships are described by propositions. The aim of theory testing is to link the theory to the real world. This is achieved by operationalizing the propositions into testable hypotheses. The hypotheses test the outcomes of the relationships between the elements.

Each of the papers answering the four research questions of this dissertation try to contribute to theory in one of those two ways. However, it needs to be stated that theory building is a highly complex task. Hence, this dissertation does not claim to build new holistic theories, but rather to make a contribution to building theories.

The four research questions of this dissertation are:

- RQ1: What are the design principles for designing an IT artifact, to provide students technology-mediated formative feedback?*
- RQ2: What are the effects of technology-mediated formative feedback on student performance?*
- RQ3: Which characteristics determine the quality of technology-mediated peer feedback?*
- RQ4: What are the effects of technology-mediated formative feedback on the students' perceptions of their skill at providing peer feedback?*

## 9.1 Design Knowledge for Research on Formative Feedback IT Artifacts

Over the last two decades, in IS, a approach called DSR emerged to build so-called *design theories* (Hevner et al. 2004). A design theory is a combination of three interrelated elements: a set of user requirements, a set of design elements and a set of design principles deemed effective for guiding the process of development (Markus et al. 2002). According to Walls et al. (1992) a design theory is based on a theory and has the purpose to provide guidance to practitioners. A design theory builds on a “kernel theory”, enabling the formulation of empirically testable predictions. At the same time, a design theory is normative; hence they are prescriptive and evaluative.

Gregor and Jones (2007) built on this work and developed a taxonomy including the eight components of an IS design theory. I used this taxonomy to document the emerging design knowledge during the process of the development of the formative feedback tool. The purpose of this design knowledge is to provide abstract knowledge for designing IT artifacts of the general class of formative feedback tools. Researchers or practitioners can create an instance of this knowledge and build a concrete artifact. Hence, the theoretical contribution of the first research question is the design knowledge for building formative feedback IT artifacts.

I derived requirements from scientific literature and semi-structured interviews with lecturers and students and developed design principles for the general class of formative feedback tools. Formative feedback is becoming increasingly important, since the skills graduates need today shift more and more towards higher-order thinking skills. To develop these skills, formative feedback is one of the most important factors (Hattie and Timperley 2007). However, if such feedback needs to be provided to a large number of learners, then manual feedback by experts is nearly impossible due to resource scarcity in terms of time and finance. One solution to overcome this problem is to use IT, which tries to provide formative feedback at scale.

However, the question arises: how should we build these tools? To the best of my knowledge, there is no available design knowledge for building formative

feedback tools. The benefit of contributing design knowledge to the IS body of literature is to reduce developers' uncertainty by restricting the range of allowable design elements, design principles and development activities to a more manageable set (Markus et al. 2002). Additionally, documented design knowledge increases the reliability of development and the likelihood of success as well as stimulating future research. In IS, DSR should develop usable design knowledge and theory to apply to solve classes of IS problems (Carlsson et al. 2011). This includes abstract design knowledge for designing and implementing specific IS initiatives.

The type of theory I contributed to is “design and action” theory (Gregor and Jones 2007). The purpose of this theory is to provide prescriptions of principles of form and function for constructing IT artifacts. Hence, the description implies that following this recipe will lead to an instantiation of the class of formative feedback tools.

This research shed light on how to design formative feedback tools in general, by using, on the one hand, well established theories and, on the other hand, real world requirements from interviews with lecturers and students. The overall aim is to derive design principles and testing them in a real-world setting. Feedback is a well established research stream in education and an important element in IS education more specifically. Yet, little is known about the design of systems providing feedback, even though designing and developing socio-technical systems are the heart of IS-research (March and Smith 1995). For example, Niehaves and Ortbach (2016) presented an initial design for a system which provides feedback on an anagram-solving task. The authors used the human embodiment and media richness theory, which is impacting social presence. However, their design comprises only two design principles and is, therefore, a rather illustrative case of their inner and outer models of explanatory design theories instead of being a holistic DSR project.

Following a rigorous DSR approach, I built an IT artifact providing formative feedback. Evaluating this instance in a real world field experiment, the results showed that using the IT artifact had a significant influence on the students' objective and perceived learning performance. Thus, these findings suggest that the composition of the design principles has benefits for students.



Furthermore, the design science approach that was used contributes to Human Computer Interaction (HCI) and mobile application usability research, because the research provides valuable insights into how to design assessments using click events. These differ significantly when using a laptop compared to a smartphone. The carried out research showed that on smartphones users were more likely to accidentally click on the screen and lose their first attempt when conducting Flaw Diagram assessments than when using a laptop.

## **9.2 Effects of Technology-Mediated Formative Feedback on Student Performance**

Higher-order thinking skills such as critical thinking, communication and self-reflection are becoming more and more important for today's graduates leaving university. Formative feedback is one of the most important factors in developing these higher-order thinking skills. To develop self-reflection skills, students need to be able to evaluate their own current state of knowledge or skill level and compare it to an objective evaluation. Thus, when a student receives an internal and an external assessment of their skill, the difference between them may lead to cognitive dissonance followed by the desire to try to resolve this internal conflict, which ultimately initiates self-reflection processes.

However, due to increasing class sizes and an increasing student to lecturer ratio, formative feedback is a scarce good. Hence, a solution is to use technology-mediated formative feedback. Nevertheless, the question arises as to whether such formative feedback actually has an impact on student performance. There are studies showing the effectiveness of either internal or external assessments of formative feedback on student performance. However, there are two gaps in the literature: a) most of the studies have been carried out in small or medium-sized classes at schools (Hattie and Yates 2013) but not in large classes and b) the existing studies have not investigated the simultaneous use of internal assessment such as SA and external assessment such as CbA.

Hence, empirically founded knowledge is needed to show the extent to which technology mediated formative feedback has an impact on student performance. Testing the applicability of design theory in practice is a highly important step in the design process (Rosemann and Vessey 2008). Hence, I tested the IT artifact,

based on the design knowledge of RQ1, in a field experiment. The results showed that students using the IT artifact had significantly higher objective and perceived student performance.

In most studies the independent and dependent variables relating to IT artifacts were tested through surveys or user interviews. However, these methods are prone to *common method variance* if the same methods are used for testing the independent and the dependent variables (Sharma et al. 2009). Hence, this study provides more robust insights by using two independent methods for measuring the variables. To measure the independent variable, system-captured data is used while the dependent variable uses perceptually anchored data. The combination of this method-method pair has the least susceptibility to common-method variance (Sharma et al. 2009). Additionally, I tested the system in a longitudinal field experiment with eight time-points.

This research adds empirically grounded knowledge to the IS-body of literature by providing valuable insights into the use of the kernel theory of cognitive dissonance to formative feedback. The presented model sheds light on how formative feedback based on internal and external assessment can impact student performance. Thus, this research contributes to theory by moving the rather “hidden” element of internal assessment (such as SA) to the “center” as an equally important element with external assessment (such as CbA).

Additionally, the fact that through the use of the IT artifact, objective and perceived student performance was increased by giving feedback, this study has implications for the larger field of the perception of assessment in universities. The results suggest that the interplay between the students’ own perception and an objective measurement of performance is beneficial. Thus, it is not only important that objective performance increases but that students also perceive their learning to have increased. The increasing importance of, and focus on, higher-order thinking skills such as self-reflection has implications for researchers conducting studies in this area. Additionally, this study supports the call for defining feedback as a process rather than providing feedback once, which often has only a summative purpose (Boud and Molloy 2013). The IT artifact was used eight times within the field experiment. The findings of a positive coefficient implied that using the system more frequently led to higher performance in terms of objective and perceived student performance.

However, the important aspect of the experiment was that the use of the tool was not mandatory but rather was an option to test students' current state of knowledge within the learning process. This has implications for how lecturers design teaching-learning environments. The focus today often lies on a summative final grade rather than providing formative feedback during the learning process. Thus, the results support the need for a shift in mindset of how assessments are perceived, changing from being viewed as a one time event to an ongoing process.

### **9.3 Prescriptive Knowledge of Measuring Peer Feedback Quality**

The higher-order thinking skill of providing feedback is of particular importance to students working in agile teams. Feedback on a person's performance or behaviour from a fellow team member can help the person to adjust their actions and steer them in the right direction (Whitworth and Biddle 2007). According to Hattie and Timperley (2007), there are two essential factors for developing higher-order thinking skills such as the ability to provide feedback: Factor A, a realistic teaching-learning scenario that allows students to apply the skills that need to be developed, and Factor B, the provision of rich and meaningful feedback of the students' progress. Factor A is addressed by using the concept of peer feedback whereby students provide each other feedback. However, Factor B remains challenging. If one wants to provide individualized formative feedback, the first steps are to understand what the characteristics of peer feedback are, how to determine its quality and how to measure them automatically.

Therefore I conducted studies measuring peer feedback quality from different angles. I applied existing knowledge from research into online reviews to the peer feedback body of literature, derived characteristics of peer feedback quality from the literature and real-world data, and operationalized these features to measure peer feedback quality automatically. Finally, I conducted a study on how to design an IT artifact for providing peer feedback in large-scale classes taken by students from multiple cultural backgrounds.

In the academic literature the characteristics determining high and low quality peer feedback have been defined in a very generic way. There is no comprehensive study that has operationalized peer feedback characteristics and measured them empirically using real-world data. However, being able to measure peer feedback, is one of the most important factors for providing automated feedback to develop the skill of providing peer feedback. Carless and Boud (2018) argues for the importance of feedback literacy and for feedback to be seen as a process rather than a one-time event. Students who develop their feedback literacy appreciate their own active role in the feedback process and over time improve their skill of providing and receiving feedback.

Without having empirically grounded knowledge of how to measure peer feedback quality, providing students with formative feedback is not possible in large-scale classes. Students need individualized recommendations of how to improve their peer feedback; if these are missing then it is very unlikely that students will develop the skill of providing high quality peer feedback (Tai et al. 2016; Patton 2012).

The studies conducted for this dissertation show that online reviews can provide valuable insights into how to measure peer feedback quality. The literature review study revealed that most research was carried out on ratings and reputation whereas reader and environmental-related factors were highly under-researched. The most studied factor was review length, which was also found to be the most important factor in my empirical analysis of 7,036 peer feedbacks. In addition to feedback length, the number and type of suggestions that are provided are also important, such as whether they are communicated using the *I-form* or *You-form*. The study on designing culturally sensitive peer feedback showed that students from individualistic or collectivist culture have different preferences such as whether they want the same reviewer in each peer feedback round or prefer the reviewer to change.

The study on online review helpfulness has implications for the body of knowledge on online reviews. The results showed that research in the top IS journals neglected reader- and environment-related factors of online review helpfulness. Thus, there may be hidden factors having an influence on how feedback receivers perceive feedback quality which have up to now been concealed.

Furthermore, feedback receivers have usually been treated as a homogeneous group, without consideration of receivers' individual characteristics. However, these characteristics form their psychological profiles, which include their level of risk aversion or open mindedness, their current stage in the search process (e.g. initial search or already comparing products), and underpin the biases of that might be primed by the first read reviews (e.g. latest reviewer talks pretty badly about the service quality). All of these features may have a significant impact on perceived review helpfulness. Thus, unveiling the individual characteristics of review providers and review receivers is an important next step.

The study on measuring peer feedback quality has multiple implications. The primary implication of this study is that the skill of providing feedback should be introduced as a core element in peer learning, which is of high importance for students graduating and subsequently working in agile teams or management positions. Most research and discussion within higher education focus only on the aspect of grading and neglect the necessity of formative feedback to develop higher-order thinking skills. Thus, this study provides valuable insights for the ongoing discussion about whether automated essay scoring or calibrated peer feedback is better. In most cases, students receive feedback on their final essay and, if they are lucky, they receive qualitative feedback but often nothing is done with it. Hence, the results of these studies support the arguments of Carless and Boud (2018) that formative feedback is a process in which the provider and the receiver are responsible for the outcome.

Thus, this dissertation proposes a hybrid model of intelligent thinking, in which the human mind is needed to evaluate creative tasks, such as essay writing, and artificial intelligence is used to help the person write the feedback in the most helpful way. This has implications for the underlying theoretical concepts of peer learning and student performance, which are used in MOOCs as well as in large-scale classes. The findings of this study provide the foundation for the development of such a hybrid intelligence teaching-learning environment.

Furthermore, the developed text features provide novel insights into what students really consider to be helpful feedback. Most of the carried out studies have investigated what is high and low quality feedback based on the lecturers' perspective. However, this does not necessarily mean that students share the

lecturers' view. Thus, this study has implications for lecturers regarding how they provide feedback to students.

The final study on culturally sensitive peer feedback has implications for the design of teaching-learning scenarios using peer learning. The research by Reinecke and Bernstein (2013) shows the importance of adapting user interfaces to different cultures. I build on this research and show that not only does the user interface need to be adapted but so does the whole peer feedback process. This has implications for which information is provided, including the reviewer selection process, competitiveness, instructions and interactions between the students.

Thus, if a student with a Chinese cultural background receives feedback from a reviewer with a Swiss cultural background, then the Chinese student would require much more detailed information about the Swiss student. This includes, for example, how many feedbacks the Swiss student has already written, what the quality of these feedbacks were, and whether the Swiss student has enough knowledge to provide correct and high quality feedback. These factors have strong implications for trust building mechanisms and for the acceptance of the feedback.

### **9.4 Effects of Technology-Mediated Formative Feedback on the Students' Perceptions of their Skill at Providing Peer Feedback**

As well as providing objective formative feedback on the students' ability to provide feedback, the students' perception of their own skill development is important (Alavi 1994). With regard to the perceived skill development, the formative feedback consisted of an internal and external assessment of the student's own perceived skill level. If these two assessments differed to a specific degree, then this would lead to cognitive dissonance and, hence, initiate self-reflection learning processes.

There is a gap in the literature concerning the effects of providing a student with these specific forms of formative feedback: internal and external assessment on the perceived development of the ability to provide feedback. To the best of

my knowledge, there is currently no research which has explored the effects of presenting students with these two assessments on their ability to provide feedback. The results showed that students participating in the ITPF process had significantly higher perceived ability to provide feedback from participating only once and multiple times.

