Implementation and Validation of an Item Response Theory Scale for Formative Assessment

Stéphanie Berger

IMPLEMENTATION AND VALIDATION OF AN ITEM RESPONSE THEORY SCALE FOR FORMATIVE ASSESSMENT

Stéphanie Berger

IMPLEMENTATION AND VALIDATION OF AN ITEM RESPONSE THEORY SCALE FOR FORMATIVE ASSESSMENT

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. T.T.M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on Wednesday the 26th of June 2019 at 10:45 hours

by

Stéphanie Berger
born on the 5th of February 1983
in Muri AG, Switzerland

This dissertation has been approved by:

Supervisors:

prof. dr. ir. T.J.H.M. Eggen

prof. dr. U. Moser

Co-supervisor:

dr. ir. A.J. Verschoor

Printed by: Ipskamp Drukkers, Enschede

Cover designed by: Stéphanie Berger

ISBN: 978-90-365-4793-2

DOI: 10.3990/1.9789036547932

Copyright © 2019 S. Berger, Zurich, Switzerland.

All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

Graduation Committee

Chairman	prof. dr. T.A.J. Toonen	University of Twente, BMS
Supervisors	prof. dr. ir. T.J.H.M. Eggen prof. dr. U. Moser	University of Twente, BMS University of Zurich
Co-supervisor	dr. ir. A.J. Verschoor	Cito
Members	prof. dr. ir. B.P. Veldkamp prof. dr. C.A.W. Glas prof. dr. L.A. van der Ark prof. dr. M.F. van der Schaaf dr. K. Schildkamp dr. D. Joosten-ten Brinke	University of Twente, BMS University of Twente, BMS University of Amsterdam Utrecht University University of Twente, BMS The Open University

Contents

Chapter 1. General Introduction	1
1.1 Educational Assessment in Northwestern Switzerland.....	1
1.2 Vertical Scaling and Efficient Testing Based on IRT Methods	3
1.3 Research Objectives and Thesis Outline.....	4
1.4 References	8
Chapter 2. Linking Standardized Tests and an Online Item Bank for Formative Assessment	11
2.1 Introduction	12
2.2 Description and Comparison of the Standardized Tests and the Online Item Bank ..	13
2.2.1 Target Population.....	13
2.2.2 Assessment Types and Purposes.....	14
2.2.3 Content Specifications	16
2.2.4 Measurement Conditions	17
2.2.5 Summary of Similarities and Differences	19
2.3 IRT as Methodological Approach.....	19
2.3.1 Measurement Model	19
2.3.2 Item Calibration	21
2.4 Designs for Efficient Item Calibration and Testing	22
2.4.1 Data-Collection Designs for IRT-Based Linking.....	22
2.4.2 Targeted and Adaptive Testing for Efficient Parameter Estimation	24
2.5 Concept for Implementing a Common Vertical Scale for Mathematics.....	27
2.5.1 Establishing a Scale with Four Calibration Steps	27
2.5.2 Step 1: Calibrating the Standardized Third-Grade Test	29
2.5.3 Step 2: Calibration Assessments for Establishing the Vertical Scale.....	30
2.5.4 Step 3: Linking the Standardized Tests for Sixth, Eighth, and Ninth Grades	34
2.5.5 Step 4: Extending the Online Item Bank for Formative Assessment	36

2.6 Discussion and Further Research	37
2.6.1 Practical Challenges to Implementing the Proposed Concept.....	37
2.6.2 Research on Increasing Measurement Efficiency Under Practical Constraints	39
2.6.3 Conclusion	40
2.7 References	42
Chapter 3. Efficiency of Targeted Multistage Calibration Designs under Practical Constraints: A Simulation Study	51
3.1 Introduction	52
3.1.1 Accuracy and Efficiency in Rasch Model-Based Item Calibration.....	53
3.1.2 Designs for Item Calibration.....	55
3.1.3 Uncertainty of Item Difficulty during Test Construction.....	58
3.1.4 The Present Study	59
3.2 Method	59
3.2.1 Ability Distributions and Item Pool	59
3.2.2 Test Designs.....	60
3.2.3 Simulation with Limited Knowledge about Item Difficulty during Test Development.....	63
3.2.4 Item Response Generation and Calibration.....	64
3.2.5 Evaluation Criteria	64
3.3 Results	65
3.3.1 Distribution of Item Difficulty per Booklet or Module.....	65
3.3.2 Bias($\hat{\beta}$), Mean RMSE($\hat{\beta}$) and Mean Number of Observations per Simulation Condition	66
3.3.3 RMSE($\hat{\beta}$) and Mean Number of Observations per Item	69
3.4 Discussion	73
3.5 References	77
Chapter 4. Improvement of Measurement Efficiency in Multistage Tests by Targeted Assignment	81
4.1 Introduction	82
4.1.1 Efficient Measurement Based on Item Response Theory	83
4.1.2 Designs for Targeted Testing.....	84
4.1.3 The Present Study	87

4.2 Method	88
4.2.1 Ability Distributions and Population Distribution Conditions	88
4.2.2 Test Designs	89
4.2.3 Item Pools	92
4.2.4 Item Response Generation and Ability Estimation	92
4.2.5 Evaluation Criteria	93
4.3 Results	95
4.3.1 RMSE($\hat{\theta}$) of the Mixture Population per Design and Distribution Condition	95
4.3.2 RMSE($\hat{\theta}$) of the Ability Groups per Design and Distribution Condition	99
4.3.3 RMSE($\hat{\theta}$) in Relation to the Starting Module Length	104
4.4 Discussion	108
4.4.1 Limitations and Future Research	110
4.4.2 Conclusion and Practical Implications	110
4.5 References	112
Chapter 5. Development and Validation of a Vertical Scale for Formative Assessment in Mathematics	117
5.1 Introduction	118
5.1.1 Curriculum Lehrplan 21 as Content Framework	120
5.1.2 Vertical Scaling Based on Item Response Theory Methods	122
5.1.3 The Present Study	124
5.2 Method	125
5.2.1 Content-Related Item Difficulty	125
5.2.2 Calibration Design	127
5.2.3 Item Calibration	129
5.2.4 Item Analysis	130
5.2.5 Data Analysis	131
5.3 Results	133
5.3.1 Item Calibration	133
5.3.2 Content-Related Validation of the Vertical Scale	138
5.4 Discussion	145
5.4.1 Limitations	148

5.4.2 Conclusion	149
5.5 References	150
Chapter 6. Epilogue	157
6.1 Insights into Implementing a Vertical Scale for Mathematics in Northwestern Switzerland.....	157
6.2 Insights into Efficient Item Calibration Under Practical Constraints	158
6.3 Insights into Relative Efficiency of Targeted and Multistage Testing	160
6.4 Insights into the Vertical Scale's Validation from a Content Perspective	161
6.5 Conclusion and Outlook.....	162
6.7 References	164
Summary	169
Samenvatting	173
Acknowledgments.....	177
Research Valorization.....	179

Chapter 1. General Introduction

A vertical measurement scale is required to repeatedly assess and monitor students' abilities throughout their school careers (Young, 2006). Advanced computer technology, which is available today, serves as a foundation for implementing vertical scales and related complex measurement models, such as item response theory models (IRT; de Ayala, 2009; Lord, 1980), in computer-based assessment systems used to assess students within the classroom and provide formative feedback on a regular basis (Brown, 2013; Glas & Geerlings, 2009; Hattie & Brown, 2007; Wauters, Desmet, & Noortgate, 2010). Nevertheless, practical implementation of a vertical measurement scale is a challenging endeavor. The underlying IRT models are based on strict assumptions (e.g., Kolen & Brennan, 2014; Strobl, 2012; Wainer & Mislevy, 2000), which are not always perfectly met in practice. Furthermore, practical constraints like time and financial resources; willingness of schools, teachers, and students to participate in calibration studies; or differences in students' test-taking motivation across different assessment occasions can complicate the development and validation of a vertical IRT scale. Moreover, the available literature about IRT-based vertical scaling provides limited guidance for how to deal with the variety of practical settings and constraints (e.g., Briggs & Weeks, 2009; Ito, Sykes, & Yao, 2008; Tong & Kolen, 2007). Consequently, it is challenging to implement a vertical scale that accurately reflects the abilities specified in the underlying content specification or curriculum.

This thesis was motivated by practical challenges related to the implementation of a vertical scale to measure students' mathematics abilities throughout compulsory school in Northwestern Switzerland. This chapter provides an overview of educational assessment in Northwestern Switzerland as the practical context for this thesis, and introduces vertical scaling and efficient testing based on IRT methods as the major common theoretical themes of the studies presented in this thesis. At the end of this chapter, the research objectives and research questions for subsequent chapters are outlined.

1.1 Educational Assessment in Northwestern Switzerland

In 2012, four cantons (i.e., districts) in Northwestern Switzerland—Aargau, Basel-Landschaft, Basel-Stadt, and Solothurn—initiated a joint project to develop a new assessment system to measure and monitor students' abilities in mathematics¹ from grade three (in the middle of primary school) through grade nine (at the end of secondary school; Bildungsraum Nordwestschweiz, 2012). These four cantons commissioned the development of a system consisting of two different assessment instruments: (1) a set of four compulsory standardized

¹ Besides mathematics, the new assessment system also aims to assess students' German language abilities, the language used in schools, as well as students' English and French language abilities, the two foreign languages taught.

tests, called *Checks* (www.check-dein-wissen.ch), to assess students' abilities in grades 3, 6, 8, and 9; and (2) an online item bank for formative assessment, called *Mindsteps* (www.mindsteps.ch), with unrestricted access for all students and their teachers (Tomasik, Berger, & Moser, 2018). The goal of this system is to provide schools, teachers, and their nearly 100,000 students with objective results showing students' abilities and progress. Following the approach of data-based decision making in formative assessment (Schildkamp, Lai, & Earl, 2013; van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015), these results support teachers and students in defining appropriate learning goals; evaluating students' progress over time; and adjusting teaching, learning environments, or goals, when necessary (Hattie, 2009; Hattie & Timperley, 2007).

An important requirement for the system is a common vertical measurement scale that allows for comparing assessment results between the two instruments, as well as across cohorts and for individual students throughout different grades. Such a common measurement scale facilitates interpretation of the results for students, teachers, principals, and other stakeholders, and ensures self-regulated monitoring of students' progress between official measurement occasions (i.e., the four standardized tests). The scale should reflect students' abilities in accordance with the new curriculum, called *Lehrplan 21*, for the German-speaking area of Switzerland (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014, 2016b). For mathematics, *Lehrplan 21* calls for a continuous development of students' mathematics competencies from kindergarten through the end of compulsory school (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016a), which is in line with a domain definition of growth (cf. Kolen & Brennan, 2014), and thus fits into the content-related requirements of a vertical scale (Young, 2006).

To assess students' mathematics abilities with two instruments over seven school years, several thousand assessment items are required to build the core of the assessment system. However, calibration of these thousands of items is challenging due to several practical constraints. First, although the two instruments have overlapping purposes, are dedicated to the same target population, and are based on the same content specifications, differences in measurement conditions complicate the development of a common vertical scale (Kolen, 2007; Kolen & Brennan, 2014). Second, the total target population consists of approximately 13,000 students per school grade. At the same time, students and teachers are not obligated to participate in calibration studies, and time and financial resources are limited. Third, the available time to complete each standardized test is limited to two school lessons, and teachers and students might also have limited time to engage with the online item bank for formative assessment. Nevertheless, the assessment results should provide as much information as possible about students' current abilities. In this thesis, these practical constraints are contrasted with the theory of IRT-based vertical scaling, with the aim of identifying the most suitable approach to establish and validate a vertical scale that covers seven school grades, links two instruments, and represents the competencies described in the curriculum.

1.2 Vertical Scaling and Efficient Testing Based on IRT Methods

Vertical scaling, i.e., establishing a scale to measure abilities across multiple school years through item calibration, requires a powerful and flexible measurement approach. Often, vertical scales are based on IRT, which refers to a family of models that incorporate students' responses to each item in order to estimate students' abilities. Specifically, IRT models express the probability that a student answers an item correctly as a function of student ability and item parameters such as difficulty or discrimination. This thesis concentrates on the Rasch model (Rasch, 1960; Strobl, 2012), the most basic unidimensional IRT model. The Rasch model states that the probability of answering an item correctly is given by

$$P(X_{ij} = 1 | \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}, \quad (1.1)$$

where θ_i corresponds to the ability of the student i , and β_j refers to the difficulty of the item j . One advantage of IRT models in general, and of the Rasch model in particular, regarding vertical scaling, is that different item sets can represent the same underlying latent ability (Rost, 2004; Wainer & Mislevy, 2000). Consequently, assessment results can be linked and compared across individual students and over time, even though students worked on different item sets or test forms. Ideally, test forms are targeted to each student's ability level, so that students from higher grades are assigned to more difficult test forms than students from lower grades (Mislevy & Wu, 1996). Targeted testing is relevant for vertical scaling in order to adequately and efficiently calibrate a large item pool and assess the variation in students' abilities across multiple grades. Administering the same items to all students would be inefficient, especially when either the size of the calibration sample or the testing time are limited. A fixed number of items provides the most accurate information about a student's true ability under the Rasch model, if the difficulty of the items corresponds to the student's ability (Lord, 1980). Items that are too easy or too difficult provide not only limited information about the student's ability but might cause boredom, demotivation, or overstraining (e.g., Asseburg & Frey, 2013; Wise, 2014). Similarly, a fixed number of students in the calibration sample provides the most accurate information about the items' true difficulties if the students' abilities correspond to the difficulty of the items (Berger, 1991; Eggen & Verhelst, 2006; Stocking, 1988).

A good match between item difficulty and student ability will not only improve the efficiency of item difficulty and student ability estimation across grades, but also within grades. The match between item difficulty and student ability within each grade can be increased by dividing each grade group into more homogenous ability groups by means of additional ability-related background variables, such as marks provided by the teacher or performance-related school types. The match between item difficulty and student ability can also be improved based on student performance during test-taking, as in computerized adaptive testing (CAT; van der Linden & Glas, 2010; Wainer, 2000) or multistage testing (MST; Hendrickson, 2007; Yan, von Davier, & Lewis, 2014).

In theory, an IRT-based vertical scale serves as an ideal basis for flexible and efficient testing across different ability levels and over time. However, IRT models in general, and the Rasch model in particular, are based on strong statistical assumptions (e.g., Kolen & Brennan, 2014; Strobl, 2012; Wainer & Mislevy, 2000), which are not always met in practice. Furthermore, targeted testing, as well as adaptive testing by means of CAT or MST, require preliminary knowledge about the distribution of abilities within the student sample as well as the difficulty of the available items in order to ensure a good match between item difficulty and student ability. However, this knowledge is not always available in practice, especially when developing data collection designs to calibrate new items. Finally, efficient data collection designs and the resulting vertical scale, which meets the assumptions of the underlying IRT model, do not guarantee the scale's content validity.

1.3 Research Objectives and Thesis Outline

This thesis includes four studies related to the practical challenges of implementing and validating a vertical mathematics scale to assess third through ninth grade students in Northwestern Switzerland using two different assessment instruments. The studies review and extend the literature about IRT-based vertical scaling from a practical perspective by analyzing the similarities and differences between the two instruments and by evaluating the implications of possible differences on the vertical scaling procedure by comparing the efficiency of different calibration and test designs under practical constraints and suggesting an approach to investigate the content validity of the final vertical scale.

Chapter 2 lays the foundation for subsequent chapters by answering the following two questions:

- Q2.1 Do the four standardized tests and the online item bank for formative assessment share enough similarities to justify a common vertical scale across seven school grades?
- Q2.2 How could such a scale be realized in practice?

To address these two questions, Chapter 2 provides a detailed overview of the two assessment instruments and evaluates their similarities and differences. Chapter 2 also introduces different data collection designs for horizontal (i.e., within each grade) and vertical scaling (i.e., across different grades), and the related calibration procedures to estimate item difficulty on a common Rasch scale. Moreover, Chapter 2 presents targeted testing, based on ability-related background variables, and adaptive testing, based on performance, as two strategies to increase the accuracy of item difficulty and student ability estimates within restricted student and item samples. In order to calibrate and link the underlying item pool, a theoretical concept is suggested to implement the vertical scale and assess students' mathematics abilities in Northwestern Switzerland through four calibration steps: by using the standardized tests (steps one and three), dedicated calibration assessments (step two), and online calibration within the online item bank for formative assessment (step four). Furthermore,

topics for further research are identified in Chapter 2, with the aim of providing more guidance regarding the implementation of vertical scales under practical constraints. Three of the identified research topics serve as the basis for the studies presented in Chapters 3, 4, and 5.

Chapter 3 investigates the efficiency of different calibration designs for estimating item difficulties from the Rasch model under the practical constraint of limited knowledge about the items' true difficulties. This chapter provides an overview of three different calibration designs: (1) targeted calibration designs, which rely on ability-related background variables for assigning test forms of different difficulty levels; (2) multistage calibration designs, which assign the most appropriate modules based on student performance on a preliminary test part or module; and (3) targeted multistage calibration designs, a new design type, which uses both ability-related background variables and student performance to optimize the match between item difficulty and student ability, and thus refers to an extension of traditional targeted calibration designs. Chapter 3 focuses on this new design type by addressing the question:

Q3.1 Are targeted multistage calibration designs more efficient for item calibration than traditional targeted calibration designs?

Chapter 3 also points out that most previous studies on the efficiency of incomplete calibration designs have neglected the practical constraint of limited knowledge about the items' true difficulty when assembling the calibration design (Berger, 1991; Stocking, 1988). This knowledge is important when creating test forms or modules targeted to specific difficulty levels, as uncertainty about an item's true difficulty might result in items being misplaced into test forms or modules which are either too easy or too difficult. Thus, the second research question answered by Chapter 3 is:

Q3.2 How does limited a priori knowledge about item difficulty affect the efficiency of both targeted calibration designs and targeted multistage calibration designs?