While previous research has exclusively examined whether peer grading is similar to expert grading, the proposed “perceived skill development” extends the literature and opens new research areas. For example, it would be important to investigate how this technique can be used to develop the skill of providing high quality peer feedback. This extension is particularly relevant in today’s increasingly large-scale classes such as at Universities and in MOOCs. In this context, my research shows that technology-mediated formative feedback consisting of an internal and an external assessment of the perceived skill development could have implications not only for the skill of providing feedback but also for the ability to receive the feedback.

Furthermore, this research provides valuable insights into, and provides new knowledge, for the theory of cognitive dissonance. Applying the theoretical lens of cognitive dissonance provides an explanation as to why the presentation of an internal assessment and an external assessment might initiate self-reflection processes, encouraging the student to close this gap in next rounds of peer feedback. Even if this finding has been herein shown to apply to a business class at a university, there is little reason to doubt its applicability to other contexts, such as MOOCs or online reviews. Hence, the findings that perceived skill development can be developed using peer learning concepts and the provision of formative feedback consisting of an internal and external assessment can be generalized to further contexts.

Furthermore, the conducted study provides valuable insights for the design of teaching-learning scenarios. The results show that formative feedback has a long-term effect on the students’ ability to provide feedback, which suggests that developing the ability to provide feedback needs to be practiced over time. Thus, by showing how the perceived skill development improves over time, this study contributes to the IS literature where longitudinal studies are rare.





# 10

## **Practical Implications**

## 10.1 Designing Formative Feedback Tools

Despite the large number of available learning management systems (Blackboard<sup>1</sup>, Canvas<sup>2</sup>, Moodle<sup>3</sup>, etc.) or MOOCs using IT-enabled assessment (Coursera<sup>4</sup>, edX<sup>5</sup>, Udacity<sup>6</sup>), or available apps (such as Kahoot<sup>7</sup>), there is still no general knowledge available about how such tools should be designed. Hence, as a first step, the conducted study provides an initial set of design principles for formative feedback IT artifacts.

The results revealed that users tried to access the tools from various devices such as smartphones, tablets or laptops. Therefore, IT tools needed to be adapted to different screen sizes. The study provided valuable insights into how the students used the user interface and what the challenges were. One finding was that students who conducted a flow diagram using a tablet or smartphone were more likely to accidentally click in the screen and thus lose their first attempt than students who used a laptop. Hence, one of the design principles coming from that finding was that the first click should focus on the flow diagram and the second click should actually be used as an attempt.

Furthermore, the proposed design principle using *hierarchical structures* to organize courses has implications for practitioners designing their teaching-learning scenario. The hierarchical structure of a course, learning units and learning outcomes provide practitioners with an aggregated view of student performance. This provides practitioners with the possibility of comparing courses from year to year (course level) as well as deriving actions about which learning units students struggled with (learning unit level) and even finer granularity regarding which learning outcomes the students had most problems with (learning outcome level). This enables practitioners to derive actions immediately; for example, lecturers can decide, in the next learning session, to rehearse content related to the learning outcomes with which the majority of the students found difficult.

---

<sup>1</sup> [www.blackboard.com](http://www.blackboard.com)

<sup>2</sup> [www.canvas.net](http://www.canvas.net)

<sup>3</sup> [www.moodle.org](http://www.moodle.org)

<sup>4</sup> [www.coursera.org](http://www.coursera.org)

<sup>5</sup> [www.edx.org](http://www.edx.org)

<sup>6</sup> [www.udacity.com](http://www.udacity.com)

<sup>7</sup> [www.kahoot.com](http://www.kahoot.com)

A further design principle is important for trust building. Since lecturers are only able to view the average performances of the whole course, students keep their anonymity and are not concerned that the use of the system might influence their final grade. Overall, the design knowledge arising from these studies has direct implications for practitioners, since the research provides directly usable and empirically tested design principles, which enables practitioners to instantiate their own IT artifact.

## **10.2 Improving Student Performance**

The conducted research in this dissertation shows that there is no single silver bullet for feedback. Instead, it shows that, to improve student performance, the teaching-learning scenario needs to be created and adapted according to the target learning outcome. The studies shed valuable insights into how to use the concept of technology-mediated learning and peer learning to develop higher and lower-order thinking skills. In particular, this research focused on the development of self-reflection skills and self-regulated learning through the use of internal and external assessments. The gap between the results of those two assessment is important, as it reveals that the students' perceived skill development differs to their objective skill development, which ultimately triggers cognitive dissonance and the desire to reduce this internal conflict.

Overall, the precise nature of the formative feedback and the channel of delivery needs to be adjusted according to the skills to be learned. This has important implications for practitioners. The lecturers need to plan carefully in advance what the learning outcomes of each learning unit and the overall course are. The learning outcomes are the determining factors for the IT-tool and the teaching-learning environment. Thus, this research supports the transition from a tool and lecturer-centered mindset to a more learning outcome and, therefore, student-centered mindset. This is achieved by using a strict learning outcome oriented approach. Furthermore, the link between the learning outcomes and the Blooms' revised learning Taxonomy (Krathwohl 2002) provides the lecturer valuable insights into how many learning outcomes are on which level. Hence, if a lecturer thinks that in the course students should learn higher-order thinking skills that are needed in their future career, then this should be reflected by the

Learning Taxonomy with the number of learning outcomes coming from the levels of *analyzing*, *evaluating* and *creating*.

### 10.3 Providing Guidelines for Peer Feedback

Despite the potentially important role of feedback in learning, most available practical guidelines on high and low quality feedback are still rather generic, lack knowledge of implementation and neglect the consideration of different cultural backgrounds. Thus, the studies conducted in this dissertation provide practical guidelines on how to measure peer feedback quality and how to design culturally sensitive IT artifacts.

The studies have implications for the ongoing education of future managers with regard to how to provide high quality feedback. There are hundreds of published books that try to provide guidelines for managers on how to provide feedback. However, most of the guidelines have not been empirically tested using scientific methods. Thus, those guidelines may contain claims that may not be true anymore.

For example, one of those claims emphasises the importance of the sandwich method, which states that negative statements should be positioned between two positive statements. However, I was not able to find evidence for the sandwich method. In contrast, my research suggests that feedback should not extensively flip between positive and negative statements. Hence, the conducted study provides valuable evidence of alternative, and more optimal, ways of delivering feedback than the the sandwich method.

Furthermore, a second claim within the literature is that feedback should always be communicated using the *I-form*. This means that the sentences of the feedback are written in the *I* perspective. However, the results provide evidence for the use of a much finer granular use of the *I-form* and the *You-form*. The *You-form* is especially important when communicating suggestions. Hence, these findings have important implications for management education and developing materials to teach managers about how to provide high qualitative feedback.

The study on culturally sensitive peer feedback provides directly usable design guidelines on how to build a culturally sensitive IT artifact using the concept

of peer feedback. The study highlights that not only is the adaption of the user interface important but so is the adaptation of the whole process of peer feedback. This includes the selection of the reviewers, the information provided on the reviewers, the provision of grades or pass and fail etc. The provided design guidelines should help practitioners improve students' engagement and increase their satisfaction.

Overall, the study has implications for graduates working in areas that need communication, such as agile team members, product managers, project manager etc., as they need to be able to provide constructive feedback as well as being able to receive feedback. For example, project managers need to provide constructive feedback to help steer projects in the right direction. They also need to receive feedback and derive actions from that feedback, such as receiving feedback from a customer about a current product. Product managers need to be able to assimilate such feedback and derive actions to improve their products and services.



# 11

## **Future Research**

### 11.1 A Design Theory of Formative Feedback IT Artifacts

Creating a design theory is a challenging and complex task. One cannot assume that a holistic design theory addressing a general class of problem like designing *formative feedback* IT artifacts can be developed with one study. It is rather the case that the accumulation of design knowledge across multiple studies over time is needed to form a holistic design theory. This dissertation is contributing to this ongoing process.

I derived requirements and developed design principles based on scientific knowledge and interviews with lecturers and students. The results of this study provide initial insights into what the design principles of formative feedback IT artifacts should be. However, further interviews are needed to improve their generalizability. Lecturers and students from other universities, operating in different contexts and working in different areas such as natural sciences and humanities need to be interviewed to develop a more holistic and generalizable design theory on formative feedback IT artifacts.

Furthermore, the IT artifact was tested as an encapsulated tool, meaning that the students used the tool as one unit. Testing the design principles as individual units would provide further insights into its usefulness, especially to be able to answer the following questions. Which design principles are most valuable? Which principles have the highest impact on objective and perceived student performance? These questions could be researched in a field experiment, in which students from different courses use different compositions of the IT artifact. A pre-test would be necessary to check that both groups are not different from each other in terms of critical characteristics such as pre-knowledge and expectations of formative feedback.

Furthermore, the developed IT artifact is viewed from two main perspectives, firstly the lecturer view and secondly the student view. The current studies focused exclusively on the students' view, with the aim to evaluate whether the use of the IT artifact leads to higher objective and perceived student performance. In a further study the lecturer's view should be investigated. The aim could be to find out whether the integration of the revised Bloom's Taxonomy (Krathwohl 2002) and the link of the learning outcomes and assessments to this taxonomy actually support the lecturers' planning process. Moreover, it is important to de-



termine whether the hierarchical structure actually leads to lecturers' rehearsing content that the feedback revealed was not understood by the students.

## 11.2 Developing the Feedback Providing Skill Holistically

Peer learning and consequently peer feedback are providing the foundation for developing higher-order thinking skills. Students participating in this process are able to put the learned skills into action. Students need to analyse the submissions of their peers and provide feedback as to what was good, what was not good and how to improve the submission overall. This dissertation focuses on one very specific higher-order thinking skill, namely the *ability to provide feedback*. This skill is of significant importance for members of agile teams, product or project managers and leaders more generally. The conducted research provides a contribution to the research field and to higher education practitioners, as it developed a teaching-learning scenario which enables students to develop this skill in large-scale university lectures. Nevertheless, there are still open questions, with further research needed in this field.

### 11.2.1 Feedback Writing or Feedback Receiving?

Carless and Boud (2018) argues that students benefit more from feedback writing than feedback receiving. Evaluating a student's submission, which involves trying to find its strengths and weaknesses, developing ideas on how to overcome those weaknesses, and conveying this with qualitative feedback, is a highly cognitively engaging task. The cognitive thinking processes that are involved form the foundation for learning based on principles of constructivism. It would be very interesting to carry out future work on measuring the effects of feedback writing and feedback receiving separately in order to understand which has a higher impact on the development of higher-order thinking skills.

Furthermore, the distinction between feedback receiving and writing is interesting when considering one specific higher-order thinking skill, namely the skill of providing high quality feedback. An interesting research question would be: does this skill improve more strongly by reading feedback or by writing it? This could be investigated using a 2x2 field experiment with a randomized control group. Group one would not participate in the peer feedback process and would

only upload a submission. Group two would only write feedback but would not receive any feedback. Group three would only receive feedback without writing any. Group four would write and receive feedback. This study would contribute to answering the question as to whether feedback writing or receiving has a higher impact on the development of feedback providing skills.

### 11.2.2 Individualized Recommendations

AES and CPR both have strengths and weaknesses. The strength of AES is that it can scale and therefore can be used in large-scale classes with no need for human-based resources. The main weakness of AES is, up to now, its limited ability to grade creative tasks such as essay writing (Attali and Burstein 2006). CPR has the benefit of humans scoring the essays, who are seen to be better at evaluating creativity. The weakness of CPR is that it is still human-resource intensive. The lecturer needs to calibrate the students' level of skill to that of the lecturer's, which often involves multiple rounds (Balfour 2013).

Overall, both approaches have the main drawback that they are summative oriented and provide only grades without qualitative formative feedback which is important for developing higher-order thinking skills. Thus, the idea is to provide a take a hybrid intelligence approach, which combines human and artificial intelligence. Human intelligence is used to evaluate the submission and write a qualitative feedback while artificial intelligence is used to support the student by helping them write feedback in the most optimal and helpful way. This is achieved by automatically analyzing the peer feedback in close-to real-time and providing individualized recommendations about what can be changed to increase the quality of the feedback.

My study on measuring peer feedback quality sets the foundation for this approach. Thus, the next step is to use the developed text features to measure peer feedback quality and train an artificial intelligent algorithm. One of the challenges of writing this algorithm is how to create meaningful recommendations. For example, if 20 features are low on feedback quality, it might be overwhelming to provide 20 individualized recommendations to the feedback provider. This overload of information might cause all recommendations to be rejected. Furthermore, the recommendations need to be written in a way such

that they are perceived to be helpful, trustworthy and not too authoritative. In a field experiment the hybrid intelligence approach could be tested against the existing approach, when students only carry out an internal assessment involving self-assessment and an external assessment involving peer-assessment.

### 11.2.3 How Students Process Feedbacks

In the past decade a new approach called NeuroIS has emerged, which focuses on how people process information (Dimoka et al. 2012). For example, it asks: what are the readers' emotions while processing information, which brain areas are stimulated when reading information and how do humans absorb information provided by user interfaces? The aim of NeuroIS is to use approaches from cognitive neuroscience and apply them to problems within the IS field (Dimoka et al. 2011). NeuroIS involves neuroscience tools like eye tracking, functional Magnetic Resonance Imaging (fMRI) or Electroencephalography (EEG) (vom Brocke et al. 2013).

In my research I derived requirements for high and low quality peer feedback based on the literature and real-world data. Additionally, it would be interesting to measure how students read and process the feedback they receive. This could be investigated through performing an experiment in which students read feedback while wearing an eye-tracking and EEG device. One interesting research question would be to what extent do the first paragraphs of a feedback cause priming effects on readings of later paragraphs? Thus, when feedback starts with a negative statement, or a statement that the feedback receiver does not agree to, does this influence how later feedback statements are read? This could lead students not to accept feedback statements that are in the middle or end of a feedback, even if they are of high relevance and importance.

Further interesting research questions to investigate relate to whether students have a limited capacity to absorb feedback statements and how students respond to generic or specific feedback. Is generic or specific feedback equally helpful or do students prefer one over the other? Hence, using advanced technologies within the field of neuroscience can open up a whole new research stream and can stimulate researchers to conduct further research on feedback.



# 12

## **Conclusion**

This dissertation was conducted to understand how to enhance student performance through technology-mediated formative feedback in large-scale university lectures. To answer the main research question, it was subdivided into four more specific research questions. Question one addressed the design of the formative feedback IT artifacts. Question two addressed the measurement of the effects of formative feedback on student performance. Question three addressed the design of high and low peer feedback quality. Question four addressed the perceived development of the skill of providing peer feedback.

To categorize these four questions, I developed a two by two framework (see Figure 1.1). Questions one and three focused on design and questions two and four focused on the measurement of effects. For questions one and two, formative feedback was operationalized as the combination of SA for the internal assessment and CbA for the external assessment while, for questions three and four, SA was used for the internal assessment but PA was used for the external assessment. To answer these four research questions I conducted five research studies.

The study to answer research questions one and two involved the creation of design knowledge and the empirical test of an instantiation based on this design knowledge in a large-scale class. I followed the DSR approach of Hevner (2007) and conducted six interviews with lecturers and students. Based on the interviews and the scientific literature I derived nine requirements and developed seven design principles, which were important for the design of the formative feedback IT artifacts. This study makes an important, initial contribution to developing a holistic theory of design and action which addresses the general problem of designing *formative feedback tools*.

To measure the effects of the formative feedback consisting of SA and CbA on the student performance, I tested the instance of the design knowledge named LOOM in a large-scale class with 192 Master's students. LOOM was used four times within the lecture and four times after the lecture. The results showed that the use of the IT artifact had a significant influence on objective and perceived student performance. This study contributes to the research literature by moving the rather "hidden" element of internal assessment to the "center" and positioning it as equally important to external assessment. Thus, this study highlights the importance of both internal and external assessment.

---

In cases where the outcomes of these two assessments differ to a significant degree, students will experience cognitive dissonance. In trying to resolve this dissonance, self-reflection processes will be initiated which are fundamental to developing higher-order thinking skills.

To determine the characteristics of high and low peer feedback quality, I carried out three studies. The *first* study aimed to analyse the existing body of literature on the related research stream of *online review helpfulness*. I analysed 81 journal articles from the eight most important IS journals using a systematic literature review approach. The results revealed that review- and reviewer-related factors were the most researched factors. In contrast, reader-related and environment-related factors have been highly neglected. Thus, these two factors provide room for further research. This study provides valuable contribution for researchers in the field of online reviews, through synthesizing the field and providing an in-depth overview of the investigated factors. I developed a new method called co-topic analysis, which shows which of the topics are researched together and how often (e.g., a topic is a review-related factor).

The *second* study aimed to determine the characteristics of high and low feedback quality. I derived requirements for high quality feedback, based on the research literature and a real-world data set consisting of 7,036 peer feedbacks. I created 90 text features in total. I used panel OLS regression with fixed effects to measure which of these text features had a significant influence on peer feedback quality. The results showed that 66 text features had a significant effect on peer feedback quality. The most important text features were text length, the number of suggestions the feedback provided and the form in which the feedback was written (i.e., *I-form*, *You-form* or a combination of both). This study contributes a set of prescriptive knowledge containing information about which text features have positive, negative or no effects on peer feedback quality. This set of features built the foundation for the later development of an artificial intelligence algorithm which automatically analysed peer feedback quality and provided individualized recommendations.

The *third* study aimed to develop design principles for a culturally sensitive peer feedback IT artifact. I followed the DSR approach taken by Peffers et al. (2007) and developed design principles based on the cultural dimension theory described by Hofstede and Hofstede (2005). Overall, I developed five design

principles and evaluated them by conducting 16 semi-structured interviews. The results showed a clear difference between the preferences of students with a Chinese or Swiss cultural background. One of the most important differences was their feedback expectations. Students with a Chinese cultural background expected to have much more information about the whole peer feedback process than students with a Swiss cultural background. This study contributes prescriptive knowledge and contributes to the building of a holistic design theory to address the general class of problems related to designing *formative feedback IT artifacts*.

The study to answer research question four measured the effects of formative feedback, consisting of an internal and an external assessment, on the students' perceived ability to provide feedback to their peers. I conducted a quasi-experiment in which around 100 students used an IT artifact, to take part in a peer feedback process. This process ran seven times. The results showed that the students' perceived ability to provide feedback increased over time in terms of their SA ability to provide feedback as well as the students' ability to provide peer feedback as assessed by their peers. The findings emphasized that not only the objective but also the perceived development of the skill of providing high quality feedback is important. Therefore, this study provides valuable insights into how this skill can be developed using the IT-based peer feedback process.



## References

- Aggarwal, A. 2016. "A Hybrid Approach to Big Data Systems Development," in *Managing Big Data integration in the public sector*, , University of Baltimore, MD, USA: IGI Global.
- Agnihotri, A., and Bhattacharya, S. 2016. "Online review helpfulness: Role of qualitative factors," *Psychology & Marketing* (33:11), pp. 1006–1017.
- Ahmad, S. N., and Laroche, M. 2015. "How do expressed emotions affect the helpfulness of a product review? Evidence from reviews using latent semantic analysis," *International Journal of Electronic Commerce* (20:1), pp. 76–111.
- Alavi, M. 1994. "Computer-Mediated Collaborative Learning: An Empirical Evaluation," *MIS Quarterly* (18:2), pp. 159–174.
- Alavi, M., and Leidner, D. E. 2001. "Research commentary: Technology-mediated learning - A call for greater depth and breadth of research," *Information Systems Research* (12:1), pp. 1–10.
- Alavi, M., Yoo, Y., and Vogel, D. R. 1997. "Using information technology to add value to management education," *Academy of Management Journal* (40:6), pp. 1310–1333.
- Allen, I. E., and Seaman, J. 2015. "Grade Level: Tracking Online Education in the United States," . (<https://www.babson.edu/academics/faculty/provost/babson-survey-research-group/>, accessed December 30, 2019).
- Alred, G. J., Brusaw, C. T., and Oliu, W. E. 2015. *Handbook of technical writing*, Boston, MA, USA: Bedford/St. Martin's, eleventh edition ed.

## REFERENCES

---

- Anderson, L. W., Krathwohl, D. R., and Bloom, B. S. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, New York, NY: Addison-Wesley Longman.
- Arbaugh, J. B. 2001. "How instructor immediacy behaviors affect student satisfaction and learning in web-based courses," *Business Communication Quarterly* (64:4), pp. 42–54.
- Attali, Y., and Burstein, J. 2006. "Automated essay scoring with e-rater® V. 2," *The Journal of Technology, Learning and Assessment* (4:3), pp. 1–31.
- Attali, Y., and van der Kleij, F. 2017. "Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving," *Computers & Education* (110), pp. 154–169.
- Ba, S., and Pavlou, P. A. 2002. "Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior," *MIS Quarterly* (26:3), pp. 243–268.
- Backhaus, K., Erichson, B., and Weiber, R. 2015. *Fortgeschrittene Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*, Berlin and Heidelberg: Springer Gabler, 3rd ed.
- Bae, S., and Lee, T. 2011. "Product type and consumers' perception of online consumer reviews," *Electronic Markets* (21:4), pp. 255–266.
- Baek, H., Ahn, J. H., and Choi, Y. 2013. "Helpfulness of online consumer reviews: Readers' objectives and review cues," *International Journal of Electronic Commerce* (17:2), pp. 99–126.
- Baek, H., Lee, S., Oh, S., and Ahn, J. 2015. "Normative social influence and online review helpfulness: Polynomial modeling and response surface analysis," *Journal of Electronic Commerce Research* (16:4), pp. 290–306.
- Balfour, S. P. 2013. "Assessing writing in MOOCs: Automated essay scoring and calibrated peer review (tm)," *Research & Practice in Assessment* (8:1), pp. 40–48.
- Ballantyne, R., Hughes, K., and Mylonas, A. 2002. "Developing procedures for implementing peer assessment in large classes using an action research process," *Assessment & Evaluation in Higher Education* (27:5), pp. 427–441.

- Bamberger, R., and Vanecek, E. 1984. *Lesen, verstehen, lernen, schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache [Reading, understanding, learning, spelling: The degrees of difficulty of texts in German language]*, Vienna: Jugend und Volk.
- Bandura, A. 1982. "Self-efficacy mechanism in human agency," *American Psychologist* (37:2), pp. 122–147.
- Banerjee, S., Bhattacharyya, S., and Bose, I. 2017. "Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business," *Decision Support Systems* (96), pp. 17–26.
- Banerjee, S., and Chua, A. 2014. "A theoretical framework to identify authentic online reviews," *Online Information Review* (38:5), pp. 634–649.
- Baranik, L. E., Barron, K. E., and Finney, S. J. 2007. "Measuring goal orientation in a work domain: Construct validity evidence for the 2 × 2 framework," *Educational and Psychological Measurement* (67:4), pp. 697–718.
- Barbaresi, A. 2018. "A corpus of German political speeches from the 21st century," in *Proceedings of European Language Resources Association (ELRA)*, , Miyazaki, Japan.
- Bauman, K., and Tuzhilin, A. 2018. "Recommending Remedial Learning Materials to the Students by Filling their Knowledge Gaps," *MIS Quarterly* (42:1), pp. 313–332.
- Baumeister, R. F., Campbell, J. D., Krueger, J. I., and Vohs, K. D. 2003. "Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles?" *Psychological Science in the Public Interest* (4:1), pp. 1–44.
- Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., and Jeffries, R. 2001. "Manifesto for agile software development," . (<http://agilemanifesto.org>, accessed December 30, 2019).
- Bennis, W., and Nanus, B. 1985. *The strategies for taking charge*, New York, NY, USA: Harper and Row.

## REFERENCES

---

- Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., and Trautwein, U. 2017. "Reading Demands in Secondary School: Does the Linguistic Complexity of Textbooks Increase With Grade Level and the Academic Orientation of the School Track?" *Journal of Educational Psychology* (110:4), pp. 518–543.
- Biggs, J. 1999. "What the student does: teaching for enhanced learning," *Higher Education Research & Development* (18:1), pp. 57–75.
- Biggs, J. B. 2011. *Teaching for quality learning at university: What the student does*, Berkshire: McGraw-Hill Education (UK).
- Bitzer, P., Söllner, M., and Leimeister, J. M. 2016. "Design principles for high-performance blended learning services delivery," *Business & Information Systems Engineering* (58:2), pp. 135–149.
- Black, P., and Wiliam, D. 1998. "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice* (5:1), pp. 7–74.
- Black, P., and Wiliam, D. 2009. "Developing the theory of formative assessment," *Educational Assessment, Evaluation and Accountability* (formerly: *Journal of Personnel Evaluation in Education*) (21:1), pp. 5–31.
- Blanton, H., Buunk, B. P., Gibbons, F. X., and Kuyper, H. 1999. "When Better-Than-Others Compare Upward: Choice of Comparison and Comparative Evaluation as Independent Predictors of Academic Performance," *Journal of Personality and Social Psychology* (76:3), pp. 420–430.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent dirichlet allocation," *Journal of Machine Learning Research* (3:Jan), pp. 993–1022.
- Bloom, B. S. 1956. *Taxonomy of educational objectives. Vol. 1: Cognitive domain*, Ann Arbor, MI: Longmans.
- Bloom, B. S., Krathwohl, and Masia, B. B. 1984. *Bloom taxonomy of educational objectives*, Boston, MA: Allyn and Bacon.
- Boud, D. 1990. "Assessment and the promotion of academic values," *Studies in Higher Education* (15:1), pp. 101–111.
- Boud, D. 1991. *Implementing student self-assessment*, Canberra: Higher Education Research and Development Society of Australia, 2nd ed.

- Boud, D. 2013. *Enhancing learning through self-assessment*, Abingdon: Routledge.
- Boud, D., Cohen, R., and Sampson, J. 1999. "Peer learning and assessment," *Assessment & Evaluation in Higher Education* (24:4), pp. 413–426.
- Boud, D., Cohen, R., and Sampson, J. 2014. *Peer Learning in Higher Education: Learning from and with Each Other*, Abingdon: Routledge.
- Boud, D., and Falchikov, N. 2007. *Rethinking assessment in higher education: Learning for the longer term*, Abingdon: Routledge.
- Boud, D., and Molloy, E. 2013. "Rethinking models of feedback for learning: The challenge of design," *Assessment & Evaluation in Higher Education* (38:6), pp. 698–712.
- Breiman, L. 2001. "Random forests," *Machine Learning* (45:1), pp. 5–32.
- Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z., and Wong, F. M. F. 2014. "Learning about social learning in MOOCs: From statistical analysis to generative model," *IEEE Transactions on Learning Technologies* (7:4), pp. 346–359.
- Brookhart, S. M. 2008. *How to give effective feedback to your students*, Alexandria, VA, USA: Association for Supervision and Curriculum Development.
- Brown, M. 2013. "Moving Into the Post-MOOC Era," . (<https://er.educause.edu/blogs/2013/7/moving-into-the-postmooc-era>, accessed 14.08.2019).
- Brown, P., and Levinson, S. C. 1987. *Politeness: Some universals in language usage*, Cambridge: Cambridge University Press.
- Buholzer, F., Rietsche, R., and Söllner, M. 2018. "Knowing what Learners Like? Developing a Cultural Sensitive Peer Assessment Process in MOOCs," in *Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI)*, , Lüneburg, Germany.
- Butler, D. L., and Winne, P. H. 1995. "Feedback and Self-Regulated Learning: A Theoretical Synthesis," *Review of Educational Research* (65:3), pp. 245–281.