Both questions are addressed in a simulation study, in which the calibration design and the accuracy of item distribution across the different test forms or modules within each design (i.e., number of misplaced items) are varied.

Chapter 4 addresses the fact that neither targeted testing based on ability-related background variables nor adaptive testing based on performance, by means of MST, can ensure that all students receive items that completely match their true abilities. Under the condition of targeted testing, some students might be disadvantaged by receiving a test form that is either too easy or too difficult because they significantly differ in their abilities from their group's mean abilities. The number of disadvantaged students might depend on the correlation between the ability-related background variable and students' true abilities. MST designs, on the other hand, mostly begin with a general starting module, which doesn't take into account the differences in students' abilities. The degree of discrimination of low- and high-ability students

by a general starting module might depend on the length of the starting module compared to the total test length.

Chapter 4 introduces targeted multistage test (TMST) designs, in which students are assigned to items based on ability-related background variables in the first stage and based on performance in subsequent stages in order to increase measurement efficiency. In particular, this chapter focuses on the question:

Q4.1 Do TMST designs achieve more accurate, and therefore more efficient, ability estimates than traditional targeted test designs or MST designs with one starting module?

In addition to this general question, Chapter 4 explores the efficiency of TMST designs from three specific perspectives by investigating the following three research questions:

Q4.2 To what extent does the efficiency gain through TMST designs depend on the correlation between the ability-related background variable and students' true ability?

Q4.3 To what extent do different ability groups profit or are disadvantaged by TMST designs compared to targeted and MST designs?

Q4.4 To what extent does the efficiency gain through TMST designs depend on the length of the starting module compared to the total test length?

All four questions are addressed in a simulation study, in which the test design, the correlation between students' abilities and the ability-related background variable, and the length of the starting module in relation to the total test length are varied.

Chapter 5 directs attention toward validating a vertical scale from a content perspective. Specifically, this chapter reports on the actual implementation and validation of a vertical Rasch scale for assessing third through ninth grade students' mathematics abilities in Northwestern Switzerland, based on the calibration assessments described in Chapter 2 (i.e., step two of the suggested calibration procedure). To validate the scale from a psychometric perspective, item analysis is performed, and two different calibration procedures (i.e., concurrent and grade-by-grade calibration) are applied to detect potential calibration problems related to multidimensionality. To validate the decisions made during test development and item calibration, from a content perspective, the empirical item difficulty parameters are contrasted with the items' content-related difficulties, according to their assignment to specific competence levels as described in the curriculum, *Lehrplan 21*. The following three research questions are addressed in this chapter:

Q5.1 Do the items developed on the basis of the curriculum, *Lehrplan 21*, and targeted to third through ninth grade, fit a unidimensional vertical Rasch scale?

- Q5.2 Do the item calibration's empirical outcomes—i.e., item difficulty estimates—match the theoretical, content-related item difficulties that reflect the curriculum's underlying competence levels?
- Q5.3 Does the match between the empirical item difficulty estimates and the theoretical, content-related item difficulties differ for items related to different curriculum cycles, domains, or competencies?

Empirical data from a cross-sectional calibration study that includes 520 mathematics items and 2,733 third through ninth grade students from Northwestern Switzerland serves as the basis for answering these three research questions.

This thesis concludes with an Epilogue, in which the primary research questions posed by the preceding chapters are reconsidered. The main findings of related studies are summarized, and an outlook for further research is provided.

1.4 References

- Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, 55(1), 92–104.
- Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, 15(3), 293–306. doi:10.1177/014662169101500310
- Bildungsraum Nordwestschweiz (BR NWCH). (2012). *Checks und Aufgabensammlung im Bildungsraum Nordwestschweiz: Porträt. Bildungsraum Nordwestschweiz (BR NWCH)*. Retrieved from Bildungsraum Nordwestschweiz (BR NWCH) website: <http://www.bildungsraum-nw.ch/medien/dokumente-pdf>
- Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3–14. doi:10.1111/j.1745-3992.2009.00158.x
- Brown, G. T. L. (2013). AsTTle – A national testing system for formative assessment: How the national testing policy ended up helping schools and teachers. In S. Kushner, M. Lei, & M. Lai (Eds.), *Advances in Program Evaluation: Volume 14. A national developmental and negotiated approach to school self-evaluation* (Vol. 14, pp. 39–56). Bradford, UK: Emerald Group Publishing Limited. doi:10.1108/S1474-7863(2013)0000014003
- de Ayala, R. J. (2009). *The theory and practice of item response theory. Methodology in the social sciences*. New York: Guilford Press.
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2014). *Lehrplan 21: Rahmeninformationen*. Luzern. Retrieved from http://www.lehrplan.ch/sites/default/files/lp21_rahmeninformation_%202014-11-06.pdf
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2016a). *Lehrplan 21: Mathematik*. Luzern. Retrieved from https://v-fe.lehrplan.ch/container/V_FE_DE_Fachbereich_MA.pdf
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2016b). *Lehrplan 21: Überblick*. Luzern. Retrieved from https://v-fe.lehrplan.ch/container/V_FE_Ueberblick.pdf
- Eggen, T. J. H. M., & Verhelst, N. D. (2006). Loss of information in estimating item parameters in incomplete designs. *Psychometrika*, 71(2), 303–322. doi:10.1007/s11336-004-1205-6
- Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, 35(2), 83–88. doi:10.1016/j.stueduc.2009.10.006
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.
- Hattie, J. A. C., & Brown, G. T. L. (2007). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, 36(2), 189–201. doi:10.2190/ET.36.2.g

- Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26(2), 44–52. doi:10.1111/j.1745-3992.2007.00093.x
- Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*, 21(3), 187–206. doi:10.1080/08957340802161741
- Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
- Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Reports Series No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* [Textbook test theory, test construction] (2nd ed.). Psychologie Lehrbuch. Bern, Switzerland: Verlag Hans Huber.
- Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education: Challenges and opportunities*. Dordrecht, the Netherlands: Springer.
- Stocking, M. L. (1988). Scale drift in on-line calibration. *ETS Research Report Series*, 1988(1), i-122. doi:10.1002/j.2330-8516.1988.tb00284.x
- Strobl, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* [The Rasch model: A coherent introduction for students and practitioners] (2nd ed.). Mering, Germany: Rainer Hampp Verlag.
- Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, 9, 2245. doi:10.3389/fpsyg.2018.02245
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227–253. doi:10.1080/08957340701301207
- van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, 22(3), 324–343. doi:10.1080/0969594X.2014.999024
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.

- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (2nd ed., pp. 61–100). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wauters, K., Desmet, P., & Noortgate, W. van den. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, 26(6), 549–562. doi:10.1111/j.1365-2729.2010.00368.x
- Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, 2(1), 1–17. doi:10.7333/1401-0201001
- Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.
- Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–485). Mahwah, NJ: Lawrence Erlbaum Associates.

Chapter 2. Linking Standardized Tests and an Online Item Bank for Formative Assessment

Abstract

Item response theory (IRT) refers to a powerful and flexible measurement approach that allows for linking various test forms within and across different ability levels. In this paper, we elaborate a concept for implementing IRT-related calibration procedures and data-collection designs in a practical context for assessing students' mathematics ability. More specifically, we introduce a set of four standardized tests and an online item bank for formative assessment that require a common vertical measurement scale to assess and monitor students' abilities throughout seven years of compulsory school. We describe the standardized tests and online item bank by evaluating their similarities and differences regarding target population, assessment types and purposes, content specifications, and measurement conditions. Furthermore, we provide an overview of different IRT-related calibration procedures for linking, data-collection designs for horizontal and vertical scaling, and we introduce the idea of targeted and adaptive testing based on the Rasch model to increase measurement efficiency when available student or item samples are limited. By integrating the two instruments' similarities and differences with the theoretical background on data-collection designs and item calibration in a Rasch framework, we define four calibration steps to establish a vertical scale that links the two instruments. In the discussion, we summarize the main practical challenges for implementing our concept in our specific context. Moreover, we stress the need for validating the final scale from a psychometric, as well as content, perspective, and we point out the need for empirical research on efficient calibration and test designs under consideration of practical contexts and related practical constraints.

2.1 Introduction

In recent years, school administrators and teachers' need for objective instruments to assess individual students, evaluate classes and schools, and monitor educational systems has increased in Switzerland (Moser, 2009). At the beginning of the 21st century, Switzerland started to develop minimal educational standards (Schweizerischen Konferenz der kantonalen Erziehungsdirektoren, 2007). Therefore, a need arose for instruments that allow for evaluating whether students meet these minimal standards. However, the Swiss school system's decentralized organization and its numerous curricula impeded monitoring students' abilities on a national level. The first initiative to assess and monitor students' learning progress in a wider area started in 2012 in Northwestern Switzerland. Four cantons (i.e., districts) with a population of approximately 13,000 students per school grade started a joint project with the objective of providing students, teachers, and schools with instruments to assess and monitor students' abilities as a basis for advancing their development, improving teaching, and evaluating school-development programs (Bildungsraum Nordwestschweiz, 2012). Specifically, the four cantons initiated the development of two distinct assessment instruments: (1) a set of compulsory standardized tests with well-defined administrative time points, called *Checks* (www.check-dein-wissen.ch) and (2) an online item bank for formative assessment, called *Mindsteps* (www.mindsteps.ch), which offers computer-based linear and adaptive assessments on demand (Tomasik, Berger, & Moser, 2018). Both instruments intend to assess students' abilities in primary and secondary school (i.e., from third through ninth grades) within four subjects: German, the schools' language; English and French, the two foreign languages taught; and mathematics.

One important requirement of the two instruments is that they share a common reporting scale that allows for directly comparing the two instruments' outcomes (Bildungsraum Nordwestschweiz, 2012). Furthermore, the reporting scale is conceptualized as a vertical scale (Young, 2006) that not only allows for comparing consecutive cohorts' performance (cross-sectional comparison), but also enables the comparison of assessment results over time to analyze individual students' progress (longitudinal comparison). A scale with these two features is supposed to ensure that teachers and students can use the online item bank for formative assessment autonomously to monitor their progress between two standardized tests in relation to previous results. Furthermore, a joint reporting scale for both instruments and for several age groups not only facilitates the interpretation of assessment outcomes for students, teachers, and other stakeholders, but also enhances assessment instruments' acceptance.

In this paper, we focus on the level of individual students² and elaborate, from a theoretical perspective, on the psychometric challenges related to the development of a vertical reporting scale that links the two instruments using the example case of mathematics. In doing

² The implications of aggregated results on the class, school, or system levels are beyond this paper's scope.

so, we identified three principal challenges to establishing a common vertical reporting scale: First, the two instruments require substantial similarities to justify a common scale. Second, a measurement approach is required that allows for linking results from various assessment forms that different students answer at one point in time, and that identical students answer at different points in time. Finally, suitable and efficient data-collection designs are required for norming (i.e., calibrating) the standardized tests and the items within the online item bank for formative assessment.

Against this backdrop, in this paper's first section, we present the two instruments in more detail by elaborating on their similarities and differences in assessing individual students' mathematics abilities. To this end, we describe and compare the two instruments regarding their target population, assessment types and purposes, content specifications, and measurement conditions (Kolen, 2007; Kolen & Brennan, 2014). In the second section, we introduce item response theory (IRT, e.g., Hambleton, Swaminathan, & Rogers, 1991; Rost, 2004; Wainer & Mislevy, 2000) or rather the Rasch model (Rasch, 1960; Strobl, 2012), as a measurement approach to link different test forms, and we elaborate on related calibration procedures for linking purposes. In the third section, we describe IRT-related data-collection designs for establishing scales that cover different test forms targeted to one school grade (i.e., horizontal scales), as well as different school grades (i.e., vertical scales; Young, 2006). Furthermore, we elaborate on targeted testing based on ability-related background variables and on adaptive testing based on performance as methods for increasing these designs' efficiency in estimating item difficulty and student ability under the Rasch model. In the fourth section, we integrate the three previous sections by illustrating a concrete concept for the practical implementation of a common vertical Rasch scale for the two assessment instruments to assess students' mathematics ability from third through ninth grade. Finally, we conclude the paper with a summarizing discussion and an outlook on options for further research.

2.2 Description and Comparison of the Standardized Tests and the Online Item Bank

Linking multiple assessments related to different instruments and school grades to one reporting scale is justified only if the assessments share enough similarities. Following Kolen (2007) and Kolen and Brennan (2014), we used four categories of assessment features to describe the core features of the standardized tests and online item bank for formative assessment and to illustrate their similarities and differences: (1) target population; (2) assessment types and purposes; (3) content specifications; and (4) measurement conditions.

2.2.1 Target Population

The target population is an important assessment feature because the applicable linking processes depend on the composition of the student population for which a scale is developed

(Kolen, 2007). Differences in gender, race, geographic region, or age might affect the linking relationship of two assessments or assessment instruments (Dorans, 2004; Kolen, 2004) and need to be considered when selecting a data-collection design and related calibration procedures. On a general level, the standardized tests and online item bank for formative assessment are both dedicated to the same target population: Both instruments intend to measure and monitor students' ability in Northwestern Switzerland from third through ninth grade (i.e., from primary school through the end of compulsory school; Bildungsraum Nordwestschweiz, 2012). Nevertheless, as illustrated in Figure 2.1, the two instruments differ in the number of related administration occasions. The standardized tests take place at four predefined points in time: at the beginning of third and sixth grade in primary school and at the end of eighth and ninth grade in secondary school. Furthermore, the standardized tests are compulsory for all students in these four school grades. Conversely, teachers—and to some extent, students—are free to choose whether, when, and how often they engage with the online item bank for formative assessment. This flexibility might result in different user behavior among teachers and students, depending on various factors, e.g., students' age, class composition, the school's information technology (IT) infrastructure, or teachers' IT user knowledge that, in turn, might influence whether a particular subpopulation of those taking the standardized assessments actually uses the online item bank.

Figure 2.1. Administration of the four standardized tests and the formative online assessments during primary and secondary school. $N \approx 13,000$ students per school grade.

2.2.2 Assessment Types and Purposes

Generally, extant literature distinguishes between two different assessment types with specific purposes for assessing individual students' abilities: *summative* and *formative assessments* (e.g., Bloom, Hastings, & Madaus, 1971). Summative assessments aim to measure what students have learned over a certain period of time and provide results (i.e., a summary) at the end of a learning process (Sadler, 1989). Often, summative assessments result in a diploma or certification, and serve as a basis for selection or placement decisions. In contrast, formative

assessments take place at the beginning of or during the learning process, with the objective of providing information for guiding and improving learning (van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). Particularly, formative-assessment outcomes serve as a basis for defining appropriate learning goals, evaluating progress toward these goals, and determining the next steps along students' learning paths (Black & Wiliam, 1998; Hattie, 2009; Hattie & Timperley, 2007). Notably, van der Kleij et al. (2015) identified three different approaches to formative assessment: (1) *Data-based decision making* originates from the No Child Left Behind Act in the United States and places a strong emphasis on monitoring the attainment of specific learning targets through objective data (Schildkamp & Kuiper, 2010; Schildkamp, Lai, & Earl, 2013). (2) *Assessment for learning* focuses on the quality of the learning process and emphasizes the importance of providing students with feedback (Stobart, 2008). (3) *Diagnostic testing* originated from the intention to identify students with special educational needs and focuses on the detailed assessment of students' problem-solving processes (Crisp, 2012).

Unlike this clear distinction between summative and formative assessments, Bennett (2011) argues that both assessment types often share similarities in their purposes, yet differ in their primary purpose. In his opinion, "... summative tests, besides fulfilling their primary purposes, routinely advance learning, and formative assessments routinely add to the teacher's overall informal judgments of student achievement" (Bennett, 2011, p. 7). Following Bennett's line of argumentation, we classify the standardized tests as summative assessments whose primary purpose is to assess the outcome of learning and whose secondary purpose is to provide assessment outcomes for guiding and fostering further learning activities (Bennett, 2011). The standardized tests' objective in mathematics is to provide information about students' current ability in mathematics and in related mathematics domains at four selected points in time during compulsory school (Bildungsraum Nordwestschweiz, 2012). Their test results indicate students' competency levels at these junctures (criterion-referenced information; Betebenner, 2009; Bundesinstitut für Bildungsforschung, Innovation & Entwicklung, 2011; Reusser, 2014). Simultaneously, these test results help students compare themselves with their reference groups (norm-referenced information; Betebenner, 2009; Moser, 2009). In line with summative assessments' primary purpose, both criterion-referenced and norm-referenced information helps teachers make fair and accurate selection decisions at the end of primary school and provide students with a certificate on their abilities at the end of secondary school to use when applying for apprenticeships. Furthermore, standardized tests' results also support students and teachers in identifying individual students' strengths and knowledge gaps, serving as a starting point for defining new learning goals and planning upcoming learning and teaching activities. Thus, in line with the approach of data-based decision making in formative assessment (Schildkamp et al., 2013; van der Kleij et al., 2015), the objective data collected through standardized tests also foster future learning. Consequently, we argue that the standardized tests' secondary assessment purpose is formative.