## REFERENCES

---

- Cao, Q., Duan, W., and Gan, Q. 2011. "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach," *Decision Support Systems* (50:2), pp. 511–521.
- Carbone, E., and Greenberg, J. 1998. "Teaching large classes: Unpacking the problem and responding creatively," *To improve the academy* (17:1), pp. 311–326.
- Carless, D., and Boud, D. 2018. "The development of student feedback literacy: enabling uptake of feedback," *Assessment & Evaluation in Higher Education* (43:8), pp. 1315–1325.
- Carless, D., Salter, D., Yang, M., and Lam, J. 2011. "Developing sustainable feedback practices," *Studies in Higher Education* (36:4), pp. 395–407.
- Carlsson, S. A., Henningsson, S., Hrastinski, S., and Keller, C. 2011. "Socio-technical IS design science research: developing design theory for IS integration management," *Information Systems and e-Business Management* (9:1), pp. 109–131.
- Casaló, L. V., Flavián, C., Guinalíu, M., and Ekinci, Y. 2015. "Avoiding the dark side of positive online consumer reviews: Enhancing reviews' usefulness for high risk-averse travelers," *Journal of Business Research* (68:9), pp. 1829–1835.
- Chan, T. Y., Li, J., and Pierce, L. 2014. "Learning from Peers: Knowledge Transfer and Sales Force Productivity Growth," *Marketing Science* (33:4), pp. 463–484.
- Chapman, P., Clinton, J. M., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R., and Wirth, R. 2000. "CRISP-DM 1.0: Step-by-step data mining guide," *SPSS inc* (16).
- Chemers, M. M., Hu, L.-t., and Garcia, B. F. 2001. "Academic self-efficacy and first year college student performance and adjustment," *Journal of Educational Psychology* (93:1), pp. 55–64.
- Chen, C. C., and Tseng, Y.-D. 2011. "Quality evaluation of product reviews using an information quality framework," *Decision Support Systems* (50:4), pp. 755–768.

- Chen, Y., and Xie, J. 2008. "Online consumer review: Word-of-mouth as a new element of marketing communication mix," *Management Science* (54:3), pp. 477–491.
- Cheng, Y.-H., and Ho, H.-Y. 2015. "Social influence's impact on reader perceptions of online reviews," *Journal of Business Research* (68:4), pp. 883–887.
- Cheung, C. M.-Y., Sia, C.-L., and Kuan, K. K. Y. 2012. "Is this review believable? A study of factors affecting the credibility of online consumer reviews from an ELM perspective," *Journal of the Association for Information Systems* (13:8), pp. 618–635.
- Cho, K., Cho, M.-H., and Hacker, D. J. 2010. "Self-monitoring support for learning to write," *Interactive Learning Environments* (18:2), pp. 101–113.
- Cho, K., and MacArthur, C. 2010. "Student revision with peer and expert reviewing," *Unravelling Peer Assessment* (20:4), pp. 328–338.
- Choi, J. D., Tetreault, J., and Stent, A. 2015. "It depends: Dependency parser comparison using a web-based evaluation tool," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, , Beijing, China.
- Chong, A. Y. L., Khong, K. W., Ma, T., McCabe, S., and Wang, Y. 2018. "Analyzing key influences of tourists' acceptance of online reviews in travel decisions," *Internet Research* (28:3), pp. 564–586.
- Christensen, C. M., Horn, M. B., Caldera, L., and Soares, L. 2011. "Disrupting College: How Disruptive Innovation Can Deliver Quality and Affordability to Postsecondary Education," . (<https://www.americanprogress.org/issues/economy/reports/2011/02/08/9034/disrupting-college/>, accessed December 30, 2019).
- Chua, A. Y. K., and Banerjee, S. 2014. "Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth," *Journal of the Association for Information Science and Technology* (66:2), pp. 354–362.
- Chua, A. Y. K., and Banerjee, S. 2016. "Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality," *Computers in Human Behavior* (54), pp. 547–554.

## REFERENCES

---

- Colquitt, J. A., and Zapata-Phelan, C. P. 2007. "Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal," *Academy of Management Journal* (50:6), pp. 1281–1303.
- Cunningham, D., and Duffy, T. 1996. "Constructivism: Implications for the design and delivery of instruction," in D. H. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology*, , New York, NY, USA: Macmillan.
- Dai, G., de Meuse, K. P., and Peterson, C. 2010. "Impact of Multi-Source Feedback on Leadership Competency Development: A Longitudinal Field Study," *Journal of Managerial Issues* (22:2), pp. 197–219.
- Davis, J. M., and Agrawal, D. 2018. "Understanding the role of interpersonal identification in online review evaluation: An information processing perspective," *International Journal of Information Management* (38:1), pp. 140–149.
- Dawson, P., Henderson, M., Ryan, T., Mahoney, P., Boud, D., Phillips, M., and Molloy, E. 2016. "Technology and Feedback Design," in *Learning, design, and technology*, J. M. Spector, B. B. Lockee, and M. Childress (eds.), Switzerland: Springer International Publishing.
- Derks, D., Bos, A. E. R., and von Grumbkow, J. 2008. "Emoticons in computer-mediated communication: Social motives and social context," *CyberPsychology & Behavior* (11:1), pp. 99–101.
- Dickson, S. V., Simmons, D. C., and Kameenui, E. J. 1995. *Text organization and its relation to reading comprehension: A synthesis of the research*, Oregon: National Center to Improve the Tools of Educators, College of Education, University of Oregon.
- Dimoka, A., Davis, F. D., Gupta, A., Pavlou, P. A., Banker, R. D., Dennis, A. R., Ischebeck, A., Müller-Putz, G., Benbasat, I., and Gefen, D. 2012. "On the use of neurophysiological tools in IS research: Developing a research agenda for NeuroIS," *MIS Quarterly* (36:3), pp. 679–702.
- Dimoka, A., Pavlou, P. A., and Davis, F. D. 2011. "Research commentary—NeuroIS: the potential of cognitive neuroscience for information systems research," *Information Systems Research* (22:4), pp. 687–702.



- Dobre, I. 2015. "Learning Management Systems for Higher Education - An Overview of Available Options for Higher Education Organizations," *Procedia-Social and Behavioral Sciences* (180), pp. 313–320.
- Dubin, R. 1976. *Theory Building in Applied Areas*, Handbook of industrial and organizational psychology, Palo Alto, CA: Consulting Psychologists Press.
- Eckert, W. A. 2000. "Situational enhancement of design validity: The case of training evaluation at the World Bank Institute," *American Journal of Evaluation* (21:2), pp. 185–193.
- Elliot, A. J., and Devine, P. G. 1994. "On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort," *Journal of Personality and Social Psychology* (67:3), pp. 382–394.
- Ennis, R. 2011. "Critical Thinking," *Inquiry: Critical Thinking Across the Disciplines* (26:2), pp. 5–19.
- Exeter, D. J., Ameratunga, S., Ratima, M., Morton, S., Dickson, M., Hsu, D., and Jackson, R. 2010. "Student engagement in very large classes: The teachers' perspective," *Studies in Higher Education* (35:7), pp. 761–775.
- Falchikov, N. 2001. *Learning together: Peer tutoring in higher education*, London: Routledge.
- Falchikov, N. 2005. *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education*, New York: Routledge.
- Fang, B., Ye, Q., Kucukusta, D., and Law, R. 2016. "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics," *Tourism Management* (52), pp. 498–506.
- Farag, Y., Yannakoudakis, H., and Briscoe, T. 2018. "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input," *arXiv* . (<http://arxiv.org/pdf/1804.06898v2>, accessed December 30, 2019).
- Felder, R. M. 1997. "Beating the numbers game: effective teaching in large classes," in *ASEE Annual Conference*, , Milwaukee, WI.

## REFERENCES

---

- Feldman, R., and Sanger, J. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*, New York, NY, USA: Cambridge University Press.
- Festinger, L. 1954. "A Theory of Social Comparison Processes," *Human Relations* (7:2), pp. 117–140.
- Festinger, L. 1962. *A theory of cognitive dissonance*, Standford, CA: Stanford University Press.
- Filieri, R. 2016. "What makes an online consumer review trustworthy?" *Annals of Tourism Research* (58), pp. 46–64.
- Fong, C. J., Warner, J. R., Williams, K. M., Schallert, D. L., Chen, L.-H., Williamson, Z. H., and Lin, S. 2016. "Deconstructing constructive criticism: The nature of academic emotions associated with constructive, positive, and negative feedback," *Learning and Individual Differences* (49), pp. 393–399.
- Formanek, M., Wenger, M. C., Buxner, S. R., Impey, C. D., and Sonam, T. 2017. "Insights about large-scale online peer assessment from an analysis of an astronomy MOOC," *Computers & Education* (113), pp. 243–262.
- Fortes, P. C., and Tchantchane, A. 2010. "Dealing with Large Classes: A Real Challenge," *Procedia-Social and Behavioral Sciences* (8), pp. 272–280.
- Gao, B., Hu, N., and Bose, I. 2017. "Follow the herd or be myself? An analysis of consistency in behavior of reviewers and helpfulness of their reviews," *Decision Support Systems* (95), pp. 1–11.
- Gensim 2019. "Latent Dirichlet Allocation via Mallet," . (<https://radimrehurek.com/gensim/models/wrappers/ldamallet.html>, accessed December 30, 2019).
- Ghose, A., and Ipeirotis, P. G. 2011. "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE Transactions on Knowledge and Data Engineering* (23:10), pp. 1498–1512.
- Gibbs, G. 1992. *Teaching large classes in higher education: How to maintain quality with reduced resources*, Londong: Kogan Page.
- Gibbs, G., and Simpson, C. 2005. "Conditions under which assessment supports students' learning," *Learning and teaching in higher education* (:1), pp. 3–31.

- Ginzberg, M. J. 1981. "Early diagnosis of MIS implementation failure: promising results and unanswered questions," *Management Science* (27:4), pp. 459–478.
- Glaserfeld, E. v. 1989. "Constructivism in education," *The international encyclopedia of education* (1), pp. 162–163.
- González-Rodríguez, M. R., Martínez-Torres, R., and Toral, S. 2016. "Post-visit and pre-visit tourist destination image through eWOM sentiment analysis and perceived helpfulness," *International Journal of Contemporary Hospitality Management* (28:11), pp. 2609–2627.
- Graesser, A. C., and McNamara, D. S. 2012. "Automated analysis of essays and open-ended verbal responses," *APA Handbook of Research Methods in Psychology* (1), pp. 307–325.
- Gregor, S., and Hevner, A. R. 2013. "Positioning and presenting design science research for maximum impact," *MIS Quarterly* (37:2), pp. 337–355.
- Gregor, S., and Jones, D. 2007. "The Anatomy of a Design Theory," *Journal of the Association for Information Systems* (8:5), pp. 312–335.
- Gregory, R. W., and Muntermann, J. 2014. "Research Note—Heuristic Theorizing: Proactively Generating Design Theories," *Information Systems Research* (25:3), pp. 639–653.
- Gunning, R. 1952. *The technique of clear writing*, Toronto: McGraw-Hill.
- Guo, B., and Zhou, S. 2017. "What makes population perception of review helpfulness: an information processing perspective," *Electronic Commerce Research* (17:4), pp. 585–608.
- Gupta, S., and Bostrom, R. 2013. "An Investigation of the Appropriation of Technology-Mediated Training Methods Incorporating Enactive and Collaborative Learning," *Information Systems Research* (24:2), pp. 454–469.
- Gupta, S., Bostrom, R. P., and Huber, M. 2010. "End-user training methods," *ACM SIGMIS Database* (41:4), pp. 9–39.
- Hair, J. F., Hult, T. M., Ringle, C. M., and Sarstedt, M. 2014. *A primer on partial least squares structural equation modeling (PLS-SEM)*, Thousand Oaks, CA: Sage.

## REFERENCES

---

- Hardin, A. M., Looney, C. A., and Fuller, M. A. 2014. "Self-efficacy, learning method appropriation and software skills acquisition in learner-controlled CSSTS environments," *Information Systems Journal* (24:1), pp. 3–27.
- Hattie, J. 2008. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*, New York: Routledge, Taylor & Francis Group.
- Hattie, J. 2015. "The applicability of Visible Learning to higher education," *Scholarship of Teaching and Learning in Psychology* (1:1), pp. 79–91.
- Hattie, J., and Gan, M. 2011. "Instruction based on feedback," in *Handbook of research on learning and instruction*, , Educational psychology handbook series, New York: Routledge, pp. 249–271.
- Hattie, J., and Timperley, H. 2007. "The power of feedback," *Review of Educational Research* (77:1), pp. 81–112.
- Hattie, J., and Yates, G. C. R. 2013. *Visible learning and the science of how we learn*, Abingdon: Routledge.
- Hevner, A. 2007. "A Three Cycle View of Design Science Research," *Scandinavian Journal of Information Systems* (19:2), pp. 87–92.
- Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design science in information systems research," *MIS Quarterly* (28:1), pp. 75–105.
- Higgins, C. A., and Bligh, B. 2006. "Formative computer based assessment in diagram based domains," *ACM SIGCSE Bulletin* (38:3), p. 98.
- Highsmith, J., and Cockburn, A. 2001. "Agile software development: the business of innovation," *Computer* (34:9), pp. 120–127.
- Hofstede, G. 2013. "The VSM 2013 (Values Survey Module) for cross-cultural research is free for download in many languages," . (<https://geerthofstede.com/research-and-vsm/vsm-2013/>, accessed December 30, 2019).
- Hofstede, G., and Hofstede, G. J. 2005. *Cultures and organizations: Software of the mind*, New York: McGraw-Hill.
- Holmberg, B. 1986. *Growth and structure of distance education*, London: Croom Helm.