Conversely, we classify the online item bank as a formative assessment instrument whose primary purpose is to provide assessment outcomes for guiding and fostering further

learning activities and whose secondary purpose is to assess learning outcomes (Bennett, 2011). The online item bank's key advantages are that it allows for repeated on-demand administration of tailored assessments for students with different ability levels, and it provides immediate reports (e.g., Hattie & Brown, 2007; Wainer, 2000b). Based on Hattie's concept of visible learning (Hattie, 2009), the online item bank for formative assessment aims to help students and teachers during the school year identify students' current strengths and weaknesses in mathematics in general, and in related mathematics domains and competencies in particular. This criterion-referenced information can serve as a starting point for defining individual learning goals, evaluating progress toward these goals, and defining appropriate subsequent learning steps (Hattie, 2009; Hattie & Timperley, 2007; van der Kleij et al., 2015). Furthermore, periodic assessments allow for measuring students' progress throughout the school year and across compulsory school grades. We claim that data-based decision making, characterized by the collection of objective data on students' current abilities best describes the online item bank's formative approach. Simultaneously, we argue that the online item bank also has a summative function as secondary purpose because information about students' current abilities is useful in evaluating learning and teaching activities that took place before an assessment. In addition, periodic assessments allow for analyzing the relationship between students' progress and different learning and teaching interventions over time.

2.2.3 Content Specifications

Content specifications refer to the framework that defines the specific content areas that an assessment or test aims to cover (Webb, 2006, p. 155). Such a framework is a crucial factor in ensuring content validity of an assessment or assessment instrument and the related measurement scale. Usually, a curriculum or content standards serve as the foundation for developing content specifications (Webb, 2006). Due to the Swiss school system's decentralized organization, each canton had its own curriculum until a few years ago. In 2014, experts published the first intercantonal curriculum for all German-speaking cantons of Switzerland (i.e., 21 out of 26 cantons), called *Lehrplan 21* (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014; see also www.lehrplan.ch). The curriculum describes the competencies that students should acquire from kindergarten through the end of compulsory school, thereby serving as a basis on which teachers and schools should plan their teaching and evaluate students' progress (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014, 2016b). Within the subject of mathematics, the curriculum is structured hierarchically into three domains, 26 competencies, and various competence levels (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016a). Within each competency, the curriculum calls for a continuous development of the competency over the school years, whereas mastering lower competence levels is a precondition for mastering more advanced competence levels (see also Bundesinstitut für Bildungsforschung, Innovation & Entwicklung, 2011; Reusser, 2014). Furthermore, the curriculum distinguishes between three different cycles, ranging from kindergarten to second grade, third to sixth grade, and seventh to ninth grade. For each cycle,

it defines basic requirements that refer to the minimal competence levels that students need to master by the end of the cycle. In addition, it states two points of orientation at the end of the fourth and eighth school grades. The cycles, basic requirements, and orientation points anchor the competence levels—and, thus, the development of competencies—across kindergarten and the nine compulsory school years. However, the curriculum focuses much more on the development of students' competence levels across grades than on specific competencies within a particular school grade; thus, the curriculum follows a domain definition of growth (Kolen & Brennan, 2014, pp. 429–431).

Thanks to its intercantonal scope, clear hierarchical structure, and domain definition of growth, the curriculum, *Lehrplan 21*, serves as an ideal basis for general content specifications of both the standardized tests and the online item bank, as well as a framework for developing and classifying all related assessment items. However, on the level of single tests or assessments, content specifications differ between the two instruments. Content experts assemble the four standardized tests annually to assess general mathematics ability among the four aforementioned target school grades on subject and domain levels reliably within the available testing time of two school hours (i.e., 90 minutes). Students, teachers, and school principals cannot change assessment content. In contrast, the online item bank for formative assessment is conceptualized as a big pool of thousands of items (i.e., approximately 10,000 items for mathematics alone) that are accessible through a web-based application, with teachers having three options for creating their own customized online assessments based on this item pool (Tomasik et al., 2018). First, they can assess their students on domain level through computerized adaptive tests (CATs; van der Linden & Glas, 2010; Wainer, 2000a), in which the system selects the most informative items based on each student's test performance. These assessments' content mixture is very similar to that within a single domain of the standardized tests. Second, teachers can narrow assessment content down to one, two, or three specific competence levels. Out of this selection, the system creates linear assessments that comprise a higher volume of similar items than standardized tests, thereby covering more specific content. Third, teachers also can create assessments by filtering the item bank by specific content topics and item difficulty, and by manually selecting preferred items. For assessments created through this third creation option, it largely depends on teachers' specifications on whether their content is comparable to that of standardized tests.

2.2.4 Measurement Conditions

Kolen (2007) distinguishes measurement conditions that test developers directly control from those that lie outside their control. Examples of directly controllable conditions include test design, administration mode (i.e., computer- vs. paper-based), instructions, and scoring procedures. Examples of indirectly controllable measurement conditions include the stakes that students or teachers associate with an assessment (i.e., low vs. high stakes) and students' motivation to take a test and make an effort while taking it. Differences between controllable

and uncontrollable measurement conditions might influence assessment outcomes and need to be considered during the linking process (e.g., Eignor, 2007; Mittelhaeuser, Béguin, & Sijtsma, 2011, 2013, 2015).

The standardized tests and online item bank for formative assessment differ under several of these measurement conditions. Furthermore, differences also exist in measurement conditions between the standardized tests. The two standardized tests for primary school are designed as linear, paper-based tests because primary-school students mainly are accustomed to working on paper. Moreover, primary schools often lack an IT infrastructure that is large and modern enough for administering computer-based tests in class (Bättig, Gut, & Schwab, 2011; Petko, Mitzlaff, & Knüsel, 2007; Petko, Prasse, & Cantieni, 2013). Conversely, the two standardized tests for secondary school are conceptualized as computer-based, multistage tests (MSTs; Hendrickson, 2007; Yan, von Davier, & Lewis, 2014; Zenisky, Hambleton, & Luecht, 2010) to address variations in abilities across the three performance-related school types within secondary school. In particular, all students start with a general test part of intermediate difficulty (i.e., starting module), then subsequently are routed, based on their performance, to one of five modules varying in difficulty in the three additional test parts (i.e., stages). The administration of all four standardized tests is regulated strictly to achieve objective results. The tests are compulsory, take place during a predefined period (i.e., two weeks for primary-school tests and six weeks for secondary-school tests), and test-administration time is limited to two school hours. Teachers are responsible for test administration, and they must follow specific administration instructions that aim to ensure a standardized test administration across all schools and classes. The tests' development, analysis of students' answers, and reporting are centralized. Due to the obligation to participate in the standardized tests, a high degree of standardization, and external organization of the assessment analysis, we argue that teachers and students might perceive the standardized tests as rather high-stakes assessments, even though no direct decisions or consequences related to their outcomes exist.

All assessments that are created based on the online item bank for formative assessment refer to linear computer-based assessments, or CATs, independent of whether they are targeted to primary- or secondary-school students. Furthermore, the system scores the assessments automatically and provides immediate reports at the end of each assessment. However, no regulations or instructions exist for administration of the assessments. Instead, teachers have much flexibility in the type and number of assessments they create, and in whether they assign assessments to individual students, groups of students, or entire classes. Some teachers might use the assessments like class exams, while others might assign assessments as exercises or homework to individual students. This flexibility might result in a broad range of measurement conditions regarding test designs and instructions. Consequently, students' perceptions of the assessments and their test-taking motivations might vary depending on the specific conditions that each teacher creates. Nevertheless, we claim that students and teachers generally might perceive these assessments as low-stakes assessments because of their primary formative function.

2.2.5 Summary of Similarities and Differences

In sum, we identified similarities and differences between the standardized tests and the online item bank for formative assessment in all four categories that we used for our comparison (i.e., target population, assessment types and purposes, content specifications, and measurement conditions; see Kolen, 2007; Kolen & Brennan, 2014). Regarding similarities, we elaborated on how both instruments are targeted to similar populations, namely students in Northwestern Switzerland. Furthermore, the summative standardized tests and the online item bank for formative assessment share overlapping assessment purposes even though they differ in their primary purposes. Moreover, the two instruments are based on the same content framework, namely the German-speaking region of Switzerland's new curriculum, *Lehrplan 21*. In addition, the standardized tests for secondary school and the assessments from the online item bank are computer-based and partially allow for adapting item selection to individual students' abilities.

Regarding differences between the two instruments, we discussed how the student samples within the target population, which take assessments related to the two instruments, differ to a certain extent due to differences in the number of administration occasions (i.e., four vs. unlimited) and degree of obligation (i.e., compulsory vs. optional). In addition, the two instruments differ in their primary assessment purposes (i.e., summative vs. formative), as well as in their broadness in assessment content (general vs. specific), depending on whether teachers focus on domains, competencies, or topics when creating assessments from the online item bank. Furthermore, we identified differences in administration mode on the primary-school level (i.e., paper-based vs. computer-based) and general differences in administration conditions (i.e., standardized vs. flexible) that, in turn, might affect the association of assessment stakes (i.e., high vs. low) and result in differences in students' test-taking motivation (i.e., high vs. low).

We conclude, from our comparison, that the similarities—especially overlapping assessment purposes, shared content framework, and similar target populations—justify linking the two instruments and developing a common reporting scale. Nevertheless, we argue that it is important to factor in similarities and differences when selecting adequate measurement and linking methods. More precisely, the identified differences are especially important in detecting potential challenges in the linking process and in identifying areas for further research.

2.3 IRT as Methodological Approach

2.3.1 Measurement Model

Establishing a vertical measurement scale for assessing students with tests related to different assessment instruments across multiple school grades requires a powerful and flexible measurement model that allows for linking various test forms. IRT offers the advantage that

different item sets can represent the same unidimensional latent construct (i.e., ability; e.g., Rost, 2004; Wainer & Mislevy, 2000). In contrast to classical test theory, which concentrates on the sum score of a test, IRT factors in each student's specific responses to each item when estimating a student's ability (Kolen & Brennan, 2014). As long as test forms are linked and related to the same underlying construct or scale, students can work on different test forms, yet get comparable results. Consequently, it is possible to update test forms by exchanging items between administrations without changing the original reporting scale, and it is possible to link tests dedicated to different school grades and report their results on the same scale. Moreover, IRT is not limited to linking different test forms, allowing for linking various subsets of items that measure the same latent construct and originate from the same item pool. This property is fundamental to enabling CATs, in which each student answers a unique set of items out of a calibrated item bank depending on his or her performance while taking a test (van der Linden & Glas, 2010; Wainer, 2000a).

IRT models express the probability that a student will answer an item correctly as a function of student ability and item parameters, such as difficulty or discrimination. In this paper, we focus on the Rasch model (Rasch, 1960; Strobl, 2012), the most basic unidimensional IRT model. For the Rasch model, the probability of a student answering an item correctly is given by

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}, \quad (2.1)$$

in which θ_i corresponds to the ability of student i , and β_j represents the difficulty of item j . According to this model, a student can answer an item correctly with a difficulty equal to his or her ability with a probability of 50 percent. The probability of correctly answering items that are easier than one's ability is higher, while the probability of correctly answering items that are more difficult is lower. Thus, item difficulty and student ability are represented on the same scale. Due to this special relationship, we can describe, through item content, what students with specific abilities most likely already know and what could be the next steps along their learning paths (i.e., what content they are not yet mastering sufficiently), thereby creating meaningful scales.

Besides its advantages, it is important to point out that IRT models in general and the Rasch model in particular are based on strong statistical assumptions (e.g., Kolen & Brennan, 2014; Strobl, 2012; Wainer & Mislevy, 2000). First, local independence must hold, i.e., a student's responses to different items must be statistically independent from each other when controlling for the student's ability (Kolen & Brennan, 2014). Second, item parameters need to be invariant across different administrations or age groups (Rupp & Zumbo, 2016). Third, unidimensional IRT models assume that all items refer to the same underlying unidimensional construct. Finally, IRT models assume a specific monotonic increasing function between students' ability and the probability of correctly answering an item, whereas the Rasch model

additionally states equal slopes for all items. When applying IRT methods, we carefully need to evaluate whether these assumptions are met sufficiently during data analysis.

2.3.2 Item Calibration

From a psychometric perspective, calibrating items is one of the core steps when implementing a new measurement scale based on IRT methods. For the Rasch model, item calibration refers to establishing model fit and estimating item difficulty parameters based on response data through maximum likelihood estimation procedures (Eggen & Verhelst, 2011; Vale & Gialluca, 1988). Generally, three calibration procedures exist for mapping parameters of items administered to different groups of students to a common IRT scale: (1) *concurrent calibration*; (2) *separate calibration with equating*; and (3) *fixed parameter calibration* (Kim, 2006; Kolen & Brennan, 2014). Under the concurrent procedure (Wingersky & Lord, 1983), item parameters of all items are estimated in one single calibration run, whereby different underlying population ability distributions need to be specified if we apply this procedure to different ability groups (e.g., students from different school grades; DeMars, 2002; Eggen & Verhelst, 2011). This procedure directly maps all item parameters to one common scale through linking items (i.e., items shared by multiple test forms). The second approach entails estimating parameters for each test form separately and equating different forms subsequently by transforming the parameters into a common scale through linear transformations. The transformation constants can be estimated using different methods, e.g., the mean/sigma method, mean/mean method, or different characteristic curve-transformation methods (see Kolen & Brennan, 2014, for an overview). Finally, under the fixed parameter calibration procedure (Keller & Hambleton, 2013; Keller & Keller, 2011; Kim, 2006), item calibration starts with a base assessment or scale with known item parameters. When calibrating new related test forms, all parameters of linking items (i.e., items included in the old and new test form) are fixed to their known parameters when calibrating new test forms. Thus, the linking items serve as anchors for aligning additional items to the base scale.

From a theoretical perspective, concurrent calibration might be superior to the other two procedures. Kolen and Brennan (2014, p. 444) argue that concurrent calibration leads to more stable results because it processes all available response data at once for estimating the parameters. Furthermore, the concurrent procedure is less error-prone because estimating transformation constants, creating the potential for additional estimation errors, is unnecessary, and it is more efficient because it requires only one calibration run (Briggs & Weeks, 2009). However, in a practical context, it often is impossible to postpone calibration until data from all test forms are available. Instead, new items or tests often need to be aligned with already-operational calibrated tests or item banks. Such an alignment is possible by equating the separately calibrated test forms or with fixed parameter calibration, but it contradicts the idea of concurrent calibration (Kim, 2006). An additional advantage of separate calibration and fixed parameter calibration is that these two procedures use smaller and simpler data sets than

concurrent calibration. As a result, the estimation procedure converges faster, and convergence problems are less likely. Furthermore, in a longitudinal context, separate calibration might offer two additional advantages compared with the other two procedures (Hanson & Béguin, 2002; Kolen & Brennan, 2014). First, it allows for directly comparing item parameter estimates between two different administrations, thereby facilitating investigation of potential deficits in parameter invariance. Investigating parameter invariance is also possible under the other two calibration procedures, but it requires—depending on calibration software—additional calibration runs for data subsamples. Second, separate calibration might be more robust against violation of the unidimensionality assumption because it only considers data from two ability groups in one equating step (Béguin & Hanson, 2001; Béguin, Hanson, & Glas, 2000; Hanson & Béguin, 2002). Thus, separate calibration might be the first choice if the empirical data do not sufficiently fit the IRT model.

Results from previous empirical studies that have compared the three calibration procedures are mixed. Some studies have found concurrent calibration to be superior to separate calibration when establishing a vertical scale (e.g., Hanson & Béguin, 1999; Kim & Cohen, 1998), while others have reported the opposite finding (e.g., Béguin et al., 2000; Ito, Sykes, & Yao, 2008). Similarly, the fixed parameter calibration procedure outperformed the other two procedures under some conditions (e.g., Keller & Hambleton, 2013; Keller & Keller, 2011, 2015). Generally, these mixed results suggest that, in practice, the most suitable calibration procedure might depend largely on the scale's specific measurement objectives, changes in ability distribution across grades (e.g., Keller & Hambleton, 2013), or the extent to which the data meet strict assumptions of unidimensionality and parameter invariance concerning the underlying IRT model (Béguin et al., 2000; Pohl, Haberkorn, & Carstensen, 2015; Tong & Kolen, 2007). Due to these complex interactions, it is difficult to draw general conclusions that are applicable to concrete practical situations. Therefore, different researchers recommend applying and comparing several calibration procedures in practice, especially when developing vertical scales (Hanson & Béguin, 2002; Kolen & Brennan, 2014). In addition, a need exists for external, content-related validation criteria to justify the scale, resulting from concrete calibration decisions in a concrete practical situation (Briggs & Weeks, 2009; Dadey & Briggs, 2012; Harris, 2007; Ito et al., 2008; Tong & Kolen, 2007).

2.4 Designs for Efficient Item Calibration and Testing

2.4.1 Data-Collection Designs for IRT-Based Linking

The exact procedure for calibrating and linking different item sets based on IRT strongly depends on the ability range that the scale intends to cover and on the selected data-collection design (i.e., calibration design). Below, we describe different data-collection designs, in more detail, for *horizontal scaling* to establish IRT scales that link different test forms dedicated to

one target population, as well as data-collection designs for *vertical scaling* to link test forms that cover a broad ability range, namely several school grades.

2.4.1.1 *Horizontal scaling*

Horizontal, or within-grade, scales link different test forms of comparable content and difficulty that are dedicated to the same population or school grade (Kolen & Brennan, 2014; Young, 2006). Generally, we can distinguish between three different designs for linking different test forms to one horizontal IRT scale: (1) *random group designs*; (2) *single group designs with counterbalancing*; and (3) *common-item nonequivalent group designs* (Kolen & Brennan, 2014, pp. 182–183). Random group designs' basic assumption is that the different groups that complete different test forms are randomly equivalent regarding their ability distributions. Under this condition, we can calibrate each form separately by constraining the mean ability to zero and the standard deviation to one for all groups. The resulting item parameters are directly compared without any further transformation or parameter fixation. In contrast, single group designs with counterbalancing imply that the same students answer multiple test forms while the order of the forms is balanced to prevent ordering effects. This design has the advantage of smaller sample-size requirements (Kolen & Brennan, 2014). However, it is only applicable if the number of test forms is manageable so that each student has enough time and motivation to answer all test forms. Item calibration usually is performed for all items together (i.e., concurrent calibration), directly resulting in a joint measurement scale for the different forms. Finally, the common-item nonequivalent group design allows for connecting test forms even if the target groups differ in ability. The link between the different test forms is established through common items (i.e., linking items) included in multiple test forms. Concurrent, separate, and fixed parameter calibrations are suitable for calibrating the resulting response data.