- Hong, K., Gu, B., Burtch, G., Huang, N., Liang, C., Wang, K., and Fu, D. 2016. "Effectiveness of performance feedback in stimulating user-generated content," in *International Conference on Information Systems (ICIS)*, , Dublin, Ireland.
- Hsia, L.-H., Huang, I., and Hwang, G.-J. 2016. "Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course," *Computers & Education* (96), pp. 55–71.
- Hu, Y.-H., and Chen, K. 2016. "Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings," *International Journal of Information Management* (36:6), pp. 929–944.
- Huang, A. H., Chen, K., Yen, D. C., and Tran, T. P. 2015. "A study of factors that contribute to online review helpfulness," *Computers in Human Behavior* (48), pp. 17–27.
- Huang, L., Tan, C.-H., Ke, W., and Wei, K.-K. 2013. "Comprehension and assessment of product reviews: A review-product congruity proposition," *Journal of Management Information Systems* (30:3), pp. 311–343.
- Huang, N., Hong, Y., and Burtch, G. 2017. "Social network integration and user content generation: Evidence from natural experiments," *MIS Quarterly* (41:4), pp. 1035–1058.
- Hutto, C. J., and Gilbert, E. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, , Ann Arbor, Michigan, USA.
- IBM SPSS 2018. "Transforming different Likert scales to a common scale," . (<http://www-01.ibm.com/support/docview.wss?uid=swg21482329>, accessed December 30, 2019).
- Ilgen, D. R., Fisher, C. D., and Taylor, M. S. 1979. "Consequences of individual feedback on behavior in organizations," *Journal of Applied Psychology* (64:4), pp. 349–371.
- Janson, A., Söllner, M., and Leimeister, J. M. 2019. "Ladders for Learning: Is Scaffolding the Key to Teaching Problem Solving in Technology-mediated

## REFERENCES

---

- Learning Contexts?” *Academy of Management Learning & Education* (online first) .
- Jarvis, C. B., MacKenzie, S. B., and Podsakoff, P. M. 2003. “A critical review of construct indicators and measurement model misspecification in marketing and consumer research,” *Journal of Consumer Research* (30:2), pp. 199–218.
- Jensen, M. L., Averbeck, J. M., Zhang, Z., and Wright, K. B. 2013. “Credibility of anonymous online product reviews: A language expectancy perspective,” *Journal of Management Information Systems* (30:1), pp. 293–324.
- Kim, K., Cheong, Y., and Kim, H. 2017. “User-generated product reviews on the internet: the drivers and outcomes of the perceived usefulness of product reviews,” *International Journal of Advertising* (36:2), pp. 227–245.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*, Millington: Defense Technical Information Center.
- Klašnja-Milićević, A., Vesin, B., Ivanović, M., and Budimac, Z. 2011. “E-Learning personalization based on hybrid recommendation strategy and learning style identification,” *Computers & Education* (56:3), pp. 885–899.
- Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. 2004. “Length, frequency, and predictability effects of words on eye movements in reading,” *European Journal of Cognitive Psychology* (16:1-2), pp. 262–284.
- Kluger, A. N., and DeNisi, A. 1996. “The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory,” *Psychological Bulletin* (119:2), pp. 254–284.
- Korfiatis, N., García-Bariocanal, E., and SáNchez-Alonso, S. 2012. “Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content,” *Electronic Commerce Research and Applications* (11:3), pp. 205–217.
- Krathwohl, D. R. 2002. “A revision of Bloom’s taxonomy: An overview,” *Theory into Practice* (41:4), pp. 212–218.

- Krishnamoorthy, S. 2015. "Linguistic features for review helpfulness prediction," *Expert Systems with Applications* (42:7), pp. 3751–3759.
- Kruger, J., and Dunning, D. 1999. "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments," *Journal of Personality and Social Psychology* (77:6), p. 1121.
- Kuan, K. K. Y., Hui, K.-L., Prasarnphanich, P., and Lai, H.-Y. 2015. "What makes a review voted? An empirical investigation of review voting in online review systems," *Journal of the Association for Information Systems* (16:1), pp. 48–71.
- Kulhavy, R. W., and Stock, W. A. 1989. "Feedback in written instruction: The place of response certitude," *Educational Psychology Review* (1:4), pp. 279–308.
- Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. 2013. "Peer and self assessment in massive online classes," *ACM Transactions on Computer-Human Interaction (TOCHI)* (20:6), pp. 1–31.
- Kusumasondjaja, S., Shanka, T., and Marchegiani, C. 2012. "Credibility of online reviews and initial trust: The roles of reviewer's identity and review valence," *Journal of Vacation Marketing* (18:3), pp. 185–195.
- Kwok, L., and Xie, K. L. 2016. "Factors contributing to the helpfulness of online hotel reviews: Does manager response play a role?" *International Journal of Contemporary Hospitality Management* (28:10), pp. 2156–2177.
- Lauterbach, D., Truong, H., Shah, T., and Adamic, L. 2009. "Surfing a web of trust: Reputation and reciprocity on couchsurfing.com," *IEEE 2009 – Computational science and engineering* (4), pp. 346–353.
- Lee, H. A., Law, R., and Murphy, J. 2011. "Helpful reviewers in TripAdvisor, an online travel community," *Journal of Travel & Tourism Marketing* (28:7), pp. 675–688.
- Lee, H. C. B., Ba, S., Li, X., and Stallaert, J. 2018. "Salience Bias in Crowdsourcing Contests," *Information Systems Research* (29:2), pp. 401–418.

## REFERENCES

---

- Lee, J. 2013. "What makes people read an online review? The relative effects of posting time and helpfulness on review readership," *Cyberpsychology, Behavior, and Social Networking* (16:7), pp. 529–535.
- Lee, S., and Choeh, J. Y. 2014. "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," *Expert Systems with Applications* (41:6), pp. 3041–3046.
- Lee, S., and Choeh, J. Y. 2017. "Exploring the determinants of and predicting the helpfulness of online user reviews using decision trees," *Management Decision* (55:4), pp. 681–700.
- Lee, S., and Choeh, J. Y. 2018. "The interactive impact of online word-of-mouth and review helpfulness on box office revenue," *Management Decision* (56:4), pp. 849–866.
- Lee, S. Y. 2014. "How do people compare themselves with others on social network sites?: The case of Facebook," *Computers in Human Behavior* (32), pp. 253–260.
- Lee, T. W., Mitchell, T. R., and Sablinski, C. J. 1999. "Qualitative research in organizational and vocational psychology, 1979–1999," *Journal of Vocational Behavior* (55:2), pp. 161–187.
- Leidner, D. E., and Kayworth, T. 2006. "A review of culture in information systems research: Toward a theory of information technology culture conflict," *MIS Quarterly* (30:2), pp. 357–399.
- Li, H., Zhang, Z., Meng, F., and Janakiraman, R. 2017. "Is peer evaluation of consumer online reviews socially embedded?—An examination combining reviewer's social network and social identity," *International Journal of Hospitality Management* (67), pp. 143–153.
- Li, L., Liu, X., and Steckelberg, A. L. 2010. "Assessor or assessee: How student learning improves by giving and receiving peer feedback," *British Journal of Educational Technology* (41:3), pp. 525–536.
- Li, M., Huang, L., Tan, C.-H., and Wei, K.-K. 2013. "Helpfulness of online product reviews as seen by consumers: Source and content features," *International Journal of Electronic Commerce* (17:4), pp. 101–136.



- Liu, B. 2012. "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies* (5:1), pp. 1–167.
- Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*, Cambridge: Cambridge University Press.
- Liu, N.-F., and Carless, D. 2006. "Peer feedback: the learning element of peer assessment," *Teaching in Higher Education* (11:3), pp. 279–290.
- Lo, A. W., Mamaysky, H., and Wang, J. 2000. "Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation," *The Journal of Finance* (55:4), pp. 1705–1765.
- Lohse, M., and Overgaard, M. 2019. "Emotional priming depends on the degree of conscious experience," *Neuropsychologia* (128), pp. 96–102.
- Luan, J., Yao, Z., Zhao, F., and Liu, H. 2016. "Search product and experience product online reviews: an eye-tracking study on consumers' review search behavior," *Computers in Human Behavior* (65), pp. 420–430.
- Lucas, H. 2014. "Disrupting and transforming the university," *Communications of the ACM* (57:10), pp. 32–35.
- Luo, C., Luo, X. R., Xu, Y., Warkentin, M., and Sia, C. L. 2015. "Examining the moderating role of sense of membership in online review evaluations," *Information & Management* (52:3), pp. 305–316.
- Malik, M. S., and Hussain, A. 2017. "Helpfulness of product reviews as a function of discrete positive and negative emotions," *Computers in Human Behavior* (73), pp. 290–302.
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. 2014. "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, , Baltimore, MD.
- Mao, Y., and Lebanon, G. 2007. "Isotonic Conditional Random Fields and Local Sentiment Flow," *Advances in Neural Information Processing Systems* (:19), pp. 961–968.
- March, S. T., and Smith, G. F. 1995. "Design and natural science research on information technology," *Decision Support Systems* (15:4), pp. 251–266.

## REFERENCES

---

- Marjanovic, O. 2016. "Designing Innovative Education through Action Design Research: Method and Application for Teaching Design Activities in Large Lecture Environments," *JITTA: Journal of Information Technology Theory and Application* (17:2), pp. 22–50.
- Markus, M. L., Majchrzak, A., and Gasser, L. 2002. "A design theory for systems that support emergent knowledge processes," *MIS Quarterly* (26:3), pp. 179–212.
- Marsden, J., Budden, D., Craig, H., and Moscato, P. 2013. "Language Individuation and Marker Words: Shakespeare and His Maxwell's Demon," *PloS one* (8:6), pp. 1–12.
- Mason, B. J., and Bruning, R. 2001. "Providing Feedback in Computer-based Instruction: What the Research Tells Us: CLASS Project Research Report No. 9," *Center for Instructional Innovation, University of Nebraska-Licolln* .
- McCallum, A. K. 2019. "MALLET: MACHine Learning for Language Toolkit," . (<http://mallet.cs.umass.edu/>, accessed July 2, 2019).
- McDonald, B., and Boud, D. 2003. "The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations," *Assessment in Education: Principles, Policy & Practice* (10:2), pp. 209–220.
- Meadows, A. J. 1998. *Communicating Research*, San Diego, CA: Academic Press.
- Memarzadeh, F., Blum, S. C., and Adams, C. 2016. "Business Travelers' Intention to Purchase: The Application of Technology Acceptance Model (TAM)," *Journal of Quality Assurance in Hospitality & Tourism* (17:4), pp. 412–424.
- Mertens, W., Pugliese, A., and Recker, J. 2017. *Quantitative data analysis: A companion for accounting and information systems research*, Switzerland: Springer.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. "Distributed Representations of Words and Phrases and their Compositionality," in *Preceedings Advances in Neural Information Processing Systems 26*, , Lake Tahoe, NV, USA.

- Millsap, R. E., and Maydeu-Olivares, A. 2009. *The SAGE handbook of quantitative methods in psychology*, London: Sage Publications.
- Min, H.-J., and Park, J. C. 2012. "Identifying helpful reviews based on customer's mentions about experiences," *Expert Systems with Applications* (39:15), pp. 11,830–11,838.
- Moore, M. G. 1989. "Editorial: Three types of interaction," *American Journal of Distance Education* (3:2), pp. 1–7.
- Moshavi, D., Brown, W. F., and Dodd, N. G. 2003. "Leader self-awareness and its relationship to subordinate attitudes and performance," *Leadership & Organization Development Journal* (24:7), pp. 407–418.
- Mruk, C. J. 2013. *Self-esteem and positive psychology: Research, theory, and practice*, New York: Springer Publishing Company.
- Mudambi, S. M., and Schuff, D. 2010. "Research note: What makes a helpful online review? A study of customer reviews on Amazon. com," *MIS Quarterly* (34:1), pp. 185–200.
- Mulryan-Kyne, C. 2010. "Teaching large classes at college and university level: Challenges and opportunities," *Teaching in Higher Education* (15:2), pp. 175–185.
- Mumford, E. 2006. "The story of socio-technical design: Reflections on its successes, failures and potential," *Information Systems Journal* (16:4), pp. 317–342.
- Myers, D. G. 2009. *Social psychology*, New York: McGraw-Hill, 10th ed.
- Nagy, W. E., Anderson, R. C., and Herman, P. A. 1987. "Learning Word Meanings From Context During Normal Reading," *American Educational Research Journal* (24:2), pp. 237–270.
- Neat, R. 2015. "Volkswagen under investigation over illegal software that masks emissions," *The Guardian* (September 18, 2015). (<https://www.theguardian.com/business/2015/sep/18/epa-california-investigate-volkswagen-clean-air-violations>, accessed December 30, 2019).

## REFERENCES

---

- Ngo-Ye, T. L., and Sinha, A. P. 2014. "The influence of reviewer engagement characteristics on online review helpfulness: A text regression model," *Decision Support Systems* (61), pp. 47–58.
- Nicol, D. 2010. "From monologue to dialogue: improving written feedback processes in mass higher education," *Assessment & Evaluation in Higher Education* (35:5), pp. 501–517.
- Nicol, D., Thomson, A., and Breslin, C. 2014. "Rethinking feedback practices in higher education: a peer review perspective," *Assessment & Evaluation in Higher Education* (39:1), pp. 102–122.
- Nicol, D. J., and Boyle, J. T. 2003. "Peer instruction versus class-wide discussion in large classes: A comparison of two interaction methods in the wired classroom," *Studies in Higher Education* (28:4), pp. 457–473.
- Nicol, D. J., and Macfarlane-Dick, D. 2006. "Formative assessment and self-regulated learning: a model and seven principles of good feedback practice," *Studies in Higher Education* (31:2), pp. 199–218.
- Niehaves, B., and Ortbach, K. 2016. "The inner and the outer model in explanatory design theory: the case of designing electronic feedback systems," *European Journal of Information Systems* (25:4), pp. 303–316.
- Nonaka, I., and Takeuchi, H. 1995. *The knowledge-creating company: How Japanese companies create the dynamics of innovation*, New York: Oxford University Press.
- OECD 2016. *Education at a Glance 2016: OECD Indicators*, Paris: OECD.
- OECD 2019. *OECD skills outlook 2019: Thriving in a digital world*, Paris: OECD.
- Otterbacher, J. 2013. "Gender, writing and ranking in review forums: a case study of the IMDb," *Knowledge and information systems* (35:3), pp. 645–664.
- Pan, Y., and Zhang, J. Q. 2011. "Born unequal: a study of the helpfulness of user-generated product reviews," *Journal of Retailing* (87:4), pp. 598–612.
- Pang, B., and Lee, L. 2008. "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval* (2:1–2), pp. 1–135.

- Park, S., and Nicolau, J. L. 2015. "Asymmetric effects of online consumer reviews," *Annals of Tourism Research* (50), pp. 67–83.
- Patton, C. 2012. "'Some kind of weird, evil experiment': student perceptions of peer assessment," *Assessment & Evaluation in Higher Education* (37:6), pp. 719–731.
- Pavlou, P. A., and Gefen, D. 2004. "Building effective online marketplaces with institution-based trust," *Information Systems Research* (15:1), pp. 37–59.
- Peppers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A design science research methodology for information systems research," *Journal of Management Information Systems* (24:3), pp. 45–77.
- Pennebaker, J. W., and King, L. A. 1999. "Linguistic styles: language use as an individual difference," *Journal of Personality and Social Psychology* (77:6), pp. 1296–1312.
- Pentina, I., Bailey, A. A., and Zhang, L. 2018. "Exploring effects of source similarity, message valence, and receiver regulatory focus on yelp review persuasiveness and purchase intentions," *Journal of Marketing Communications* (24:2), pp. 125–145.
- Piaget, J. 1985. *The equilibration of cognitive structures : the central problem of intellectual development*, Chicago: University of Chicago press.
- Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. 2013. "Tuned models of peer assessment in MOOCs," *arXiv* . (<https://arxiv.org/abs/1307.2579>, accessed December 30, 2019).
- Pintrich, P. R. 2000. "The role of goal orientation in self-regulated learning," in *Handbook of self-regulation*, , San Diego, CA: Elsevier, pp. 451–502.
- Pitt, E., and Norton, L. 2017. "'Now that's the feedback I want!' Students' reactions to feedback on graded work and what they do with it," *Assessment & Evaluation in Higher Education* (42:4), pp. 499–516.
- Poplin, M. S. 1988. "Holistic/constructivist Principles of the Teaching/Learning Process Implications for the Field of Learning Disabilities," *Journal of Learning Disabilities* (21:7), pp. 401–416.

## REFERENCES

---

- Price, M., Handley, K., Millar, J., and O'Donovan, B. 2010. "Feedback: All that effort, but what is the effect?" *Assessment & Evaluation in Higher Education* (35:3), pp. 277–289.
- Purnawirawan, N., de Pelsmacker, P., and Dens, N. 2012. "Balance and sequence in online reviews: How perceived usefulness affects attitudes and intentions," *Journal of Interactive Marketing* (26:4), pp. 244–255.
- Qazi, A., Syed, K. B. S., Raj, R. G., Cambria, E., Tahir, M., and Alghazzawi, D. 2016. "A concept-level approach to the analysis of online review helpfulness," *Computers in Human Behavior* (58), pp. 75–81.
- Quaschnig, S., Pandelaere, M., and Vermeir, I. 2015. "When consistency matters: The effect of valence consistency on review helpfulness," *Journal of Computer-Mediated Communication* (20:2), pp. 136–152.
- Racherla, P., and Friske, W. 2012. "Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories," *Electronic Commerce Research and Applications* (11:6), pp. 548–559.
- Racherla, P., Mandviwalla, M., and Connolly, D. J. 2012. "Factors affecting consumers' trust in online product reviews," *Journal of Consumer Behaviour* (11:2), pp. 94–104.
- Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. 2016. "The emotional arcs of stories are dominated by six basic shapes," *EPJ Data Science* (5:1), pp. 1–12.
- Reinecke, K., and Bernstein, A. 2013. "Knowing What a User Likes: A Design Science Approach to Interfaces that Automatically Adapt to Culture," *MIS Quarterly* (37:2), pp. 427–453.
- Remus, R., Quasthoff, U., and Heyer, G. 2010. "SentiWS-A Publicly Available German-language Resource for Sentiment Analysis," in *Proceedings of the International Language Resources and Evaluation (LREC)*, , Valletta, Malta.
- Renaud, S., Tannenbaum, E., and Stantial, P. 2007. "Student-Centered Teaching in Large Classes with Limited Resources," in *English Teaching Forum*, , Washington, DC.

- Resnick, L. B. 1987. *Education and learning to think*, Washington, DC: National Academies.
- Reuters, T. 2017. “Journal Citation Reports,” . (<http://help.incites.clarivate.com/incitesLiveJCR/overviewGroup/overviewJCR.html>, accessed December 30, 2019).
- Rietsche, R., Bauman, K., Soellner, M., and Leimeister, J. M. 2018. “What Makes Great Feedback? Understanding and Measuring Technology-Mediated Peer Feedback Quality,” *Information Systems Research* .
- Rietsche, R., Duss, K., Persch, J. M., and Soellner, M. 2018. “Design and Evaluation of an IT-based Formative Feedback Tool to Foster Student Performance,” in *Proceedings of the International Conference on Information Systems (ICIS) 2018*, , San Francisco, CA, USA.
- Rietsche, R., Frei, D., Stoeckli, E., and Soellner, M. 2019. “Not all Reviews are Equal - a Literature Review on Online Review Helpfulness,” in *Proceedings of the European Conference on Information Systems (ECIS)*, , Stockholm, Sweden.
- Rietsche, R., Lehmann, K., Haas, P., and Söllner, M. 2017a. “The Twofold Value of IT-Based Peer Assessment in Management Information Systems Education,” in *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*, , St Gallen, Switzerland.
- Rietsche, R., and Söllner, M. 2019. “Insights into Using IT-Based Peer Feedback to Practice the Students Providing Feedback Skill,” in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, , Maui, HI, USA.
- Rietsche, R., Söllner, M., and Seufert, S. 2017b. “Digital formative learning assessment tool - Towards helping students to take ownership of their learning,” in *Proceedings of the European Conference on Information Systems (ECIS)*, , Guimarães, Portugal.
- Ringle, C. M., Wende, S., and Becker, J.-M. 2015. “SmartPLS 3,” .
- Robinson, R., Goh, T.-T., and Zhang, R. 2012. “Textual factors in online product reviews: a foundation for a more influential approach to opinion mining,” *Electronic Commerce Research* (12:3), pp. 301–330.

## REFERENCES

---

- Roscoe, R. D., and Chi, M. T. H. 2008. "Tutor learning: The role of explaining and responding to questions," *Instructional science* (36:4), pp. 321–350.
- Rosemann, M., and Vessey, I. 2008. "Toward improving the relevance of information systems research to practice: the role of applicability checks," *MIS Quarterly* (32:1), pp. 1–22.
- Ryang, H., Yun, U., Pyun, G., Lee, G., and Kim, J. 2015. "Ranking algorithm for book reviews with user tendency and collective intelligence," *Multimedia Tools and Applications* (74:16), pp. 6209–6227.
- Sadler, D. R. 1989. "Formative assessment and the design of instructional systems," *Instructional science* (18:2), pp. 119–144.
- Sadler, P. M., and Good, E. 2006. "The impact of self-and peer-grading on student learning," *Educational assessment* (11:1), pp. 1–31.
- Salehan, M., and Kim, D. J. 2016. "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics," *Decision Support Systems* (81), pp. 30–40.
- Scalise, K., and Gifford, B. 2006. "Computer-based assessment in e-learning: A framework for constructing" intermediate constraint" questions and tasks for technology platforms," *The Journal of Technology, Learning and Assessment* (4:6), pp. 1–45.
- Schiefele, U. 1991. "Interest, Learning, and Motivation," *Educational Psychologist* (26:3-4), pp. 299–323.
- Schindler, R. M., and Bickart, B. 2012. "Perceived helpfulness of online consumer reviews: The role of message content and style," *Journal of Consumer Behaviour* (11:3), pp. 234–243.
- Schlosser, A. E. 2011. "Can including pros and cons increase the helpfulness and persuasiveness of online reviews? The interactive effects of ratings and arguments," *Journal of Consumer Psychology* (21:3), pp. 226–239.
- Schumacher, R. 2017. *Schreiben in den Literaturwissenschaften [Writing in literary studies]: Ein Arbeitsbuch [A workbook]*, vol. 4749 of *UTB Schlüsselkompetenzen Literaturwissenschaft*, Paderborn: Wilhelm Fink.



- Scopus 2017. "CiteScore metrics for journals and serials," . (<https://www.scopus.com>, accessed December 30, 2019).
- Searle, J. R., and Searle, J. R. 1969. *Speech acts: An essay in the philosophy of language*, Cambridge: Cambridge University Press.
- Seel, N. M. 2011. *Encyclopedia of the Sciences of Learning*, New York, NY: Springer Science & Business Media.
- Shan, Y. 2016. "How credible are online product reviews? The effects of self-generated and system-generated cues on source credibility evaluation," *Computers in Human Behavior* (55), pp. 633–641.
- Sharma, R., Yetton, P., and Crawford, J. 2009. "Estimating the Effect of Common Method Variance: The Method—Method Pair Technique with an Illustration from TAM Research," *MIS Quarterly* (33:3), p. 473.
- Shen, X.-L., Zhang, K. Z. K., and Zhao, S. J. 2016. "Herd behavior in consumers' adoption of online reviews," *Journal of the Association for Information Science and Technology* (67:11), pp. 2754–2765.
- Shute, V. J. 2008. "Focus on formative feedback," *Review of Educational Research* (78:1), pp. 153–189.
- Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., and Roy, P. K. 2017. "Predicting the "helpfulness" of online consumer reviews," *Journal of Business Research* (70), pp. 346–355.
- Smith, H. J., Milberg, S. J., and Burke, S. J. 1996. "Information Privacy: Measuring Individuals' Concerns about Organizational Practices," *MIS Quarterly* (20:2), p. 167.
- Söllner, M., Bitzer, P., Janson, A., and Leimeister, J. M. 2017. "Process is king: Evaluating the performance of technology-mediated learning in vocational software training," *Journal of Information Technology* (33:3), pp. 233–253.
- Spacy 2019. "Rule-based matching · spaCy Usage Documentation," . (<https://spacy.io/usage/rule-based-matching>, accessed 27.05.2019).
- Spagnolli, A., and Gamberini, L. 2007. "Interacting via SMS: Practices of social closeness and reciprocation," *British Journal of Social Psychology* (46:2), pp. 343–364.

## REFERENCES

---

- Spärck Jones, K. 1972. "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation* (28:1), pp. 11–21.
- Sparks-Langer, G. M., Simmons, J. M., Pasch, M., Colton, A., and Starko, A. 1990. "Reflective pedagogical thinking: How can we promote it and measure it?" *Journal of Teacher Education* (41:5), pp. 23–32.
- Srinivasan, A. 1985. "Alternative measures of system effectiveness: associations and implications," *MIS Quarterly* (9:3), pp. 243–253.
- Statista 2018a. "Cumulative number of reviews submitted to Yelp from 2009 to 2017 (in millions)," . (<https://www.statista.com/statistics/278032/cumulative-number-of-reviews-submitted-to-yelp/>, accessed December 30, 2019).
- Statista 2018b. "Number of user reviews and opinions on TripAdvisor worldwide from 2014 to 2017 (in millions)," . (<https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/>, accessed December 30, 2019).
- Stein, S. 2003. *Textgliederung [Text structure]: Einheitenbildung im geschriebenen und gesprochenen Deutsch: Theorie und Empirie [Unit formation in written and spoken German: theory and empiricism]*, vol. 69 of *Studia Linguistica Germanica*, Berlin: De Gruyter.
- Stollenwerk, T. 2013. "German-stopwords based on apache," . (<https://github.com/collective/collective.solr/blob/master/etc/german-stopwords.txt>, accessed December 30, 2019).
- Straub, D., Limayem, M., and Karahanna-Evaristo, E. 1995. "Measuring system usage: Implications for IS theory testing," *Management Science* (41:8), pp. 1328–1342.
- Strijbos, J.-W., Narciss, S., and Dünnebier, K. 2010. "Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency?" *Learning and Instruction* (20:4), pp. 291–303.
- Suen, H. K. 2014. "Peer assessment for massive open online courses (MOOCs)," *The International Review of Research in Open and Distributed Learning* (15:3), pp. 312–327.

- Tahir, I. H. 2012. "A study on peer evaluation and its influence on college ESL students," *Procedia-Social and Behavioral Sciences* (68), pp. 192–201.
- Tai, J. H.-M., Canny, B. J., Haines, T. P., and Molloy, E. K. 2016. "The role of peer-assisted learning in building evaluative judgement: opportunities in clinical medical education," *Advances in Health Sciences Education* (21:3), pp. 659–676.
- Tausczik, Y. R., and Pennebaker, J. W. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology* (29:1), pp. 24–54.
- Teng, S., Wei Khong, K., Wei Goh, W., and Yee Loong Chong, A. 2014. "Examining the antecedents of persuasive eWOM messages in social media," *Online Information Review* (38:6), pp. 746–768.
- Thelwall, M. 2000. "Computer-based assessment: a versatile educational tool," *Computers & Education* (34:1), pp. 37–49.
- Topping, K. 1998. "Peer Assessment Between Students in Colleges and Universities," *Review of Educational Research* (68:3), pp. 249–276.
- Trilling, B., and Fadel, C. 2009. *21st century skills: Learning for life in our times*, San Francisco, CA: John Wiley & Sons.
- UNESCO Institute for Statistics 2019. "Number of students in tertiary education and public spending 2014-2017," . (<http://data.uis.unesco.org/Index.aspx>, accessed December 30, 2019).
- Universität St.Gallen 2019. "Immatrikulationsstatistik 2018," . (<https://www.unisg.ch/de/universitaet/ueber-uns/portraet/hsginzahlen/lehren>, accessed December 30, 2019).
- van der Kleij, F. M., Eggen, T. J., Timmers, C. F., and Veldkamp, B. P. 2012. "Effects of feedback in a computer-based assessment for learning," *Computers & Education* (58:1), pp. 263–272.
- van Popta, E., Kral, M., Camp, G., Martens, R. L., and Simons, P. R.-J. 2017. "Exploring the value of peer feedback in online learning for the provider," *Educational Research Review* (20), pp. 24–34.