2.4.1.2 *Vertical scaling*

A vertical scale or developmental score scale refers to “an extended score scale that spans a series of grades and allows the estimation of student growth along a continuum” (Young, 2006, p. 469; see also Carlson, 2011; Kolen & Brennan, 2014; Tong & Kolen, 2007). A precondition for justifying a vertical scale is that the measured ability or competency is stimulated continuously and strengthens over time (Young, 2006). In contrast to horizontal scales, vertical scales combine test forms that vary in their mean difficulty, reflecting the broad ability range that needs to be covered when assessing abilities over several school grades. However, even though the different test forms refer to different difficulty levels, the underlying latent construct needs to remain constant from a content perspective to justify a unidimensional IRT scale.

The data-collection designs for establishing a vertical scale based on IRT methods resemble the ones for creating a horizontal IRT scale. Commonly, three different designs are distinguished, namely: (1) *equivalent group designs*; (2) *common item designs*; and (3) *scaling test designs* (Carlson, 2011; Kolen & Brennan, 2014; Young, 2006). Equivalent group designs

are very similar to random group designs. In particular, equivalent groups within a school grade are assigned randomly either to items related to their own grade, to the next lower grade, or to the next higher grade. Based on the assumption that the groups within one grade are equivalent regarding their ability distributions, it is possible to calibrate the items separately for each of the equivalent groups within one grade. However, further transformation or parameter fixation is needed to place the item parameters retrieved from the different grade groups in one common scale. Under the condition that all data are available at once, it also is possible to calibrate all items concurrently (Carlson, 2011; Kolen & Brennan, 2014).

Common item designs are very similar to common-item nonequivalent group designs. In contrast to equivalent group designs, students within one grade answer items dedicated not only to their own grade, but also to the next lower and next higher grades in common item designs. Linking items within the test forms of adjacent grade groups allow for establishing a link over all related school grades. Again, different calibration procedures are applicable. Separate calibration within each group produces item parameters that need to be transferred to a common scale through linear transformation. Alternatively, it is possible to calibrate all items together by assuming that the common items (i.e., linking items) have the same parameters independent of the school grade in which they are administered (Carlson, 2011). The same assumption is required for conducting fixed parameter calibration, in which item calibration starts within one school grade. Subsequently, the linking items' parameters are fixed to the first calibration's outcomes for calibrating the second school grade's response data, and for estimating non-common items' parameters. This procedure can be repeated from one school grade to the next.

Finally, scaling test designs resemble common item designs to the extent that students from different school grades are answering identical items. However, when applying a scaling test design, common items are shared not only between neighboring school grades. Instead, one block of items, namely the scaling test, is administered to all involved school grades (Kolen & Brennan, 2014; Young, 2006). Thus, the scaling test must include items covering a broad range of difficulties to provide suitable items at all ability levels. At the same time, the scaling test should have a reasonable length that allows for administering the whole test in one administration round. Besides the scaling test, students from each school grade answer items related to their specific grades. These grade-specific items are linked across grades through the scaling test. Concurrent, separate, or fixed parameter calibrations can be used to calibrate the items.

2.4.2 Targeted and Adaptive Testing for Efficient Parameter Estimation

2.4.2.1 Efficient parameter estimation under the Rasch model

In practical settings, it is often difficult to organize enough students who are willing to participate in calibration studies and secure sufficient testing time for administering a

satisfactory number of items to assess students' ability accurately. Therefore, a need exists for calibration and test designs that deal with these limited resources efficiently. The accuracy of item difficulty and student ability estimation depends on their relationship within the calibration or test sample. In particular, item difficulty and student ability can be estimated most accurately under the Rasch model if students' ability is close to the items' difficulty (Berger, 1991; Rost, 2004; van der Linden, 1988). Thus, assigning items that fit students' ability levels is important not only from a motivational (e.g., Asseburg & Frey, 2013; Wainer, 2000b) and content-related perspective, but also from a psychometric perspective. As stated by Lord (1980), "an examinee is measured most effectively when the test items are neither too difficult nor too easy for him" (p. 150). Test forms targeted to students' abilities increase measurement efficiency because they provide more information about students' abilities with a given number of items per student than a general overall test (Lord, 1971a, 1971b, 1971c, 1980). Simultaneously, assignment of targeted items also increases the information that students provide about the items, thereby enhancing the efficiency of item difficulty estimation during calibration based on a given number of students (Berger, 1991; Eggen & Verhelst, 2011; Stocking, 1988; Wright, 1977). Generally, two strategies can be used to align item difficulty and student ability: targeted testing, based on ability-related background variables, and adaptive testing, based on performance during test taking.

2.4.2.2 Increasing efficiency through ability-related background variables

The three designs for vertical scaling (i.e., equivalent group, common item, and scaling test design) use school grade or age as a criterion for dividing the target population into different ability groups, and for assigning different test forms accordingly. If the scale spans more than three school or age groups, we assume that the match between item difficulty and student ability is better in equivalent group designs and common item designs, in which each item is administered to a maximum of three different ability groups, than in scaling test designs, in which the scaling test is administered to students from all school grades. Thus, we hypothesize that equivalent group designs and common item designs generally are more efficient than scaling test designs for establishing scales that span more than three school years. The efficiency of common-item nonequivalent group designs, common item designs, and the grade-specific parts of scaling test designs could be increased further by using additional ability-related background variables within grades, such as exam grades or performance-related school types, to improve the alignment of student ability and item difficulty within each grade group (Mislevy & Wu, 1996). In contrast, random group designs, single group designs with counterbalancing, or equivalent group designs do not allow for increasing measurement efficiency by further dividing the samples within grades because these designs explicitly require comparable ability groups for linking within one school grade.

The efficiency gain from targeted testing depends on the correlation of the ability-related background variable with students' true ability—namely, its power to divide the target population into subsamples with abilities that are as homogenous as possible (Pohl, 2013). A

weak ability-related background variable could result in considerable overlap between different subsamples' ability distributions (e.g., Baumert, Stanat, & Watermann, 2006a; College Board, 2017; Moser, Buff, Angelone, & Hollenweger, 2011), which might impair the efficiency of item calibration or ability estimation. Furthermore, the efficiency gain also might depend on the different targeted test forms' quality. In particular, if the items are not yet calibrated and, thus, the empirical item difficulties are not yet known (as in most calibration studies), test developers might under- or overestimate some items' difficulty (e.g., Bejar, 1983; Hambleton & Jirka, 2006; Sydorenko, 2011; Wauters, Desmet, & van den Noortgate, 2012), consequently placing them in either too difficult or too easy test forms. Unfortunately, extant studies that have investigated the impact of using potentially inaccurate ability-related background variables or test forms with misplaced items on the efficiency of targeted data-collection designs are lacking, to our knowledge.

2.4.2.3 Increasing efficiency through performance-based assignment

A different approach for increasing measurement efficiency is to use students' performance during test taking to assign students adaptively to the most informative test form, as in MSTs (Hendrickson, 2007; Yan et al., 2014; Zenisky et al., 2010), or even to single items, as in CATs (van der Linden & Glas, 2010; Wainer, 2000a). CATs' highly customized item administration offers the advantage of measuring close to the student's ability, thereby increasing measurement accuracy and efficiency compared with linear tests with the same length (e.g., Lord, 1980; van der Linden & Glas, 2010; Wainer, 2000a; Weiss, 1982). Extant research on MSTs' efficiency has shown that MSTs also increase efficiency considerably compared with linear tests, even though they are slightly less efficient than CATs (for a general overview, see Hendrickson, 2007; Yan et al., 2014; Zenisky et al., 2010). Furthermore, we also assume that CATs and MSTs are more accurate and efficient than a set of linear test forms targeted based on ability-related background variables, even though we are unaware of any related empirical studies.

On the downside, algorithms in CATs—and those in MSTs to a certain extent—only can select informative items and automatically estimate students' ability if the item parameters are known (Wainer & Mislevy, 2000). Thus, the development of CATs and MSTs usually includes a calibration phase in which linear tests are used to collect information systematically about the items (Thompson & Weiss, 2011). Nevertheless, an alignment between item difficulty and student ability based on performance also could be advantageous in securing efficient item difficulty estimation (Eggen & Verhelst, 2011; Glas & Geerlings, 2009; Zwitser & Maris, 2015). However, the challenge in this context is to determine preliminary item difficulties before item calibration, which can serve as a basis for matching items and students adaptively. In particular, it is desirable to control the number of observations per item in calibration designs such that all items' difficulty parameters can be estimated with comparable accuracy (Glas & Geerlings, 2009). The lack of knowledge about item difficulty before calibration has been considered in a few studies related to CATs (e.g., Ali & Chang, 2014; Fink, Born, Spoden, & Frey, 2018; Kingsbury, 2009; Makransky & Glas, 2010), but we are unaware of any studies

related to MSTs. Furthermore, the studies mainly focused on the calibration of a limited number of field-test items during the administration of operational CATs, not on the calibration of a new item pool.

In sum, adaptive (i.e., performance-based) assignment of items to students through CATs or MSTs increases the efficiency of ability estimation if a calibrated item pool is available. Furthermore, it could increase the efficiency of item calibration under certain conditions. In particular, the efficiency of common-item nonequivalent group designs, common item designs, and scaling test designs could be increased by performance-based item assignment within school grades under the condition that we have accurate preliminary estimates of item difficulties at hand. Such preliminary estimates are required to guarantee a good match between item difficulty and student ability. In contrast, adaptive item assignment is incompatible with the precondition of equal ability groups of random group designs, single group designs with counterbalancing, or equivalent group designs.

2.5 Concept for Implementing a Common Vertical Scale for Mathematics

2.5.1 Establishing a Scale with Four Calibration Steps

In this section, we propose a concept for the practical implementation of a common vertical Rasch scale for the standardized tests and the online item bank for formative assessment to assess the mathematics abilities of students from Northwestern Switzerland. This concept is based on our analysis of the similarities and differences between the two instruments, the specific features of IRT and the Rasch model as a measurement approach, and related data-collection designs. In addition, two practical constraints are especially relevant for our concept. First, government authorities from Northwestern Switzerland specified a roadmap for gradually implementing the four standardized tests. In particular, the four standardized tests must be implemented one per year in a predefined order starting with the third-grade test, followed by the sixth-grade test, eighth-grade test, and ninth-grade test. Simultaneously, the item pool of the online item bank for formative assessment must be developed for all seven target grades. Furthermore, all four standardized tests' content needs to be refreshed partially after each administration to ensure item quality and confidentiality.

Second, the target population of approximately 13,000 students per school grade is rather small compared with the large amount of items required for the two instruments. The final item pool should include more than 10,000 items to cover seven school years with two different instruments. Thus, we would need to administer more than 20 items to each of the 91,000 students from the seven school grades to collect 200 observations for each of the 10,000 items. However, using the online item bank is optional for schools and their students, which means that we must count on their voluntary participation for any related calibration studies. Based on our experience, we expect that less than 10 percent of the target population would participate voluntarily in such studies, which would increase the number of items per student

drastically to more than 200. Consequently, it is not feasible to include all items in a single calibration study. Simultaneously, the standardized tests are compulsory, but limited to four of the seven target school grades and separated by up to three school years. In addition, testing time is limited to two school hours, which automatically limits the number of items that we can administer to each student. Based on these preconditions, directly linking the four standardized tests to establish a vertical scale would decrease the tests' efficiency for estimating item difficulty and student ability significantly. For example, if we were to include sixth-grade items on the third-grade test, the sixth-grade items would be very difficult for most third-grade students to answer. Therefore, the third-grade students would provide very limited information about the difficulty of the sixth-grade items, and the sixth-grade items would provide very limited information about third-grade students' abilities.

Figure 2.2. Calibration steps for establishing a vertical scale and for linking the standardized tests and online item bank for formative assessment. (1) Calibrating the standardized third-grade test; (2) calibration assessments for establishing the vertical scale; (3) linking the standardized tests for sixth, eighth, and ninth grades; (4) extending the online item bank for formative assessment.

Against this backdrop, we propose developing a vertical scale in four dedicated calibration steps that are summarized in Figure 2.2. For step 1, we suggest establishing a horizontal third-grade scale based on the standardized third-grade test through a common-item nonequivalent group design with eight linked test forms. The resulting calibrated third-grade item pool serves as a basis for future standardized third-grade tests and for extending the third-grade scale to a vertical scale. For step 2, we recommend administering dedicated calibration assessments within the online item bank to extend the scale from third through ninth grades and link the two instruments. These assessments comprise test forms targeted to specific grades with linking items to adjacent school grades. Furthermore, we suggest including items

dedicated to the standardized tests in these calibration assessments to establish the link between the four standardized tests and between the two instruments. Thus, the calibration assessments are the central element of the linking process. Step 3's objective is to relate the three additional standardized tests (i.e., the tests for sixth, eighth, and ninth grades) to the vertical scale. Finally, in step 4, all remaining items included in the calibration assessments serve as a basis for further extending the calibrated item pool of the online item bank through online calibration and for administering CATs. In the following sections, we describe and justify the first three steps in more detail and provide a brief overview of the fourth step.

2.5.2 Step 1: Calibrating the Standardized Third-Grade Test

We propose using the standardized third-grade test as a starting point for establishing the scale because the standardized tests generally offer several advantages compared with the online item bank for formative assessment. First, the standardized tests are compulsory, guaranteeing a large sample size. Second, administration of standardized tests does not depend necessarily on calibrated items. Content experts assemble the two paper-based tests, as well as the two MSTs. Thanks to their expertise, experts roughly can rate the items' difficulty (e.g., Bejar, 1983; Hambleton & Jirka, 2006; Sydorenko, 2011; Wauters et al., 2012) and assign the items accordingly to different grade-specific test forms or to different modules of the MSTs. Third, the centralized administration, somewhat higher stakes, and especially the paper-based test administration used with primary-school students justify more processing time between test administration and the release of assessment results, allowing for an in-depth item analysis to ensure the test results' quality. Moreover, the standardized tests are administered under well-defined, standardized conditions that enhance the validity of the resulting item difficulty parameters for measuring students' true ability in mathematics. Thus, all four standardized tests would be suitable for starting the scale. In our particular practical situation, we propose using the standardized third-grade test as a starting point because this test is the first one that should be implemented, according to the roadmap.

To benefit from the large sample of students who must take the test, and to get the maximum use out of the limited testing time of two school lessons, we suggest developing multiple parallel test forms for the first administration of the standardized third-grade test and distributing them randomly across classes. This strategy allows us to calibrate a wide selection of items as a basis for developing future third-grade tests and for linking the third-grade test with the online item bank. In detail, we propose a common-item nonequivalent group design with eight parallel test forms (i.e., test forms with equally difficult items), whereas each item is included in two test forms for linking purposes. We prefer this design to the two other designs for horizontal scaling because it allows for an economical handling of testing time per student. A single group design with counterbalancing is not applicable in our context because it requires that all students answer all items or test forms. Given that extending testing time is not an option, we could administer only one test form and, thus, only a very limited selection of third-grade

items with such a design. Theoretically, a random group design would allow for administering test forms without any overlap and, thus, for administering more items than a common-item nonequivalent group design. However, we did not consider this option because it is difficult in educational measurement to ensure that groups have equal abilities (Keizer-Mittelhaeuser, 2014), which is the random group design's basic requirement.

As described earlier, common-item nonequivalent group designs theoretically allow for increasing the efficiency of item calibration and ability estimation through targeted and adaptive testing. However, we decided against following this option in our concept. First, we have no access to any ability-related student variables prior to test administration that we could use as a criterion for the targeted assignment of test versions that vary in difficulty. It would be possible to ask teachers to select the most suitable test form for each of their students, but in such a scenario, we would lose control of the number of observations per test form and risk failing to provide reports for selected test forms in which we have too few observations for calibrating related items. Second, an adaptive assignment of different test forms or test parts (i.e., an MST design) would increase the complexity of the distribution and administration of paper-based test booklets significantly. Finally, we expect only a very marginal efficiency gain through targeted or adaptive testing for item calibration, given the large sample size.

Given that the eight linked test forms are administered during the same period, we suggest calibrating all test forms concurrently. Concurrent calibration processes all available data in one calibration run and directly results in one horizontal scale for representing third-grade students' abilities and items' difficulties. Thus, it is more efficient and stable than separate calibration and fixed parameter calibration, which estimate item difficulties based on single test forms, i.e., limited information (Briggs & Weeks, 2009; Kolen & Brennan, 2014). Items that do not fit the underlying Rasch model (e.g., items with divergent parameters across test forms or items representing different dimensions) need to be excluded before ability estimation to ensure the quality of student reports.

2.5.3 Step 2: Calibration Assessments for Establishing the Vertical Scale

Additional linking steps are needed to expand the horizontal third-grade scale into a vertical scale and establish a link between the standardized tests and the online item bank. Furthermore, we require a calibrated item pool to offer CATs within the online item bank for formative assessment. To develop the vertical scale, we propose including predefined, computer-based calibration assessments in the first version of the online item bank for formative assessment. Only the online item bank allows us to reach students from all seven target school grades. Furthermore, most of the items are dedicated to the online item bank. Consequently, it is reasonable to calibrate these items under related measurement conditions. Last, but not least, a computer-based administration through the online item bank is much less complicated and cheaper than administering paper-based calibration assessments.

Like the calibration of the standardized third-grade test, we advise using a common item design as a basis for the calibration assessments. Compared with other data-collection designs for vertical scaling, a common item design does not require equal groups within school grades, facilitates the administration of grade-specific items, and foresees administering linking items to adjacent school grades only. Equivalent group designs rely on equal groups, which are difficult to achieve when participation in calibration assessments is optional, thereby lowering the stakes for students and teachers (Keizer-Mittelhaeuser, 2014). Besides, we agree with Carlson (2011) that scaling test designs are less suitable for educational assessments because they require administering a wide range of items, whereas most items are too easy or too difficult for a specific school grade. Furthermore, a large deviation between item difficulty and student ability is inefficient from a measurement perspective and might frustrate students (Young, 2006).

Table 2.1. Common Item Design for Vertical Scaling Through Calibration Assessments on Macro Level.

Item blocks	Samples per school grade						
	3	4	5	6	7	8	9
3	ST3						
3–4							
4							
4–5							
5							
5–6							
6				ST6			
6–7							
7							
7–8							
8						ST8	
8–9							
9							ST9

Note. Dark blocks refer to grade-specific items, light blocks refer to linking items between adjacent grade groups. ST3, ST6, ST8, and ST9 represent linking items to the standardized tests for third, sixth, eighth, and ninth grades, respectively.

Table 2.1 provides a macro-level overview of the proposed common item design. In detail, we suggest that content experts develop dedicated test forms for each of the seven target school grades that generally include two different types of items: (1) grade-specific items targeted to students' current school grade, and (2) linking items targeted to the next lower and next higher grades. Furthermore, we recommend substituting some grade-specific items in the

calibration assessments for the third, sixth, eighth, and ninth grades by linking items to the related standardized tests.

Using the same items for both instruments is possible because they both share the same content framework. Thus, it is important that the linking items cover all relevant content domains and competencies (Kolen, 2007; Kolen & Brennan, 2014). In particular, the curriculum, *Lehrplan 21*, calls for a continuous development of 26 competencies within mathematics (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016a) and specifies multiple consecutive competence levels for each competency over the course of compulsory school. This means that all 26 competencies are relevant in each school grade and need to be reflected in the composition of the standardized tests and the calibration assessments, especially in the selection of the linking items. Furthermore, it is important to consider potential mode effects and limitations in item formats when selecting linking items for the third and sixth grades (e.g., Flaugher, 2000; Kröhne & Martens, 2011; Robitzsch et al., 2017). For these two grades, the linking items need to be produced as paper-based and computer-based versions. To serve as linking items, their difficulty must be comparable between the two administration modes. Furthermore, the item formats applicable for these linking items must allow for representing all relevant competencies within the mathematics curriculum.

To calibrate a maximum number of items through calibration assessments, we recommend combining the common item design with a common-item nonequivalent group design. Namely, we suggest developing multiple linked test forms for each school grade, whereas grade-specific items link between two test forms within one grade, and linking items link test forms across adjacent school grades. Consequently, grade-specific items are included in two test forms per school grade, and linking items are included in two test forms targeted to two different school grades. However, it is important to adjust the number of items included in the design and, thus, the complexity of the design to the available calibration sample (Thompson & Weiss, 2011). Items are only usable for CATs if their item parameters are estimated based on a satisfactory amount of observations (e.g., Thompson & Weiss, 2011; Wainer & Mislevy, 2000). Because students and classes are not obliged to participate in calibration assessments, we suggest developing only five parallel test forms per school year (i.e., 35 test forms in total) with a moderate test length of 32 mathematics items answerable in one school lesson. Within each school grade, we recommend distributing the forms randomly across all participating students to ensure an equal number of observations per test form. This design allows for calibrating up to 560 mathematics items across all seven school grades. Furthermore, a participation rate of 5 percent of the total population (i.e., $N = 4,375$) would be sufficient to collect 250 observations per item, under the condition that the sample is distributed evenly over all seven school grades.

Given the low expected participation rate, adaptive calibration designs, namely multistage calibration designs, could be an option for increasing the efficiency of item difficulty estimation. Multistage calibration designs would allow us to increase measurement efficiency

based on students' performance. Thus, they do not require preliminary information about students' ability levels to select the optimal test form. However, we do not consider this option in our concept because no previous research has provided enough information about multistage calibration designs' expected efficiency gains and related risks of efficiency loss due to potential misspecifications of the design. We know from IRT that we could increase item-calibration efficiency only if we ensure equal numbers of observations per item within a school grade (cf. Glas & Geerlings, 2009). To develop multistage calibration tests accordingly, we would require substantial knowledge about the items' difficulty, as well as about the target population's ability distribution. The fact that this knowledge is limited in our situation could complicate the definition of routing rules that ensure an equal number of observations per module and item. Given the expected low number of participating students, it is crucial to distribute these students evenly across different test forms within school grades. In sum, we rate the risk of efficiency loss due to an unbalanced number of observations per item as higher than a potential efficiency gain through adaptive item assignment.

Regarding item calibration, we propose a combination of concurrent and fixed parameter calibration. To establish the vertical scale, we advise calibrating all 35 test forms concurrently while specifying separate population distributions for each of the seven school grades (DeMars, 2002; Eggen & Verhelst, 2011). Under the condition that the data represent a unidimensional construct and fit the Rasch model, this procedure is the most economical and accurate for determining item difficulty parameters (Kolen & Brennan, 2014). To verify these assumptions, we recommend additionally following a separate calibration procedure with equating to establish the vertical scale. Differences between concurrent and separate calibration procedures would indicate potential problems in calibration procedures, such as multidimensionality (e.g., Hanson & Béguin, 2002). Furthermore, separate calibration allows for directly comparing item parameters between grades to detect potential parameter variance. Items with varied difficulties between grades need to be excluded from the linking-item set.

To link the vertical scale to the horizontal third-grade scale, we suggest fixing linking items between the two scales (i.e., the third-grade items included in the standardized third-grade test and the third-grade calibration assessments) based on the calibration of the standardized third-grade test while concurrently calibrating the 35 forms of the calibration assessments. A fully concurrent calibration is not applicable because the third-grade test already would be operational at this stage. The advantage of fixed parameter calibration compared with separate calibration through equating is that the parameters of the new items (i.e., the items from the calibration assessments) are placed directly on the same scale as the items from the standardized third-grade test (Kim, 2006). In particular, the fixed items serve as anchors for adjusting the vertical scale to the established third-grade scale (e.g., Keller & Hambleton, 2013; Keller & Keller, 2011; Kim, 2006), thereby establishing the first link between the two assessment instruments.

2.5.4 Step 3: Linking the Standardized Tests for Sixth, Eighth, and Ninth Grades

For the third calibration step, we propose linking the standardized tests for the sixth, eighth, and ninth grades to the vertical scale developed based on the calibration assessments. Because the standardized sixth-grade test is very similar to the standardized third-grade test in terms of assessment type and purpose, content specifications, and measurement conditions, we suggest a test design for the sixth-grade test that strongly resembles the standardized third-grade test. In particular, we recommend developing a common-item nonequivalent group design with eight linked parallel test forms and assigning these test forms randomly across all sixth-grade classes. This test design allows for benefitting from the large sample of sixth-grade students by administering and calibrating a large amount of items that can be used to develop future sixth-grade tests. Furthermore, we can include linking items with the calibration assessments in these test forms to link the standardized sixth-grade test with the vertical scale. These linking items already have empirical item difficulties from step 2 of the calibration process. Consequently, we advise fixing these items while concurrently calibrating the additional sixth-grade items included in the eight parallel test forms. Furthermore, we recommend checking parameter invariance between the standardized test and calibration assessments, and performing in-depth item analysis. Items with variant parameters need to be excluded from linking, and items that do not fit the underlying Rasch model need to be excluded completely prior to ability estimation to ensure the quality of student reports.

In contrast to the paper-based linear third- and sixth-grade tests, the standardized tests for the eighth and ninth grades are conceptualized as computer-based MSTs. Both standardized MSTs include four test parts (i.e., stages). The first stage contains a starting module of intermediate difficulty as a basis for preliminary ability estimation, and three subsequent stages include five modules, each differing in difficulty. In stages 2 to 4, students are routed based on their preliminary performance to one of the five modules. In total, the two MSTs include 20 different modules each, and each participating student sees four of them. Because the two standardized tests for secondary school differ only in their target population (i.e., eighth vs. ninth grade), we propose identical designs and calibration approaches for both tests.

Given that the MST design itself already is rather complex, we refrain from suggesting parallel forms for the MSTs. However, to balance the number of observations per item between the first and subsequent stages, and to prevent students from cheating by copying answers from their neighbors, we foresee developing five different parallel versions of the starting module for each test, then assigning these versions randomly across students. As an alternative to developing parallel starting modules, it also would be possible to develop targeted starting modules for each of the three performance-related school types distinguished on the secondary school level in Northwestern Switzerland. Each student is assigned to one of these three school types, and this information is available before test administration. Related, targeted starting modules could optimize the efficiency of the MST designs for item calibration based on the first administration, as well as for ability estimation in general. However, several studies have

found heavy overlap between the abilities of students assigned to different secondary school types (e.g., Angelone, Keller, & Moser, 2013; Baumert, Stanat, & Watermann, 2006b). Based on this research, we cannot rule out that students with abilities that differ from their group's mean ability could be disadvantaged by such a design. Furthermore, we could not find any research on the expected efficiency gain through a targeted assignment of starting modules in an MST design for item calibration and ability estimation. Consequently, we propose refraining from distinguishing different difficulty levels during the first stage of standardized eighth- and ninth-grade tests.

Usually, a calibrated item pool is available for distributing items within an MST design and for defining related routing rules. However, due to limited time and financial resources, as well as related limited options for running calibration studies, we propose using the first administration of the standardized MSTs for the eighth and ninth grades to calibrate the items. In contrast to the calibration assessments, for which we refrain from using an MST design due to limited knowledge about item difficulty and students' ability distribution within the target school grades, the outcome of the calibration assessments (i.e., step 2) provide a basis for developing the MSTs for the eighth and ninth grades. Namely, the calibration assessments' outcomes will provide an indication of eighth- and ninth-grade populations' ability distributions, and the calibrated eighth- and ninth-grade items from the calibration assessments can serve as references for adjusting content experts' difficulty ratings for the eighth- and ninth-grade items. Moreover, we will have empirical item difficulty estimates for linking items between the standardized tests and calibration assessments. Finally, the expected sample size is much larger for the compulsory standardized eighth- and ninth-grade tests than for the calibration assessments.

Against this backdrop, we suggest developing the two standardized MSTs for the eighth and ninth grades based on content experts' expertise. For each of the two tests, experts need to develop five parallel test forms for the first stage and five test forms targeted to five different difficulty levels for stages 2 to 4. Linking items between the different modules are not required for item calibration. Instead, we can link the different modules within one MST through overlapping paths (i.e., various combinations of different modules from the four stages). The link between the two MSTs can be established by linking them to the vertical scale of the calibration assessments, specifically through the linking items between each standardized test and the calibration assessments. We also suggest asking content experts to determine rules to route students based on their raw scores from one stage to the next. Therefore, it is important to define routing rules that guide a comparable number of students to all five modules within a stage. A low number of observations would impair the calibration of related items. Thanks to the large sample size (i.e., 13,000 students per test), we rate the risk of having an insufficient number of observations as being very small. Ideally, approximately 2,600 students would reach each module. Nevertheless, only 10 percent of this sample would be sufficient for accurately estimating item difficulty parameters based on the Rasch model (Wright, 1977).

For calibrating the items, we again recommend a procedure similar to that of the standardized third- and sixth-grade tests. Specifically, for each of the two standardized tests, we propose calibrating all modules of the MST concurrently while fixing the difficulty parameters of the linking items to the outcomes of the calibration assessments. Furthermore, we advise investigating parameter invariance between the standardized tests and calibration assessments, getting a close look at item fit statistics during item analysis, and excluding misfitting items prior to ability estimation and reporting. Even though we expect some efficiency loss due to limited knowledge about true item difficulty parameters compared with MSTs, which are developed based on calibrated item pools, we still expect an efficiency gain for item calibration and ability estimation compared with linear tests. Furthermore, measurement efficiency is limited only during the first administration of the test. In subsequent years, when we have a calibrated item pool at hand, we can improve item-difficulty consistency within the MST modules and adjust the routing rules if needed. However, the detailed interactions between limited knowledge about item difficulty parameters during test construction, consequential specification errors in test modules, and measurement efficiency are subject to further research.

2.5.5 Step 4: Extending the Online Item Bank for Formative Assessment

For the final calibration step, we propose using an online calibration algorithm to develop the calibrated item pool further (Verschoor & Berger, 2015; see also Ban, Hanson, Wang, Yi, & Harris, 2001; Fink et al., 2018; Makransky & Glas, 2010; Stocking, 1988; Thompson & Weiss, 2011; Wainer & Mislevy, 2000). The detailed specifications of this online calibration procedure are beyond this paper's scope, but we want to provide a general overview of the procedure to complete the concept for calibrating and linking the two instruments.

Several hundred items from the calibration assessments (i.e., all items except the linking items to the standardized tests) serve as the basis for the item pool of the online item bank for formative assessment. However, the online item bank requires several thousand items to provide teachers and students with a broad range of assessments suitable for assessing students from different school grades and ability levels repeatedly. Given that the online item bank serves as a formative instrument with a low-stakes character, we suggest administering the calibrated items together with a large proportion of uncalibrated items, as well as calibrating new items gradually after each assessment. In doing so, the calibrated items can serve as anchors to link the new items to the vertical scale and as references to prevent bias in item difficulty estimates of new items. In addition, we propose exporting response data from time to time for offline calibration to perform item analysis. Thanks to this online calibration strategy, we can start the online item bank for formative assessment after a brief offline calibration phase based on calibration assessments (i.e., step 3) while students and teachers engage with the system.

2.6 Discussion and Further Research

In this paper, we introduced a set of four standardized tests and an online item bank for formative assessment as two different assessment instruments for measuring students' ability repeatedly across compulsory school (Bildungsraum Nordwestschweiz, 2012; Tomasik et al., 2018). Both instruments intend to provide objective reports reflecting individual students' current ability levels, are based on the same content specification, namely the curriculum, *Lehrplan 21* (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014), and both are dedicated to the same population of third- through ninth-grade students in Northwestern Switzerland. These similarities between the two instruments elicited the need for a common underlying reporting scale that allows for comparing test results independent of the assessment instrument or assessment occasion, and for tracking students' progress over the course of their school careers. Moreover, these similarities are the basis for justifying linking both instruments through IRT-related horizontal and vertical scaling methods.

Considering the similarities and differences between the instruments, as well as additional practical constraints, we proposed a concept that included four calibration steps to establish a vertical Rasch-based scale for measuring students' mathematics ability from third through ninth grade, and for linking the two instruments. More specifically, we suggested: (1) starting with a horizontal scale for the standardized third-grade test; (2) extending this scale to a vertical scale through dedicated calibration assessments; (3) linking the standardized tests for the sixth, eighth, and ninth grades to the vertical scale; and (4) calibrating and linking additional items to the scale through online calibration while students engage with the online item bank for formative assessment. While developing this concept, we came across several challenges and interesting research questions related to the implementation of concrete data-collection designs and calibration procedures in our specific practical context. On one hand, differences between the two instruments might complicate the implementation of our concept and impose a careful validation of the final vertical scale from a content perspective. On the other hand, empirical studies should investigate how targeted and adaptive testing could increase calibration efficiency and test designs if the size of the expected calibration sample and testing time are limited. Furthermore, these studies also should evaluate potential risks related to the implementation of such designs in practical contexts.

2.6.1 Practical Challenges to Implementing the Proposed Concept

The first challenge to successful implementation of our concept is that students might differ significantly in their motivation to make an effort in answering items in various assessments that serve as a basis for item calibration. Differences in test-taking motivation might result from differences between the two instruments in degree of obligation, degree of standardization of test administration, and primary assessment purpose. Generally, assessments related to both instruments exert no direct consequences for teachers or students. However, the standardized tests are official compulsory tests that are centrally organized and administered, following clear

administration guidelines. Thus, we assume that teachers and students are motivated to provide a solid performance as far as effort. In contrast, we hypothesize that at least some students could be less motivated to provide a good performance while voluntarily engaging with the online item bank for formative assessment and the calibration assessments (e.g., Mittelhaeuser et al., 2013; Wise & DeMars, 2005). Students with low test-taking motivation might perform at levels that are lower than their actual abilities. Presumably, they guess correct answers or skip some items. Such behavior might influence the linking process negatively as reported by Mittelhaeuser et al. (2011). To ensure the vertical scale's validity, it is important to consider students' test-taking motivation and analyze the impact of potential differences on the estimation of item difficulty parameters and the linking between the two instruments and different school grades.

A second challenge refers to differences in administration modes on the primary-school level. Specifically, the standardized third- and sixth-grade tests are conceptualized as paper-based tests, while the two standardized tests for secondary school and all assessments related to the online item bank, including the calibration assessments, are administered through computers. Thus, we need to transform linking items from one mode into the other to establish a common scale. This requirement might restrict suitable item formats for the linking items, which might, in turn, conflict with the requirement that the linking items should cover all relevant domains and competencies described in the content specifications. Furthermore, the transformation might distort the scale if one mode facilitates or hinders answering an item correctly compared with the other mode (e.g., Flaugher, 2000; Kröhne & Martens, 2011; Robitzsch et al., 2017; Wainer & Mislevy, 2000). Therefore, it is essential to analyze the item difficulty parameters' invariance between the two administration modes carefully when linking the two paper-based standardized tests to the computer-based calibration assessments. Items with divergent parameters are not suitable as linking items, but when excluding linking items, we need to ensure that item exclusion is unrelated to particular domains or competencies within mathematics so that the remaining linking items are still covering all relevant competencies, thereby representing underlying content specifications.