## REFERENCES

---

- van Zundert, M., Sluijsmans, D., and van Merriënboer, J. 2010. "Effective peer assessment processes: Research findings and future directions," *Unravelling Peer Assessment* (20:4), pp. 270–279.
- Venable, J., Pries-Heje, J., and Baskerville, R. 2016. "FEDS: a framework for evaluation in design science research," *European Journal of Information Systems* (25:1), pp. 77–89.
- vom Brocke, J., Riedl, R., and Léger, P.-M. 2013. "Application strategies for neuroscience in information systems design science research," *Journal of Computer Information Systems* (53:3), pp. 1–13.
- vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., and Cleven, A. 2015. "Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research," *Communications of the Association for Information Systems* (37:9), pp. 205–224.
- Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an information system design theory for vigilant EIS," *Information Systems Research* (3:1), pp. 36–59.
- Wand, M. P., and Jones, M. C. 1994. *Kernel smoothing*, London: Chapman and Hall.
- Wang, W., Zhao, Y., Qiu, L., and Zhu, Y. 2014. "Effects of Emoticons on the Acceptance of Negative Feedback in Computer-Mediated Communication," *Journal of the Association for Information Systems* (15:8), pp. 454–483.
- Weathers, D., Swain, S. D., and Grover, V. 2015. "Can online product reviews be more helpful? Examining characteristics of information content by product type," *Decision Support Systems* (79), pp. 12–23.
- Weaver, M. R. 2006. "Do students value feedback? Student perceptions of tutors' written responses," *Assessment & Evaluation in Higher Education* (31:3), pp. 379–394.
- Webster, J., and Watson, R. T. 2002. "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quarterly* (26:2), pp. xiii–xxiii.
- Weiß, Z., and Meurers, D. 2018. "Modeling the Readability of German Targeting Adults and Children: An empirically broad analysis and its cross-corpus

- validation,” in *Proceedings of the International Conference on Computational Linguistics*, , Santa Fe, NM.
- Weitzel, S. R. 2000. *Feedback that works: How to build and deliver your message*, Ideas into action guidebooks, Greensboro, NC: Center for Creative Leadership.
- Whitworth, E., and Biddle, R. 2007. “The Social Nature of Agile Teams,” in *Agile 2007*, J. Eckstein (ed.), Los Alamitos, CA: IEEE Computer Soc.
- Wikimedia 2019. “Wikimedia Downloads: dewiki-20190401-pages-articles-multistream.xml.bz2,” . (<https://dumps.wikimedia.org/dewiki/>, accessed December 30, 2019).
- Willemsen, L. M., Neijens, P. C., Bronner, F., and de Ridder, J. A. 2011. “Highly recommended! The content characteristics and perceived usefulness of online consumer reviews,” *Journal of Computer-Mediated Communication* (17:1), pp. 19–38.
- Williams, L., and Cockburn, A. 2003. “Agile software development: it’s about feedback and change,” *IEEE Computer* (36:6), pp. 39–43.
- Wills, T. A. 1981. “Downward comparison principles in social psychology,” *Psychological Bulletin* (90:2), pp. 245–271.
- Winstone, N. E., Nash, R. A., Parker, M., and Rowntree, J. 2017. “Supporting learners’ agentic engagement with feedback: A systematic review and a taxonomy of recipience processes,” *Educational Psychologist* (52:1), pp. 17–37.
- Wirth, R., and Hipp, J. 2000. “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, , Manchester, UK.
- Wu, P. F. 2013. “In search of negativity bias: An empirical study of perceived helpfulness of online reviews,” *Psychology & Marketing* (30:11), pp. 971–984.
- Xu, Q. 2014. “Should I trust him? The effects of reviewer profile characteristics on eWOM credibility,” *Computers in Human Behavior* (33), pp. 136–144.

## REFERENCES

---

- Yang, M., and Carless, D. 2013. "The feedback triangle and the enhancement of dialogic feedback processes," *Teaching in Higher Education* (18:3), pp. 285–297.
- Yang, S.-B., Hlee, S., Lee, J., and Koo, C. 2017a. "An empirical examination of online restaurant reviews on Yelp. com: A dual coding theory perspective," *International Journal of Contemporary Hospitality Management* (29:2), pp. 817–839.
- Yang, S.-B., Shin, S.-H., Joun, Y., and Koo, C. 2017b. "Exploring the comparative importance of online hotel reviews' heuristic attributes in review helpfulness: a conjoint analysis approach," *Journal of Travel & Tourism Marketing* (34:7), pp. 963–985.
- Yang, Y.-T. C. 2008. "A catalyst for teaching critical thinking in a large university class in Taiwan: Asynchronous online discussions with the facilitation of teaching assistants," *Educational Technology Research and Development* (56:3), pp. 241–264.
- Yin, D., Bond, S., and Zhang, H. 2013. "Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews," *MIS Quarterly* (38:2), pp. 539–560.
- Yin, D., Mitra, S., and Zhang, H. 2016. "Research note—When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth," *Information Systems Research* (27:1), pp. 131–144.
- Yorke, M. 2003. "Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice," *Higher Education* (45:4), pp. 477–501.
- Zhao, X., Wang, L., Guo, X., and Law, R. 2015. "The influence of online reviews to online hotel booking intentions," *International Journal of Contemporary Hospitality Management* (27:6), pp. 1343–1364.
- Zheng, X., Zhu, S., and Lin, Z. 2013. "Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach," *Decision Support Systems* (56), pp. 211–222.

- Zhou, S., and Guo, B. 2017. "The order effect on online review helpfulness: A social influence perspective," *Decision Support Systems* (93), pp. 77–87.
- Zhu, L., Yin, G., and He, W. 2014. "Is this opinion leader's review useful? Peripheral cues for online review helpfulness," *Journal of Electronic Commerce Research* (15:4), pp. 267–280.
- Ziman, J. M. 1974. *Public knowledge: An essay concerning the social dimension of science*, London: Cambridge University Press.





# **Appendices**



## Appendix A: All Tables of the OLS Regression

**Table A1:** OLS Regression with Student-Level Fixed Effects for Subcategory “1.1 General Text Structure” from *I. Text Features* Category (\*\* $p$  – value  $\leq 0.01$ , \*\*\*  $p$  – value  $\leq 0.001$ ).

| <i>Variable</i>                              | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> | <i>M 4</i> |
|----------------------------------------------|------------|------------|------------|------------|
| 1.1.1.1 Paragraph tag                        | 0.031***   |            |            | 0.002      |
| 1.1.1.2 Line break tag                       | 0.014**    |            |            | –0.005     |
| 1.1.2.1 Short (up to 3 words)                |            | 0.003      |            | –0.006     |
| 1.1.2.2 Medium (1 sentence)                  |            | 0.045***   |            | 0.046***   |
| 1.1.2.3 Long (more than one sentence)        |            | 0.093***   |            | 0.104***   |
| 1.1.2.4 Number sentences in Long             |            | 0.018***   |            | 0.016***   |
| 1.1.2.5 Average number sentences in Long     |            | 0.057***   |            | 0.053***   |
| 1.1.3.1 Bullet point list tag                |            |            | –0.124*    | 0.028      |
| 1.1.3.2 Bullet point list e.g. hyphen symbol |            |            | 0.118***   | –0.025     |
| 1.1.3.3 Entries per bullet point list        |            |            | 0.041***   | 0.017*     |
| $R^2$                                        | 0.036      | 0.122      | 0.010      | 0.125      |

**Table A2:** OLS Regression with Student-Level Fixed Effects for Subcategory “1.2 Readability” from *1. Text Features* Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>                    | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> | <i>M 4</i> | <i>M 5</i> |
|------------------------------------|------------|------------|------------|------------|------------|
| 1.2.1.1 Average word length        | 0.079      |            |            |            | 0.024      |
| 1.2.1.2 Average sentence length    | 0.023***   |            |            |            | 0.030***   |
| 1.2.2.1 Flesch-Reading-Ease German |            | −0.006**   |            |            | 0.001      |
| 1.2.2.2 Gunning Fog                |            |            | 0.021***   |            | 0.001      |
| 1.2.3.1 Noun chunks                |            |            |            | 1.429*     | 2.059***   |
| 1.2.3.2 Nouns and proper nouns     |            |            |            | 0.884*     | 1.284**    |
| $R^2$                              | 0.006      | 0.002      | 0.002      | 0.004      | 0.014      |

**Table A3:** OLS Regression with Student-Level Fixed Effects for Subcategory “1.3 Writing Style” from *1. Text Features* Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>    | <i>M 1</i> |
|--------------------|------------|
| 1.3.1.1 Stop words | 1.227***   |
| $R^2$              | 0.003      |

**Table A4:** OLS Regression with Student-Level Fixed Effects for Subcategory “2.1 Argumentation” from 2. *Feedback Features* Category (\*\* $p - value \leq 0.01$ , \*\*\* $p - value \leq 0.001$ ).

| <i>Variable</i>                     | <i>M I</i> |
|-------------------------------------|------------|
| 2.1.1.1 Phrases indicating examples | 10.907***  |
| 2.1.1.2 Topic “examples“            | 4.313***   |
| $R^2$                               | 0.005      |

**Table A5:** OLS Regression with Student-Level Fixed Effects for Subcategory “2.2 Feedback Points” from 2. Feedback Features Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>                                                      | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> | <i>M 4</i> | <i>M 5</i> | <i>M 6</i> | <i>M 7</i> | <i>M 8</i> |
|----------------------------------------------------------------------|------------|------------|------------|------------|------------|------------|------------|------------|
| 2.2.1.1 Similarity submission and feedback                           | 2.034***   |            |            |            |            |            |            | 0.366**    |
| 2.2.1.2 Similarity top 5% topics submission and feedback             | 0.223***   |            |            |            |            |            |            | 0.190***   |
| 2.2.2.1 <i>Positive</i> Sentiment                                    |            | 0.229***   |            |            |            |            |            | 0.088***   |
| 2.2.2.2 <i>Negative</i> Sentiment                                    |            | −0.129***  |            |            |            |            |            | −0.008     |
| 2.2.2.3 Ratio between <i>positive</i> and <i>negative</i> words      |            | −15.979*** |            |            |            |            |            | −6.089***  |
| 2.2.3.1 Total number of sentences                                    |            |            | 0.034***   |            |            |            |            | 0.021***   |
| 2.2.3.2 Ratio feedback length and submission length in sentence      |            |            | 0.006      |            |            |            |            | 0.052      |
| 2.2.4.1 Important words in feedback, which are not in the submission |            |            |            | −0.115***  |            |            |            | −0.074***  |
| 2.2.4.2 New topics in feedback, which are not in submission          |            |            |            | 0.161***   |            |            |            | 0.080***   |
| 2.2.4.3 Sentence containing suggestion phrases                       |            |            |            | 0.436***   |            |            |            | 0.517      |
| 2.2.5.1 Sentence suggestion and exclusive <i>I</i>                   |            |            |            |            | 0.278      |            |            | 0.225      |
| 2.2.5.2 Sentence suggestion and exclusive <i>You</i>                 |            |            |            |            | 0.709**    |            |            | 0.281      |
| 2.2.5.3 Sentence suggestion and <i>I &amp; You</i>                   |            |            |            |            | 0.711      |            |            | 0.480      |
| 2.2.6.1 Sentence suggestion and is <i>short</i>                      |            |            |            |            |            | 0.173      |            | 0.131      |
| 2.2.6.2 Sentence suggestion and is <i>long</i>                       |            |            |            |            |            | 0.669***   |            | 0.361      |
| 2.2.7.1 Sentence suggestion and <i>positive</i> Sentiment            |            |            |            |            |            |            | 0.120      | −0.371     |
| 2.2.7.2 Sentence suggestion and <i>negative</i> Sentiment            |            |            |            |            |            |            | 0.464      | 0.214      |
| $R^2$                                                                | 0.054      | 0.040      | 0.101      | 0.091      | 0.003      | 0.002      | 0.001      | 0.153      |

**Table A6:** OLS Regression with Student-Level Fixed Effects for Subcategory “2.3 Flow of Feedback Points” from 2. *Feedback Features* Category (\*\* $p - value \leq 0.01$ , \*\*\* $p - value \leq 0.001$ ).

| <i>Variable</i>                                                      | <i>M 1</i> |
|----------------------------------------------------------------------|------------|
| 2.3.1.1 Feedback organized along positive, negative and improvements | 0.201***   |
| $R^2$                                                                | 0.007      |

**Table A7:** OLS Regression with Student-Level Fixed Effects for Subcategory “2.4 Verification and Elaboration” group of features from 2. *Feedback Features* Category (\*\* $p - value \leq 0.01$ , \*\*\* $p - value \leq 0.001$ ).

| <i>Variable</i>                                            | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> | <i>M 4</i> | <i>M 5</i> |
|------------------------------------------------------------|------------|------------|------------|------------|------------|
| 2.4.1.1 Sentence containing verification (positive)        | −1.222     |            |            |            | −1.091     |
| 2.4.1.2 Sentence containing verification (negative)        | −2.194     |            |            |            | −2.035     |
| 2.4.1.3 Sentence containing elaboration phrases            | 0.261**    |            |            |            | 0.470      |
| 2.4.2.1 Sentence elaboration and exclusive <i>I</i>        |            | 0.079      |            |            | −0.280     |
| 2.4.2.2 Sentence elaboration and exclusive <i>You</i>      |            | 0.418*     |            |            | 0.061      |
| 2.4.2.3 Sentence elaboration and <i>I &amp; You</i>        |            | 0.935*     |            |            | 0.652      |
| 2.4.3.1 Sentence elaboration and is <i>short</i>           |            |            | −0.420     |            | −0.578     |
| 2.4.3.2 Sentence elaboration and is <i>long</i>            |            |            | 0.603***   |            | 0.782***   |
| 2.4.4.1 Sentence elaboration and <i>positive</i> sentiment |            |            |            | 0.197      | −0.493*    |
| 2.4.4.2 Sentence elaboration and <i>negative</i> sentiment |            |            |            | −0.102     | −0.468*    |
| $R^2$                                                      | 0.003      | 0.003      | 0.005      | 0.000      | 0.010      |