Third, extant empirical studies are lacking, i.e., research that would provide clear recommendations for selecting the most appropriate calibration procedure (i.e., concurrent, separate, or fixed parameter calibration) in specific practical settings, especially for developing vertical scales. Instead, existing studies have provided mixed results indicating that the most suitable calibration strategy might depend on various factors and their interactions (e.g., Briggs & Weeks, 2009; Dadey & Briggs, 2012; Kolen & Brennan, 2014). Consequently—as suggested by Hanson and Béguin (2002) and Kolen and Brennan (2014)—we intend to apply concurrent and separate calibration simultaneously to establish the vertical scale based on calibration assessments. Differences between the two procedures might indicate calibration issues, such as multidimensionality, that require changing the IRT model or content specifications.

Finally, even if we can consider potential impacts from differences in students' test-taking motivation and administration mode in our psychometric model, and even if the different calibration procedures yield comparable outcomes, no guarantee exists that the resulting vertical scale represents the intended latent mathematics ability. Establishing the vertical scale and ensuring that it meets psychometric assumptions, such as parameter invariance of linking items and item fit, are only the first step toward a valid vertical scale. We argue that the calibration decisions and resulting vertical scale also should be validated from a content perspective through an external content-related validation criterion (Briggs & Weeks, 2009; Dadey & Briggs, 2012; Harris, 2007; Ito et al., 2008; Tong & Kolen, 2007). Namely, we need to investigate, in follow-up studies, whether the vertical scale represents mathematics ability and its development across school grades as described in the content specifications (i.e., the curriculum, *Lehrplan 21*).

2.6.2 Research on Increasing Measurement Efficiency Under Practical Constraints

A very practical challenge to developing our concept for establishing the vertical mathematics scale was that the expected calibration sample size and available testing time per student are limited compared with the huge number of items requiring calibration. In theory, data-collection designs' efficiency can be increased under the Rasch model by improving the match between item difficulty and student ability through targeted and adaptive testing (Berger, 1991; Rost, 2004; van der Linden, 1988). However, optimizing this match might include some complications and risks in practical situations.

In theory, it would be possible to use an MST approach to improve the efficiency of item difficulty estimation within school grades in our common item design for calibration assessments. More specifically, we could route students after the first general test module, based on their performance, to either an easy or more difficult second module to increase the match between item difficulty and student ability. However, in practice, we lack important knowledge for concretizing this idea. First, we have no indication of how much efficiency we could gain by including performance-based routing in our common item design for vertical scaling because we could not find any related empirical studies. Second, we have only limited knowledge about the true difficulty of the items prior to item calibration. We assume that content experts roughly could estimate items' difficulty. Nevertheless, we expect that this somewhat-limited knowledge about the true difficulty could lead to misplaced items in some test modules (i.e., too easy or too difficult items, given the module's target difficulty) and to inaccurate routing rules that might result in unequal sample sizes for subsequent modules. Third, we are unaware of any empirical studies that have analyzed how potential misplaced items and inaccurate routing rules would affect MST designs' efficiency in item calibration, and whether such designs even involve the risk of an efficiency loss compared with common item designs without performance-based routing. In sum, studies are required that weight expected efficiency gains

through performance-based item assignments in calibration designs against potential risks of suboptimal test forms or routing rules to justify implementation of such designs in practice.

In addition, we also lack empirical evidence to justify targeted testing in the first stage of the standardized MSTs for eighth and ninth grades. Theoretically, we could assign secondary-school students based on their performance-related school type to different starting modules. Under the condition that the performance-related school type is an accurate indicator of students' true ability, this approach would increase the match between item difficulty and student ability during the first stage of the MST, thereby simultaneously increasing the efficiency of item calibration and ability estimation. However, MST designs with targeted modules during the first stage are very rare in practice (Hendrickson, 2007; Zenisky & Hambleton, 2014), and we could not find any empirical studies that have compared their efficiency with that of traditional MSTs with one general starting module. Furthermore, we question, to some extent, the suitability of the performance-related school type as an indicator of students' true ability, given that several studies have found significant overlap between ability distributions related to different school types on secondary-school level (e.g., Angelone et al., 2013; Baumert et al., 2006b). Consequently, we did not consider implementing MSTs with targeted starting modules in our concept. Moreover, we cannot recommend such designs in practice until empirical studies have provided insights into their general efficacy gain, the extent to which the efficiency gain depends on the correlation between the ability-related background variable used for assigning the targeted starting modules and students' true ability, and the extent to which students with abilities diverging from the mean ability of their ability group might be disadvantaged by inappropriate starting modules.

2.6.3 Conclusion

This paper was motivated by the objective to link a set of four standardized tests and an online item bank for formative assessments through a common vertical measurement scale. The extant literature on IRT and related data-collection designs and calibration procedures offer powerful methodologies for implementing such a vertical scale. However, by developing a concept for implementing a vertical scale to assess students' mathematics abilities across seven school grades, we showed that in practice, these methodologies need to be reconciled carefully with various practical constraints. Therefore, we pointed out the importance of validating decisions regarding calibration and linking procedures not only from a psychometric perspective through item analysis or investigation of parameter invariance, but also from a content perspective through external content-related validation criteria. Furthermore, we emphasized the need for further research on the efficiency of targeted and adaptive designs for item calibration and ability estimation in practical contexts with limited knowledge about the true difficulty of the items and the suitability of potential ability-related background variables. Empirical evidence for an efficiency gain of targeted and adaptive designs under such practical constraints would justify their implementation in practice and help deal with limited student samples or testing

time efficiently. Despite the related challenges, we are convinced that a common reporting scale for the two instruments and for the seven target school grades will enhance the assessment results' validity considerably by facilitating the interpretation of the results for students, teachers, and other stakeholders, and by offering them a unique chance to monitor students' learning trajectories across compulsory school.

2.7 References

- Ali, U. S., & Chang, H.-H. (2014). *An item-driven adaptive design for calibrating pretest items* (ETS Research Reports Series No. RR-14-38). Princeton, NJ: Educational Testing Service.
- Angelone, D., Keller, F., & Moser, U. (2013). *Entwicklung schulischer Leistungen während der obligatorischen Schulzeit: Bericht zur vierten Zürcher Lernstandserhebung zuhanden der Bildungsdirektion des Kantons Zürich* [Development of school performance during compulsory school: Report on the fourth assessment for the attention of the Zurich department of education]. Zürich, CH: Institut für Bildungsevaluation (IBE). Retrieved from Institut für Bildungsevaluation (IBE) website: http://www.ibe.uzh.ch/projekte/lezh/Lernstandserhebung_9KlasseZH_Bericht.pdf
- Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, 55(1), 92–104.
- Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38(3), 191–212. doi:10.1111/j.1745-3984.2001.tb01123.x
- Bättig, M., Gut, M., & Schwab, S. (2011). *Bericht der Evaluation über das stufenübergreifende ICT-Entwicklungskonzept für die Schulen des Kantons Solothurn 2008*. Solothurn: Fachhochschule Nordwestschweiz, Pädagogische Hochschule.
- Baumert, J., Stanat, P., & Watermann, R. (Eds.). (2006a). *Herkunftsbedingte Disparitäten im Bildungswesen: differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit: Vertiefende Analysen im Rahmen von PISA 2000* [Social-origins based disparities in education: In-depth analyses of PISA 2000] (1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften. Retrieved from <http://gbv.ebib.com/patron/FullRecord.aspx?p=747619>
- Baumert, J., Stanat, P., & Watermann, R. (2006b). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and development of differential learning and development environments]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (1st ed., pp. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Béguin, A. A., & Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.
- Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the 2000 annual meeting of the National Council of Measurement in Education, New Orleans, IL.

- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7(3), 303–310. doi:10.1177/014662168300700306
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. doi:10.1080/0969594X.2010.513678
- Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, 15(3), 293–306. doi:10.1177/014662169101500310
- Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51. doi:10.1111/j.1745-3992.2009.00161.x
- Bildungsraum Nordwestschweiz (BR NWCH). (2012). *Checks und Aufgabensammlung im Bildungsraum Nordwestschweiz: Porträt*. Bildungsraum Nordwestschweiz (BR NWCH). Retrieved from Bildungsraum Nordwestschweiz (BR NWCH) website: <http://www.bildungsraum-nw.ch/medien/dokumente-pdf>
- Black, P., & Wiliam, D. (1998). Assessment and classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. doi:10.1080/0969595980050102
- Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.
- Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3–14. doi:10.1111/j.1745-3992.2009.00158.x
- Bundesinstitut für Bildungsforschung, Innovation & Entwicklung (BIFIE) (Ed.). (2011). *Kompetenzorientierter Unterricht in Theorie und Praxis* [Competence orientation in theory and practice]. Graz, AT: Leykam. Retrieved from https://www.bifie.at/wp-content/uploads/2017/06/bist_vs_sek1_kompetenzorientierter_unterricht_2011-03-23.pdf
- Carlson, J. E. (2011). Statistical models for vertical linking. In A. A. von Davier (Ed.), *Statistics for Social and Behavioral Sciences. Statistical Models for Test Equating, Scaling, and Linking* (pp. 59–70). New York, NY: Springer Science+Business Media LLC.
- College Board. (2017). *SAT Suite of Assessments Technical Manual: Characteristics of the SAT*. New York, NY: The College Board. Retrieved from <https://collegereadiness.collegeboard.org/pdf/sat-suite-assessments-technical-manual.pdf>
- Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education*, 37(1), 33–43. doi:10.1080/02602938.2010.494234
- Dadey, N., & Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment, Research & Evaluation*, 17(14), 1–13. Retrieved from <http://pareonline.net/getvn.asp?v=17&n=14>

- DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, 15(1), 15–31. doi:10.1207/S15324818AME1501_02
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2014). *Lehrplan 21: Rahmeninformationen*. Luzern. Retrieved from http://www.lehrplan.ch/sites/default/files/lp21_rahmeninformation_%202014-11-06.pdf
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2016a). *Lehrplan 21: Mathematik*. Luzern. Retrieved from https://v-fe.lehrplan.ch/container/V_FE_DE_Fachbereich_MA.pdf
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2016b). *Lehrplan 21: Überblick*. Luzern. Retrieved from https://v-fe.lehrplan.ch/container/V_FE_Ueberblick.pdf
- Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, 41(1), 43–68. doi:10.1111/j.1745-3984.2004.tb01158.x
- Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica*, 32(1), 107–132.
- Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135–159). New York: Springer.
- Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60(3), 327–346.
- Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (2nd ed., pp. 37–60). Mahwah, NJ: Lawrence Erlbaum Associates.
- Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, 35(2), 83–88. doi:10.1016/j.stueduc.2009.10.006
- Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.
- Hanson, B. A., & Béguin, A. A. (1999). *Separate versus concurrent estimation of IRT parameters in the common item equating design* (ACT Research Report Series No. 99-8). Iowa City, IA: ACT.
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3–24. doi:10.1177/0146621602026001001

- Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–251). New York: Springer.
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.
- Hattie, J. A. C., & Brown, G. T. L. (2007). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, 36(2), 189–201. doi:10.2190/ET.36.2.g
- Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26(2), 44–52. doi:10.1111/j.1745-3992.2007.00093.x
- Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*, 21(3), 187–206. doi:10.1080/08957340802161741
- Keizer-Mittelhaeuser, M.-A. (2014). *Modeling the effect of differential motivation on linking educational tests* (Dissertation). Tilburg University, Tilburg, the Netherlands.
- Keller, L. A., & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement*, 50(4), 390–407. doi:10.1111/jedm.12025
- Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educational and Psychological Measurement*, 71(2), 362–379.
- Keller, L. A., & Keller, R. R. (2015). The effect of changing content on IRT scaling methods. *Applied Measurement in Education*, 28(2), 99–114. doi:10.1080/08957347.2014.1002922
- Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22(2), 131–143. doi:10.1177/01466216980222003
- Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43(4), 355–381.
- Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from <http://www.iacat.org/sites/default/files/biblio/cat09kingsbury.pdf>
- Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, 41(1), 3–14. doi:10.1111/j.1745-3984.2004.tb01155.x

- Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.
- Kröhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, 14(2), 169–186. doi:10.1007/s11618-011-0185-4
- Lord, F. M. (1971a). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 31, 805–813. doi:10.1177/001316447103100401
- Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika*, 36(3), 227–242. doi:10.1007/BF02297844
- Lord, F. M. (1971c). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association*, 66(336), 707–711. doi:10.2307/2284216
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
- Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, 11(1), 1–20.
- Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Reports Series No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
- Mittelhaeuser, M.-A., Béguin, A. A., & Sijtsma, K. (2011). *Comparing the effectiveness of different linking designs: The internal anchor versus the external anchor and pre-test data* (Report No. 11-01). Retrieved from http://www.cito.nl/~media/cito_nl/Files/Onderzoek%20en%20wetenschap/cito_mrd_report_2011_01.ashx
- Mittelhaeuser, M.-A., Béguin, A. A., & Sijtsma, K. (2013). Modeling differences in test-taking motivation: Exploring the usefulness of the mixture Rasch model and person-fit statistics. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *Springer proceedings in mathematics & statistics: Vol. 66. New developments in quantitative psychology. Presentations from the 77th annual Psychometric Society meeting* (pp. 357–370). New York: Springer.
- Mittelhaeuser, M.-A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, 52(3), 339–358. doi:10.1111/jedm.12080
- Moser, U. (2009). Test. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee, & J. Oelkers (Eds.), *Handwörterbuch Erziehungswissenschaft* (pp. 866–880). Weinheim und Basel: Beltz.

- Moser, U., Buff, A., Angelone, D., & Hollenweger, J. (2011). *Nach sechs Jahren Primarschule: Deutsch, Mathematik und motivational-emotionales Befinden am Ende der 6. Klasse* [After six years of primary school: German, Mathematics and motivational-emotional well-being at the end of grade 6]. Zürich, Switzerland: Bildungsdirektion Kanton Zürich. Retrieved from Bildungsdirektion Kanton Zürich website: <https://www.ibe.uzh.ch/de/projekte/lezh.html>
- Petko, D., Mitzlaff, H., & Knüsel, D. (2007). *ICT in Primarschulen: Expertise und Forschungsübersicht*. Goldau: Institut für Medien und Schulen, Pädagogische Hochschule Schwyz.
- Petko, D., Prasse, D., & Cantieni, A. (2013). *ICT im Unterricht der Primarstufe: Ergebnisse einer empirischen Bestandsaufnahme im Kanton Thurgau*. Goldau: Pädagogische Hochschule Schwyz.
- Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement*, 50(4), 447–468. doi:10.1111/jedm.12028
- Pohl, S., Haberkorn, K., & Carstensen, C. H. (2015). Measuring competencies across the lifespan: Challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Springer proceedings in mathematics & statistics: volume 145. Dependent data in social sciences research. Forms, issues, and methods of analysis* (pp. 281–308). Cham, Switzerland: Springer.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Reusser, K. (2014). Kompetenzorientierung als Leitbegriff der Didaktik [Competence orientation as a key concept of teaching]. *Beiträge zur Lehrerinnen- und Lehrerbildung*, 32(3), 325–339.
- Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien [Challenges in Estimations of Trends in Large-Scale Assessments: A Calibration of the German PISA Data]. *Diagnostica*, 63(2), 148–165. doi:10.1026/0012-1924/a000177
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* [Textbook test theory, test construction] (2nd ed.). *Psychologie Lehrbuch*. Bern, Switzerland: Verlag Hans Huber.
- Rupp, A. A., & Zumbo, B. D. (2016). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66(1), 63–84. doi:10.1177/0013164404273942
- Sadler, R. D. (1989). Formative assessment and the design of instructional systems. *Instructional science*, 18(2), 119–144.
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26(3), 482–496. doi:10.1016/j.tate.2009.06.007

- Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education: Challenges and opportunities*. Dordrecht, the Netherlands: Springer.
- Schweizerischen Konferenz der kantonalen Erziehungsdirektoren (EDK). (2007). *Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule*. Bern. Retrieved from http://edudoc.ch/record/24711/files/HarmoS_d.pdf
- Stobart, G. (2008). *Testing Times: The uses and abuses of assessment*. Abingdon, UK: Routledge.
- Stocking, M. L. (1988). Scale drift in on-line calibration. *ETS Research Report Series*, 1988(1), i-122. doi:10.1002/j.2330-8516.1988.tb00284.x
- Strobl, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* [The Rasch model: A coherent introduction for students and practitioners] (2nd ed.). Mering, Germany: Rainer Hampp Verlag.
- Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, 8(1), 34–52. doi:10.1080/15434303.2010.536924
- Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, 16(1), 1–9.
- Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, 9, 2245. doi:10.3389/fpsyg.2018.02245
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227–253. doi:10.1080/08957340701301207
- Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement*, 12(1), 53–67. doi:10.1177/014662168801200106
- van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, 22(3), 324–343. doi:10.1080/0969594X.2014.999024
- van der Linden, W. J. (1988). *Optimizing incomplete sample designs for item response model parameters* (Research Report No. 88-5). Enschede, the Netherlands: University of Twente.
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.
- Verschoor, A. J., & Berger, S. (2015, September). *Computerized adaptive testing with online JML calibration*. Paper presented at the IACAT Conference, Cambridge, England.
- Wainer, H. (Ed.). (2000a). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

- Wainer, H. (2000b). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (2nd ed., pp. 1–21). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (2nd ed., pp. 61–100). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wauters, K., Desmet, P., & van den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4), 1183–1193. doi:10.1016/j.compedu.2011.11.020
- Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Lawrence Erlbaum Associates.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4), 473–492. doi:10.1177/014662168200600408
- Wingersky, M. S., & Lord, F. M. (1983). *An investigation of methods for reducing sampling error in certain IRT procedures* (ETS Research Reports Series No. RR-83-28-ONR). Princeton, NJ: Educational Testing Service. Retrieved from Educational Testing Service website: <http://dx.doi.org/10.1177/014662168400800312>
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1–17. doi:10.1207/s15326977ea1001_1
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97–116. doi:10.1111/j.1745-3984.1977.tb00031.x
- Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.
- Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–485). Mahwah, NJ: Lawrence Erlbaum Associates.
- Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer.
- Zenisky, A. L., & Hambleton, R. K. (2014). Multistage test designs: Moving research results into practice. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 21–37). Boca Raton, FL: CRC Press.
- Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, 80(1), 65–84. doi:10.1007/s11336-013-9369-6

Chapter 3. Efficiency of Targeted Multistage Calibration Designs under Practical Constraints: A Simulation Study³

Abstract

Calibration of an item bank for computer adaptive testing requires substantial resources. In this study, we investigated whether the efficiency of calibration under the Rasch model could be enhanced by improving the match between item difficulty and student ability. We introduced targeted multistage calibration designs, a design type that considers ability-related background variables and performance for assigning students to suitable items. Furthermore, we investigated whether uncertainty about item difficulty could impair the assembling of efficient designs. The results indicated that targeted multistage calibration designs were more efficient than ordinary targeted designs under optimal conditions. Limited knowledge about item difficulty reduced the efficiency of one of the two investigated targeted multistage calibration designs, whereas targeted designs were more robust.

³ This chapter is a minor revision of Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Efficiency of targeted multistage calibration designs under practical constraints: A simulation study. *Journal of Educational Measurement*, 56 (1), 121–146. doi: 10.1111/jedm.12203

3.1 Introduction

Computer adaptive tests (CAT) have become increasingly common in educational assessment owing to substantial technological improvements in computers in recent years. CAT provide a tailored selection of items to a student, thereby measuring his or her ability more efficiently than linear tests (e.g., Lord, 1980; van der Linden & Glas, 2010; Wainer, 2000; Weiss, 1982). Tailored test administration is especially suitable for monitoring the development of student ability over time (Weiss & Kingsbury, 1984; Wright, 1977). In such longitudinal settings, students differ not only in their current ability levels but also in the growth of their ability over time. CAT consider these differences by selecting the most suitable and informative items based on students' performance during test taking.

A calibrated item bank is required to administer CAT. Item calibration involves determining item parameters, such as item difficulty or discrimination before test administration. Such preliminary knowledge about items is essential in CAT to select the most suitable items for a given student during test taking. In the context of CAT, item response theory (IRT) is a common methodological approach to calibrating item banks, as well as for administering CAT (e.g., van der Linden & Glas, 2010; Wainer, 2000). To simplify matters, we limit our study to the Rasch model, an IRT model that includes only one item parameter, namely, item difficulty (Rasch, 1960; Rost, 1996). The Rasch model is used with several operational CAT, such as the National Council Licensure Examination for Registered Nurses (NCLEX-RN; e.g., O'Neill, Marks, & Reynolds, 2005), the Measures of Academic Progress (MAP; Northwest Evaluation Association, 2011) and the STAR Math, Reading, and Early Literacy tests (Renaissance Learning, 2015).

A CAT item bank often consists of several hundred items, particularly if it intends to cover a broad ability range in a longitudinal setting. From a practical perspective, it is very unlikely that a single group of students will respond to all items in the item bank for calibration purposes because doing so would considerably increase the testing time per student (Lord, 1980). Therefore, incomplete calibration designs are very common for calibrating broad item banks from scratch. In such designs, an item bank is divided into several subsets of items, and different subsamples work on these different subsets (Eggen & Verhelst, 2011; Mislevy & Wu, 1996). Incomplete calibration designs reduce testing time per student to a manageable duration. However, such designs require a larger sample than that required by complete calibration designs to achieve the same number of observations per item.

Unfortunately, it is often difficult to find sufficient numbers of students for calibrating items. Thus, there is a need for data collection designs that calibrate items efficiently with a limited number of students. Item calibration under the Rasch model is the most efficient if the ability of the students in a calibration sample matches the difficulty of the items needing calibration (Berger, 1991). Hence, efficient incomplete calibration designs consider students' ability for building subsamples and item difficulty for assembling item subsets. Subsamples

with relatively high ability are assigned to relatively more difficult item sets, while subsamples with relatively low ability are assigned to relatively easier item sets (Eggen & Verhelst, 2011).

However, the development of efficient incomplete calibration designs presents two major challenges. On the one hand, it is difficult to build homogenous subsamples for different ability levels. Usually, ability-related background variables, such as grades in school, are used to group students by ability. This is the basic idea of targeted calibration designs (Eggen & Verhelst, 2011; Mislevy & Wu, 1996). Nevertheless, students within such subsamples still vary in their ability, which results in a loss of efficiency. On the other hand, difficulty of the items is often only vaguely known prior to calibration, which makes it challenging to assemble item sets of similar difficulty and to assign the items to students of corresponding abilities. The lack of knowledge about item difficulty prior to calibration has been considered in a few studies that have investigated the calibration of a small number of field-test items during the administration of operational CAT (e.g., Ali & Chang, 2014; Kingsbury, 2009). Furthermore, Makransky and Glas (2010) have explored different strategies for automatic online calibration of new item pools with unknown item parameters. However, this study focused on ability estimation during the calibration phase and not on item parameter estimation itself (cf. Ali & Chang, 2014). Studies that have compared the efficiency of incomplete calibration designs for calibrating an item pool from scratch have neglected the practical constraint of unknown item parameters by assuming that the difficulty of the items is known (e.g., Berger, 1991; Stocking, 1988).

We address this gap by extending previous research on the efficiency of incomplete calibration designs in two ways: first, we investigate the extent to which it is possible to improve the match between student ability and item difficulty and, therefore, calibration efficiency under the Rasch model, by narrowing subsamples based on students' performance. To this end, we introduce the concept of targeted multistage calibration designs as an extension of targeted calibration designs. Second, we explore how uncertainty about the items' difficulty affects the efficiency of targeted calibration designs, as well as that of targeted multistage calibration designs.

3.1.1 Accuracy and Efficiency in Rasch Model-Based Item Calibration

The efficiency of any item calibration design depends largely on the underlying methodological approach. IRT refers to a family of models that express the probability of a student solving an item correctly as a function of student ability and item difficulty (Lord, 1980). For the Rasch model—the simplest unidimensional IRT model—the probability of a student solving a specific item correctly is given by

$$P(X_{ij} = 1 | \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} = p_{ij}, \quad (3.1)$$

where θ_i represents the ability of student i , β_j is the difficulty of item j , and p_{ij} corresponds to the probability that student i solves item j correctly (Rasch, 1960; Rost, 1996; Strobl, 2012). Within this framework, item calibration refers to establishing model fit and estimating the item difficulty parameter β_j (Eggen & Verhelst, 2011; Vale & Gialluca, 1988). Item difficulty can be estimated by following various maximum likelihood estimation procedures, and the resulting standard error $SE(\hat{\beta}_j)$ serves as a measure of the accuracy of the estimate. The standard error of the estimated difficulty of item j can be approximated as (Rost, 1996)

$$SE(\hat{\beta}_j) \sim \sqrt{\frac{1}{\sum_{i=1}^N p_{ij}(1 - p_{ij})}}. \quad (3.2)$$

On the one hand, we can infer from Equation 3.2 that the calibration sample size is a key factor for calibrating item parameters accurately according to the Rasch model (e.g., Rost, 1996; Wright, 1977). The larger the number of students N who provide responses for a specific item, the greater is the accuracy of item difficulty estimation. According to Equation 3.2, we can conclude that the standard error $SE(\hat{\beta}_j)$ is inversely proportional to the square root of the number of students N :

$$SE(\hat{\beta}_j) \sim \frac{1}{\sqrt{N}}. \quad (3.3)$$

Thus, a reduction of 50% in the standard error $SE(\hat{\beta}_j)$ requires quadrupling the sample size subject to the constraint that the properties of the sample, as well as those of the item pool, remain constant.

On the other hand, we can deduce from Equation 3.2, that the accuracy of item difficulty estimation depends on the relationship between item difficulty and ability of the students in a calibration sample (Berger, 1991; Eggen & Verhelst, 2011; Stocking, 1988; Wright, 1977). Item difficulty can be estimated most accurately under the Rasch model if the mean ability of the sample is close to the difficulty of the items and if the standard deviation of the students' ability is small (Berger, 1991; Rost, 1996; van der Linden, 1988). If θ is equal to β , we have a probability of 50% for solving item i correctly, which in turn, results in minimization of $SE(\hat{\beta}_j)$.

In sum, we can calibrate item difficulty most accurately under the Rasch model if the calibration sample is large and if it includes students with abilities that correspond to the items' difficulty. Enhancing the calibration sample is often difficult in practical settings, which leaves optimization of the relationship between item difficulty and student ability as the only option for improving the accuracy of item difficulty estimation. As shown by Berger (1991), the fit between these two variables depends largely on the calibration design.

3.1.2 Designs for Item Calibration

We distinguish three different types of calibration designs that take item difficulty and student ability into account in optimizing item calibration: (1) targeted calibration designs, (2) multistage calibration designs, and (3) targeted multistage calibration designs, which refer to a combination of the first two designs. We describe all three designs based on an example and discuss the extent to which the designs allow for efficient item calibration under ideal conditions; that is, if the difficulty of all items in the design is known. Practical constraints of limited knowledge about difficulty of the items in an item pool will be discussed in the subsequent section.

3.1.2.1 Targeted calibration designs

Test designs for targeted testing consist of several test booklets of varying mean difficulty, and students are assigned to the most suitable booklet by means of ability-related background variables, such as grades in school (Eggen & Verhelst, 2011; Mislevy & Wu, 1996). Students can be classified and, therefore, sampled based on such background variables before test administration. The main goal of targeted testing is to estimate student ability more precisely and to prevent students from getting discouraged by items that are not suited to their ability level. However, the same mechanism, that is, assigning students to test booklets based on background variables, can also be used to improve the accuracy of item difficulty estimation during calibration.

Figure 3.1. Examples of targeted (left) and multistage calibration designs (right). y = ability-related background variable (e.g., grades in school); s = student's score in stage 1, c = cut-off score.

Figure 3.1 shows a basic example of a targeted design on the left. In the example, students are grouped by a background variable y (e.g., grades in school) such that the first subsample with $y = 1$ is assumed to have lower mean ability than the second subsample consisting of students with $y = 2$. This results in two subsamples with greater homogeneity in terms of abilities than the overall sample. Similarly, the item pool is divided into two test

booklets such that the first booklet contains mainly items that are easier than those in the second booklet. The overlap between the two booklets (i.e., linking items of intermediate difficulty) ensures that all items can be calibrated on the same scale by using IRT calibration methods (Eggen & Verhelst, 2011; Kolen, 2007; Kolen & Brennan, 2014). Variance in item difficulty is lower within a single booklet than in the total item pool.

According to the example design, students from group $y = 1$ are assigned to the easier test booklet, while students from group $y = 2$ are assigned to the more difficult test booklet. This optimization of the relationship between item difficulty and student ability allows for more accurate estimation of item difficulty over the entire item pool than random assignment of students to random sets of items from the same item pool. However, this simple example design has one significant disadvantage: linking items that are included in both booklets have twice as many observations as the other items in the design. Consequently, item difficulty of the linking items is estimated much more accurately than that of the remaining items (see Equation 3.3). This imbalance can be corrected by using more complex designs, as we will show later.

3.1.2.2 *Multistage calibration designs*

Multistage tests consist of several parts (i.e., stages), which, in turn, include multiple item sets—called modules—of varying difficulty (Yan, von Davier, & Lewis, 2014; Zenisky, Hambleton, & Luecht, 2010). Students' performance in the first stage determines whether they receive an easier or a more difficult module in the second stage. The decision is based on predefined routing rules. Several studies have shown that multistage tests are more efficient than linear tests in estimating student ability owing to closer alignment between item difficulty and student ability (e.g., Yan et al., 2014). Such an alignment is also advantageous from the viewpoint of estimating item difficulty more precisely (Eggen & Verhelst, 2011; Glas & Geerlings, 2009; Zwitser & Maris, 2015).

To define modules and routing rules that lead to an optimal match between items and students, knowledge is required about ability distribution in the sample, as well as about item difficulty. The routing rules control which and how many students reach a certain module in the second or any subsequent stage. Thus, they determine the ability distributions of the subsamples, as well as the number of observations per item for the modules. For calibration designs, it is desirable to control the number of observations per item such that the difficulty parameters of all items can be estimated with comparable accuracy (Glas & Geerlings, 2009).

Figure 3.1 shows a basic example of a multistage design on the right. In the first stage, all students start with the same module. This routing module comprises items of intermediate difficulty, and it establishes a link between the two modules in the subsequent stage. Module selection in stage 2 depends on the student's score s in the routing module and on the cut-off score c . Students with a low score (i.e., $s < c$), and thus low ability, are directed to the easy module in stage 2, while students with a high score (i.e., $s \geq c$) are directed to the difficult module. This ability-based routing leads to closer alignment between item difficulty and student

ability in the second stage, and—under the condition of an equal number of observations per item—it increases the accuracy of item difficulty estimates.

Again, one complication of this simple example design is that the number of observations per item varies among different groups of items. The items in the first stage are solved by the entire sample, but the sample is divided between the two modules in the second stage. This imbalance can be corrected by using more complex designs with parallel modules (i.e., including two equally difficult modules in stage 1), as we will elaborate later.

3.1.2.3 Targeted multistage calibration designs

Targeted multistage calibration designs refer to a combination of the two previously presented designs. Figure 3.2 shows an example of such a design that considers two different subsamples and includes two stages. Again, knowledge about item difficulty is an essential precondition for defining modules characterized by different mean difficulty, as well as for specifying appropriate routing rules. In the first stage, students are either assigned to an easier or a more difficult routing module based on the background variable y (i.e., targeted assignment). Assignment in the second stage depends on performance within the first modules and on the routing rules related to each module. According to the example design in Figure 3.2, students from group $y = 1$ with a low score in the easy routing module are guided to the easy module within stage 2. Students from group $y = 1$ with a high score in the easy routing module (i.e., $s_1 \geq c_1$) and students from group $y = 2$ with a low score in the difficult routing module (i.e., $s_2 < c_2$) are both directed to a module of intermediate difficulty, which allows for a link among all modules in the design. Last, but not least, students from group $y = 2$ with a high score in the difficult routing module are directed to the difficult module of stage 2.

Figure 3.2. Example of targeted multistage calibration design. y = ability-related background variable (e.g., grades in school); s_1 and s_2 = student's scores in stage 1, c_1 and c_2 = cut-off scores.

Targeted multistage calibration designs have the advantage that they consider preliminary background information, as well as performance information, for optimizing the fit

between item difficulty and student ability. In the first stage, background information is used to select the most suitable routing module, which results in an improved alignment between item difficulty and student ability compared to traditional multistage designs. In the second stage, the assignment no longer depends on background variables as in targeted designs, but rather on the performance of the students. Performance is most likely a better predictor of the students' true ability than an ability-related background variable such as grades in school.

Similar to multistage calibration designs, it is important to control the number of observations per item in targeted multistage calibration designs because this factor has a large influence on the accuracy of item difficulty estimation. From this viewpoint, the design in Figure 3.2 is not ideal for calibration purposes because the number of observations is higher for the items in the two modules of stage 1 (half of the total sample) than for the items in the three modules of stage 2 (one-third of the total sample). As for the other two designs, we will present options for controlling the number of observations per item by developing more complex targeted multistage calibration designs.

3.1.3 Uncertainty of Item Difficulty during Test Construction

When arranging items within a given calibration design, test developers need to know the item difficulties to locate items optimally within the design (Glas & Geerlings, 2009). For example, when implementing a targeted calibration design, as shown in Figure 3.1, a test developer would need items from three different difficulty levels, namely, easy items for group 1, difficult items for group 2, and intermediate items for linking. Knowledge about item difficulty is even more relevant when constructing modules for multistage designs and for defining the related routing rules. However, usually, no empirical information about item difficulty is available before item calibration. Therefore, the decision about item distribution within the calibration design and the definition of the routing rules depend mainly on the expertise and experience of the test developer and/or other involved experts.

Several studies have investigated the accuracy with which experts, such as test developers, content experts, or item authors, can rate item difficulty, and they have found moderate to high correlations between the ratings and the empirical item difficulties (e.g., Bejar, 1983; Hambleton & Jirka, 2006; Sydorenko, 2011; Wauters, Desmet, & van den Noortgate, 2012). The accuracy of difficulty ratings depends on several factors, such as the content, item type, training of the judges, and number of judges (Hambleton & Jirka, 2006, pp. 407–408).

We conclude from these findings that the distribution of items across modules of different target difficulties might deviate in a practical setting from the optimal distribution in a theoretical setting (i.e., optimal condition), where the difficulty of all items is known. Owing to missing empirical data, test developers might fail to assign all items to the most suitable location within a calibration design. Instead, they might include some easy items in test booklets or modules with high target difficulty or vice versa, owing to over- or underestimation of item

difficulty. This results in more heterogeneous item sets than those under the optimal condition. The number of misplaced items depends on the accuracy with which experts can predict item difficulty.

3.1.4 The Present Study

The aim of this study is to answer two research questions: first, we investigate whether we can achieve higher efficiency in item calibration with targeted multistage calibration designs than with targeted calibration designs. If item difficulty is known during test construction, it is possible to optimize the positions of items within the design and to assemble item sets of homogenous difficulty. For this optimal condition, we assume that a targeted multistage calibration design improves the fit between item difficulty and student ability and, therefore, increases the accuracy of item difficulty estimation. Furthermore, we hypothesize that the efficiency gain of a targeted multistage calibration design depends on the exact design specification (i.e., composition of the different modules and the related routing rules).