**Table A8:** OLS Regression with Student-Level Fixed Effects for Subcategory “2.5 Function” group of features from 2. *Feedback Features* Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>                                             | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> | <i>M 4</i> | <i>M 5</i> | <i>M 6</i> | <i>M 7</i> | <i>M 8</i> | <i>M 9</i> | <i>M 10</i> | <i>M 11</i> | <i>M 12</i> |
|-------------------------------------------------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|-------------|-------------|-------------|
| 2.5.1.1 Phrases liking the submission                       | 0.098**    |            |            |            |            |            |            |            |            |             |             | 0.073*      |
| 2.5.2.1 Sentence containing directive phrases               |            | 0.603***   |            |            |            |            |            |            |            |             |             | 0.552       |
| 2.5.3.1 Sentence directive and exclusive <i>I</i>           |            |            | 0.482*     |            |            |            |            |            |            |             |             | −0.537      |
| 2.5.3.2 Sentence directive and exclusive <i>You</i>         |            |            | 0.636      |            |            |            |            |            |            |             |             | −0.517      |
| 2.5.3.3 Sentence directive and <i>I &amp; You</i>           |            |            | 1.659***   |            |            |            |            |            |            |             |             | 0.831       |
| 2.5.4.1 Sentence directive and is <i>short</i>              |            |            |            | 0.313      |            |            |            |            |            |             |             | −0.080      |
| 2.5.4.2 Sentence directive and is <i>long</i>               |            |            |            | 1.159***   |            |            |            |            |            |             |             | 0.715       |
| 2.5.5.1 Sentence directive and <i>positive</i> sentiment    |            |            |            |            | 0.579*     |            |            |            |            |             |             | −0.026      |
| 2.5.5.2 Sentence directive and <i>negative</i> sentiment    |            |            |            |            | 0.342      |            |            |            |            |             |             | 0.008       |
| 2.5.6.1 Sentences containing facilitative phrases           |            |            |            |            |            | 0.485***   |            |            |            |             |             | 0.036       |
| 2.5.7.1 Sentence facilitative and exclusive <i>I</i>        |            |            |            |            |            |            | 0.533***   |            |            |             |             | 0.633*      |
| 2.5.7.2 Sentence facilitative and exclusive <i>You</i>      |            |            |            |            |            |            | 0.604**    |            |            |             |             | 0.592       |
| 2.5.7.3 Sentence facilitative and <i>I &amp; You</i>        |            |            |            |            |            |            | 1.047**    |            |            |             |             | 0.835       |
| 2.5.8.1 Sentence facilitative and is <i>short</i>           |            |            |            |            |            |            |            | 0.266      |            |             |             | 0.082       |
| 2.5.8.2 Sentence facilitative and is <i>long</i>            |            |            |            |            |            |            |            | 0.795***   |            |             |             | 0.578*      |
| 2.5.9.1 Sentence facilitative and <i>positive</i> sentiment |            |            |            |            |            |            |            |            | 0.366*     |             |             | −0.462      |
| 2.5.9.2 Sentence facilitative and <i>negative</i> sentiment |            |            |            |            |            |            |            |            | 0.022      |             |             | −0.378      |
| 2.5.10.1 At least one question mark                         |            |            |            |            |            |            |            |            |            | 0.224***    |             | 0.473***    |
| 2.5.10.2 Question mark                                      |            |            |            |            |            |            |            |            |            |             | −0.145      | −1.729***   |
| $R^2$                                                       | 0.002      | 0.003      | 0.004      | 0.004      | 0.002      | 0.003      | 0.005      | 0.004      | 0.001      | 0.010       | 0.000       | 0.035       |

**Table A9:** OLS Regression with Student-Level Fixed Effects for Subcategory “3.1 Expressions” from 3. *Emotional Features* Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>                     | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> | <i>M 4</i> |
|-------------------------------------|------------|------------|------------|------------|
| 3.1.1.1 Salutation at the beginning | 0.194***   |            |            | 0.169**    |
| 3.1.2.1 Signing-Off at the end      |            | 0.117*     |            | 0.064      |
| 3.1.3.1 Emoticons                   |            |            | 0.179*     | 0.166*     |
| $R^2$                               | 0.002      | 0.001      | 0.001      | 0.003      |

**Table A10:** OLS Regression with Student-Level Fixed Effects for Subcategory “3.2 Nothing to Improve” from 3. *Emotional Features* Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>                                 | <i>M 1</i> |
|-------------------------------------------------|------------|
| 3.2.1.1 Submission does not need to be improved | −0.233***  |
| $R^2$                                           | 0.002      |

**Table A11:** OLS Regression with Student-Level Fixed Effects for Subcategory “3.3 Sentiment Curve” from 3. *Emotional Features* Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>                                | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> | <i>M 4</i> | <i>M 5</i> |
|------------------------------------------------|------------|------------|------------|------------|------------|
| 3.3.1.1 Number of segments                     | 0.121***   |            |            |            | 0.097***   |
| 3.3.1.2 Average segment length                 | 0.190***   |            |            |            | 0.207***   |
| 3.3.1.3 Average <i>positive</i> segment length | 0.040**    |            |            |            | 0.037      |
| 3.3.1.4 Average <i>negative</i> segment length | 0.013      |            |            |            | −0.078     |
| 3.3.2.1 Starts with negative sentiment         |            | −0.173**   |            |            | −0.032     |
| 3.3.2.2 Ends with negative sentiment           |            | −0.230***  |            |            | −0.180     |
| 3.3.3.1 Position of first positive segment     |            |            | −0.281***  |            | −0.220     |
| 3.3.3.2 Length of first positive segment       |            |            | 0.065***   |            | −0.005     |
| 3.3.3.3 Position of first negative segment     |            |            | −0.469***  |            | −0.200     |
| 3.3.3.4 Length of first negative segment       |            |            | 0.010      |            | 0.011      |
| 3.3.4.1 Position of last positive segment      |            |            |            | 0.053***   | 0.006      |
| 3.3.4.1 Length of last positive segment        |            |            |            | 0.383***   | −0.019     |
| 3.3.4.3 Position of last negative segment      |            |            |            | 0.159      | 0.169      |
| 3.3.4.4 Length of last negative segment        |            |            |            | 0.030**    | 0.067      |
| $R^2$                                          | 0.107      | 0.011      | 0.023      | 0.023      | 0.113      |

**Table A12:** OLS Regression with Student-Level Fixed Effects for Subcategory “3.3 Tone” from 3. *Emotional Features* Category (\*\* $p - value \leq 0.01$ , \*\*\*  $p - value \leq 0.001$ ).

| <i>Variable</i>                                               | <i>M 1</i> | <i>M 2</i> | <i>M 3</i> |
|---------------------------------------------------------------|------------|------------|------------|
| 3.4.1.1 Motivational words                                    | −24.762*** |            | −25.849*** |
| 3.4.2.1 Pronoun exclusive I and positive sentiment            |            | 0.249      | 0.285      |
| 3.4.2.2 Pronoun exclusive you and positive sentiment          |            | 0.904**    | 0.957**    |
| 3.4.2.3 Pronoun I & you and positive sentiment                |            | 0.422      | 0.612      |
| 3.4.2.4 Pronoun exclusive third-person and positive sentiment |            | 0.036      | −0.040     |
| 3.4.2.5 Pronoun others and positive sentiment                 |            | −0.070     | −0.036     |
| $R^2$                                                         | 0.008      | 0.003      | 0.012      |

---

## Appendix B: Lebenslauf

### Roman Rietsche

---

#### Persönliche Daten

---

Kontakt                      roman.rietsche@unisg.ch  
                                      +49 176 47135474  
Publikationsliste:        Alexandria: [6662](#), Google Scholar: [0FOGvPwAAAAJ](#)

---

#### Doktorat

---

Jul. 2016 – Apr. 2020      Ph.D. Programm in Management (Business Innovation)  
                                      Title: "Enhancing Student Performance Through Formative Feedback in Large Scale University Lectures"  
  
                                      **Erstgutachter:**  
                                      Prof. Dr. Jan Marco Leimeister, Institut für Wirtschaftsinformatik, Universität St. Gallen  
                                      **Zweitgutachter:**  
                                      Prof. Dr. Jan vom Brocke, Institut für Wirtschaftsinformatik, Universität Liechtenstein

---

#### Wissenschaftliche Auslandsaufenthalte

---

Jun. 2019 – Jan. 2020      Gastforscher  
                                      Cambridge Assessment Center, University of Cambridge  
  
Jan. 2019 – Mai 2019        Gastforscher  
                                      Management Information Systems, Fox School of Business, Temple University

---

#### Berufserfahrung

---

Mai 2020 – Heute            Projektleiter am Lehrstuhl von Prof. Dr. Jan Marco Leimeister  
                                      Institut für Wirtschaftsinformatik, Universität St. Gallen  
  
Jun. 2016 – Apr. 2020        Wissenschaftlicher Mitarbeiter am Lehrstuhl von Prof. Dr. Jan Marco Leimeister  
                                      Institut für Wirtschaftsinformatik, Universität St. Gallen  
                                      Betreuer des Forschungsrats des Instituts für Wirtschaftsinformatik der Universität St. Gallen  
                                      Forschungsprojekt mit der Siemens AG  
                                      – Verantwortung für das Projekt «Feed2Bot» Erhebung von  
                                            Mitarbeiterfeedback durch Chatbots  
                                      – Verantwortung für die wissenschaftliche Auswertung des Feedbacks und  
                                            Teilverantwortung beim Erstellen des Machine Learning Models

---

|                      |                                                                                                                                                                                                                                    |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Dez. 2013 – Mai 2016 | Senior Systems Engineer<br>Leitwerk AG, Freiburg                                                                                                                                                                                   |
|                      | <ul style="list-style-type: none"> <li>- Beratung und Umsetzung von IT-Sicherheit- und Netzwerkinfrastruktur-Projekte</li> <li>- Projektmanager für Grossprojekte</li> <li>- Leiter Competence Center Projektmanagement</li> </ul> |

---

## Akademische und schulische Ausbildung

---

|                       |                                                                                       |
|-----------------------|---------------------------------------------------------------------------------------|
| Feb. 2011 – Nov. 2013 | Wirtschaftsinformatik, Master of Science<br>Universität Mannheim                      |
| Jan. 2010 – Mar. 2011 | Information Technology, Bachelor of Science Honours, Edinburgh Napier University      |
| Sep. 2007 – Feb. 2011 | Computer Networking, Bachelor of Science<br>Furtwangen University of Applied Sciences |
| Sep. 2006 – Jul. 2007 | Fachhochschulreife<br>Gewerblich-Technische Schule Offenburg                          |
| Sep. 2003 – Jul. 2006 | IT-Systemelektroniker<br>Berufskolleg Offenburg                                       |
| Sep. 2002 – Jul. 2003 | Mittlere Reife<br>Realschule Ettenheim                                                |
| Sep. 1997 – Jul. 2002 | Hauptschule<br>Hauptschule Kippenheim                                                 |

---

## Berufsausbildung

---

|                       |                                                         |
|-----------------------|---------------------------------------------------------|
| Sep. 2003 – Jul. 2006 | IT-Systemelektroniker<br>Deutsche Telekom AG, Offenburg |
|-----------------------|---------------------------------------------------------|

---

## Auszeichnungen und Stipendien

---

|             |                                                                                                                                         |
|-------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| 2019        | Einjähriges Forschungsstipendium des Schweizer National Fonds (SNF, Doc Mobility)                                                       |
| 2019        | Betreuer der besten Abschlussarbeit im Bereich e-Learning im Jahr 2019, welche durch die Gesellschaft für Informatik ausgezeichnet wird |
| 2019        | Einladung zum Doctoral Consortium der Hawaii International Conference on System Sciences (HICSS)                                        |
| 2018        | Nominierung für den Best Paper Award Award der International Conference on Information Systems in San Francisco (ICIS)                  |
| 2017        | Nominierung für den Best Research-in-Progress Paper Award der European Conference on Information Systems (ECIS)                         |
| 2012 – 2013 | Deutschland Stipendium                                                                                                                  |
| 2011 – 2012 | Universität Mannheim Stipendium                                                                                                         |
| 2009        | Furtwangen University of Applied Science Stipendium                                                                                     |

---

## Lehrtätigkeit

---

|                     |                                                                                                                                                                                                                                                                                                                           |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Jun 2016 – Dez 2019 | Veranstaltungsbetreuer der Vorlesung Business Innovation II mit ca. 220 Studierenden auf Masterlevel <ul style="list-style-type: none"><li>- Erstellung des ganzheitlichen Lehr-Lernkonzeptes</li><li>- Verantwortlich für die dezentralen Prüfungsleistungen sowie Gruppenarbeiten und Abschlusspräsentationen</li></ul> |
| Jan 2017 – Dez 2018 | Akademische Betreuung von zwei Forschungs-, Praxis- und Ventureprojekten (FPV) sowie einer Gruppe in der Veranstaltung Angewandte Forschungsmethoden                                                                                                                                                                      |
| Jun 2016 – Jan 2019 | Betreuung von Bachelor- (5) and Masterarbeiten (4)                                                                                                                                                                                                                                                                        |

---

## Zertifikate

---

|      |                                                             |
|------|-------------------------------------------------------------|
| 2019 | CAS Hochschuldidaktische Ausbildung (10 ECTS – 300 Stunden) |
| 2014 | Prince 2 Foundation                                         |
| 2009 | ITIL v3 Foundation Examination                              |

---

## Drittmittelverzeichnis

---

|              |                                            |
|--------------|--------------------------------------------|
| Projektname  | Mitarbeiter im Zeitalter von Industrie 4.0 |
| Auftraggeber | Siemens AG                                 |
| Laufzeit     | Apr. 2018 – Apr. 2019                      |