Second, we examine how limited a priori knowledge about item difficulty affects the efficiency of both targeted calibration designs and targeted multistage calibration designs. If a priori knowledge about item difficulty is limited, we expect that misplaced items impair the efficiency of all calibration designs. In addition, we hypothesize that targeted multistage calibration designs are impaired to a greater extent by misplaced items than targeted calibration designs because knowledge about item difficulty is needed not only to construct the modules but also to specify the routing rules. Finally, we assume that the degree of vulnerability of targeted multistage designs to misplaced items depends on the exact design specifications.

All hypotheses are addressed in a simulation study in which we vary the calibration design, as well as the accuracy of item distribution, across the different booklets or modules within each design (i.e., number of misplaced items).

3.2 Method

To highlight the practical constraints to item calibration, we embedded our simulation study in a practical context, namely, the development of an adaptive online item bank for formative assessment in northwestern Switzerland (Berger, Moser, Verschoor, & Eggen, 2015; Berger, Verschoor, Moser, & Oostlander, 2014). The aim of this online item bank is to assess students' ability and to monitor their progress throughout compulsory school.

3.2.1 Ability Distributions and Item Pool

For each simulation run, two samples of 1,300 simulees were drawn randomly from two normal distributions, $\theta \sim N_1(0,1)$ and $\theta \sim N_2(0.8,1)$, to simulate students from two successive grades in school. The selected sample size represented 10% of the total student population from two

successive school years in northwestern Switzerland (Berger, Moser et al., 2015), which refers to the expected response rate for a calibration study. The performance progress between the two grades in school was modeled on selected results from a longitudinal study that investigated, among other variables, the performance progress of Swiss students during primary school (i.e., progress in mathematics from grades 3 to 6; Angelone, Keller, & Moser, 2013, p. 35).

We generated an artificial item pool of 180 dichotomous Rasch items with equally spaced difficulty parameters β_j ranging from -1.5 to 2.3 to provide items that cover the range of ± 1.5 standard deviations from the means of both samples. Thus, the ability of approximately 16% of the simulees from the two samples lay outside of this range, which ensured variation in the response patterns for all items. We ensured uniform distribution of item difficulty to investigate the efficiency of calibration in relation to item difficulty. The size of the item pool was adjusted to the sample size to achieve a reasonable number of observations per item (i.e., approximately 400 observations per item; cf. Wright, 1977, p. 106). Furthermore, we selected an item pool size that allowed us to construct balanced designs for all investigated design types, given a practically relevant test length (see next paragraph for further details). The same item pool was used for all conditions within the simulation study.

3.2.2 Test Designs

The simulation study included three different calibration designs (i.e., a targeted calibration design and two variations of a targeted multistage calibration design) with a multi-matrix structure, which is often used in large calibration studies, as well as a random condition as a baseline. In all four design conditions, the item pool described above was used, and test length was fixed to 30 items, which refers to the number of items that we expect students to solve within one school lesson (i.e., 45 minutes). All designs aimed to achieve an equal number of observations per item. Keeping this factor constant allowed us to focus on the effect of the match between item difficulty and student ability on the efficiency of the different designs.

3.2.2.1 Targeted calibration designs

In the targeted calibration design, the 180 items were divided into six easy and six difficult booklets with 30 items each (see Figure 3.3). To construct equally difficult booklets for each difficulty level, the items were sorted by difficulty and split into three equally large categories: an “easy” category (items 1–60), “intermediate” category (items 61–120), and “difficult” category (items 121–180). Each category was further divided into six equally difficult modules of 10 items, which resulted in a total of 18 modules (i.e., six modules per category). All modules were included in two different booklets for linking purposes. The six easy booklets contained two easy and one intermediate module each, while the six difficult booklets included one intermediate and two difficult modules. Each simulee from the low ability group was assigned randomly to one of the six easy booklets, and each simulee from the high ability group was assigned randomly to one of the six difficult booklets.

		Modules											
Ability group	Booklets	Easy (cat1)				Intermediate (cat2)				Difficult (cat3)			
Low	B1	10	10			10							
	B2		10	10			10						
	B3			10	10			10					
	B4				10	10			10				
	B5					10	10			10			
	B6	10						10					
High	B7							10	10	10			
	B8								10	10	10		
	B9									10	10	10	
	B10										10	10	10
	B11											10	10
	B12												10

Figure 3.3. Targeted calibration design. B1–B6 = easy booklets; B7–B12 = difficult booklets. Each column represents one module of ten items. The 18 modules are classified into three difficulty categories: cat1 = “easy,” cat2 = “intermediate,” and cat3 = “difficult.”

3.2.2.2 Targeted multistage calibration designs

We investigated two different targeted multistage calibration designs in our simulation study to analyze the effect of different design specifications on calibration efficiency. Both designs were specified as simply as possible given the item pool size of 180 items, test length of 30 items, and requirement of an equal number of observations per item. The designs consisted of two stages, and each stage included half of the total item pool (i.e., 90 items). In general, items with more extreme difficulty were included in the second stage, where simulees were assigned to the different modules based on their performance in the first stage. Within each stage, the 90 items were divided into six modules of 15 items, resulting in 12 modules in total. The modules were not linked, and each item was included in one module only. Instead, links between the different modules were established by overlapping paths (i.e., various combinations of different modules from both stages).

The two designs differed in terms of module composition and routing. Figure 3.4 shows the targeted multistage calibration design A (TMST A). In this design, we grouped the 180 items into five difficulty categories. The easiest, intermediate, and most difficult categories were assigned to the second stage and included 30 items each: items 1–30, items 76–105, and items 151–180, respectively. The relatively easy and relatively difficult categories were assigned to the first stage and included 45 items each: items 31–75 and items 106–150, respectively. Within each category, we further divided the items into equally difficult modules of 15 items. In the first stage, simulees from the low ability group were assigned randomly to one of the three relatively easy modules, and simulees from the high ability group were assigned randomly to one of the three relatively difficult modules. The score in the first stage determined module selection in the second stage. Simulees from the low ability group could reach either an

easy or an intermediate module ($c_1 = 11$); simulees from the high ability group were guided to either an intermediate or a difficult module ($c_2 = 5$). The two cut-off scores were determined in preceding simulations to achieve an equal number of observations per module in the second stage. The aim was to guide one-third of each sample to the intermediate modules and to guide the remaining two-thirds of each sample to the easy or the difficult modules. The design was linked in two ways: first, simulees from both samples could reach the intermediate modules. Second, the random assignment of students to modules of equal difficulty resulted in multiple combinations of the modules over the two stages (i.e., overlapping paths).

Ability group	Stage 1		Routing	Stage 2		
	Rel. easy (cat2)	Rel. difficult (cat4)		Easy (cat1)	Interm. (cat3)	Difficult (cat5)
Low	M3 15		$s_1 < 11$	M1 15		
			$s_1 \geq 11$		M6 15	
High		M8 15	$s_2 < 5$			
			$s_2 \geq 5$		M7 15	
						M11 15
						M12 15

Figure 3.4. Targeted multistage calibration design A (TMST A). M1/M2 = easy modules (difficulty category 1); M3–M5 = relatively easy modules (category 2); M6/M7 = intermediate modules (category 3); M8–M10 = relatively difficult modules (category 4); M11/M12 = difficult modules (category 5); s_1 = score of low ability group; s_2 = score of high ability group.

In the second targeted multistage design (TMST B, see Figure 3.5), we distinguished four instead of five different difficulty levels and split the item pool into four categories of 45 items: items 1–45, items 46–90, items 91–135, and items 136–180. Each category was further divided into three equally difficult modules of 15 items. Simulees from the low ability group started with one of the three relatively easy modules, and simulees from the high ability group were assigned randomly to one of the three relatively difficult modules. Simulees' scores in the first stage determined the selection of difficulty level in the second stage. In both samples, the simulees were directed to one of the three easy modules if they scored lower than the sample-specific cut-off score ($c_1 = 11$; $c_2 = 5$). Simulees with equal or higher scores were directed to one of the three difficult modules. Again, the two cut-off scores were determined by preceding simulations to achieve equal numbers of observations per module in the second stage. The aim was to guide three-fourths of the low ability group to the easy modules and one-fourth to the difficult modules. For the high ability group, the goal was to route one-fourth of the simulees to the easy modules and three-fourths to the difficult modules. Similar to TMST A, the different

modules in TMST B were linked in two ways: first, all modules in the second stage could be reached by both samples, and, second, the random assignment of students to modules of equal difficulty resulted in multiple combinations of the modules over the two stages (i.e., overlapping paths).

Ability group	Stage 1		Routing	Stage 2	
	Rel. easy (cat2)	Rel. difficult (cat3)		Easy (cat1)	Difficult (cat4)
Low	M4 15		$s_1 < 11$	M1 15	
				M2 15	
					M3 15
			$s_1 \geq 11$		M10 15
					M11 15
					M12 15
High		M7 15	$s_2 < 5$	M1 15	
				M2 15	
					M3 15
		M8 15	$s_2 \geq 5$		M10 15
					M11 15
		M9 15			M12 15

Figure 3.5. Targeted multistage calibration design B (TMST B). M1–M3 = easy modules (difficulty category 1); M4–M6 = relatively easy modules (category 2); M7–M9 = relatively difficult modules (category 3); M10–M12 = difficult modules (category 4); s_1 = score of low ability group; s_2 = score of high ability group.

3.2.2.3 Random test

We included a random test as a baseline in our simulation study. In this condition, each simulee was assigned to a random selection of 30 items out of the entire item pool. Hence, neither the ability of the target sample nor its performance during test taking was considered during item selection.

3.2.3 Simulation with Limited Knowledge about Item Difficulty during Test Development

To simulate the case with limited knowledge about item difficulty during test development, we manipulated the order of the items in relation to item difficulty and, therefore, the distribution of the items across the different modules in the designs. In the optimal condition with complete knowledge about item difficulty, the items were ordered by difficulty, as described earlier, which resulted in a correlation of $r = 1.0$ between item order and item difficulty. In addition to this full knowledge condition, we investigated two additional conditions in our simulation study. In the first condition, item order and item difficulty were correlated at $r = .4$ to simulate a low amount of knowledge during test construction following results reported by Berger, Oostlander, Verschoor, Eggen, and Moser (2015). This condition resulted in a high number of

misplaced items. In the second condition, item order and item difficulty were correlated higher ($r = .6$) to represent a condition with a medium amount of knowledge during test construction and a resulting medium number of misplaced items. Both conditions were applied to the targeted calibration design, as well as to the two targeted multistage calibration designs, meaning that the manipulated item order was used as a basis to distribute the items within the designs. Owing to the manipulated item order, item categories were no longer homogenous, but included a few items that exceeded or fell below the envisaged difficulty range. This, in turn, affected the difficulty range and the mean difficulty of the booklets and modules.

3.2.4 Item Response Generation and Calibration

Altogether, the simulation study included 10 different conditions: the three calibration designs (i.e., targeted calibration design and two versions of targeted multistage calibration designs) were combined with the three variations of item distribution (i.e., full, medium, and low knowledge) and completed by the random condition that served as a baseline. For each condition, we generated 20,000 datasets according to the Rasch model. For each dataset, we drew new samples from the two ability distributions described and generated new response patterns consisting of $30 \times 2,600$ responses. The Multidimensional Item Response Theory (MIRT) software application (Glas, 2010) was used to generate marginal maximum likelihood (MML) estimates of the item parameters from the Rasch model. All designs were estimated using two marginal proficiency distributions, that is, one marginal distribution per sample (cf. Eggen & Verhelst, 2011).

3.2.5 Evaluation Criteria

3.2.5.1 *Distribution of item difficulty per booklet or module*

As a descriptive evaluation criterion, we explored the distribution of item difficulty within and among the different booklets and modules. This distribution was predefined under the full knowledge condition, where the item difficulty of each item was known. For the conditions with limited knowledge, the distribution of item difficulty depended on manipulation of the item order.

3.2.5.2 *Bias and root mean square error*

To determine the accuracy of item parameter estimation in the different calibration designs, the bias and the root mean square error (RMSE) of each item in each condition were computed. That are,

$$\text{Bias}(\hat{\beta}_j) = \frac{\sum_{k=1}^{20000} (\hat{\beta}_{kj} - \beta_j)}{20000} \quad \text{and} \quad (3.4)$$

$$\text{RMSE}(\hat{\beta}_j) = \sqrt{\frac{\sum_{k=1}^{20000} (\hat{\beta}_{kj} - \beta_j)^2}{20000}}, \quad (3.5)$$

where k and j represent replications and items, respectively, and $\hat{\beta}$ denotes the estimate of item difficulty β .

3.2.5.3 Mean number of observations per item

As an additional criterion for the efficiency of the different calibration conditions, we investigated the mean number of observations per item over 20,000 simulation runs within each simulation condition. That is,

$$N_j = \sqrt{\frac{\sum_{k=1}^{20\,000} N_{kj}}{20000}}, \quad (3.6)$$

where k and j represent replications and items, respectively, and N denotes the number of observations. As described earlier, the higher the number of observations per item, the higher is the accuracy of item difficulty estimation. The intention in our setting was to asymptotically achieve equal mean numbers of observations for all items (i.e., $SD_N = 0$) to ensure that the influence of this factor on calibration efficiency was constant for all items.

3.3 Results

3.3.1 Distribution of Item Difficulty per Booklet or Module

Figure 3.6 shows the distribution of item difficulty within the different booklets and modules for each design and knowledge condition by means of boxplots. For the full knowledge condition (dark gray boxplots), this distribution was given by the design specifications. Independent of the design, the booklets and modules representing different difficulty levels differed clearly in terms of difficulty range and median (e.g., B06 vs. B07). Meanwhile, the booklets and modules that represented one difficulty level had comparable item difficulty distributions (e.g., B01 vs. B02). The range of item difficulty within a single booklet or module was larger in the targeted design than in the two targeted multistage designs. This result is directly related to the number of difficulty categories defined for each design (i.e., two difficulty categories for the targeted design, five for TMST A, and four for TMST B).

By contrast, the simulation with limited knowledge about item difficulty during test construction (i.e., medium and low knowledge) resulted in more heterogeneous booklets and modules (see medium and light gray boxplots in Figure 3.6). The range of item difficulty within the booklets and modules increased under these conditions, and the medians shifted towards the mean difficulty of the total item pool ($M = 0.400$). Consequently, differences between booklets and modules intended to be equally difficult increased, and differentiation between

booklets and modules intended to represent different difficulty levels decreased. These effects were generally more prominent under the low knowledge condition than under the medium knowledge condition. From these findings, we conclude that our manipulation led to the envisaged misplacement of items in the different designs.

Figure 3.6. Boxplots of item difficulty per design, booklet/module, and knowledge condition.

3.3.2 Bias($\hat{\beta}$), Mean RMSE($\hat{\beta}$) and Mean Number of Observations per Simulation Condition

This section provides a general overview of the efficiency of the different simulation conditions. Table 3.1 presents the range of Bias($\hat{\beta}$), the mean of RMSE($\hat{\beta}$) and the related standard error based on the first 1,000 simulation runs under each condition as an indicator of overall efficiency. As indicators of the relative efficiency of the different designs, on the one hand, Table 3.1 reports the relative decrease or increase in RMSE($\hat{\beta}$) compared to the random condition and targeted design within the corresponding knowledge condition; on the other hand, the table includes for each condition the differences in the number of students needed to achieve the same overall efficiency with the random condition and targeted design (absolute number and percentage of total sample, i.e., 2,600) given Equation 3.3. This information serves as an

indicator of the practical relevance of the reported differences. Also, Table 3.1 lists the standard deviation of the mean number of observations under each simulation condition, as well as the mean number of observations per difficulty category (i.e., design specific categorization of the modules' difficulty levels).

In all conditions, we observed very little bias and the highest absolute value of $\text{Bias}(\hat{\beta})$ was found in the Random condition (i.e., $\text{Bias}(\hat{\beta}) = 0.012$). Independent of the design, the mean $\text{RMSE}(\hat{\beta})$ was generally lower under the full knowledge condition than under the random or the limited knowledge conditions. The lowest mean $\text{RMSE}(\hat{\beta})$ was found for TMST B under the full knowledge condition ($M = 0.110$). The efficiency gain of TMST B was 7% compared to the random condition and 4% compared to the targeted design. To achieve the same efficiency with the baseline designs, we would need to increase sample size by 336 and 200 students, respectively. The overall efficiency gain of TMST A under the full knowledge condition was slightly lower ($M = 0.112$), yet the design outperformed the targeted design ($M = 0.115$), which, in turn, yielded lower mean $\text{RMSE}(\hat{\beta})$ than the random condition ($M = 0.119$). All differences were statistically significant, as can be derived from the small standard errors of $\text{RMSE}(\hat{\beta})$ in Table 3.1.

Under the medium knowledge condition, the efficiency gain of TMST B ($M = 0.116$) and the targeted design ($M = 0.117$) were considerably smaller than that under the full knowledge condition. Nevertheless, the gain of TMST B was still statistically significant and practically relevant in terms of the number of students: 114 (4%) additional students would be required to achieve the same efficiency with the random condition, and 52 (2%) additional students would be required for the targeted design. The mean $\text{RMSE}(\hat{\beta})$ of TMST A was even higher than that of the two related baselines ($M = 0.120$), thereby indicating a significant efficiency loss. Similar but more prominent results were found for the low knowledge condition: TMST B ($M = 0.118$) was slightly more efficient than the two baseline conditions, which showed a comparable mean $\text{RMSE}(\hat{\beta})$ ($M = 0.119$), whereas TMST A resulted in a considerable efficiency loss (i.e., -6% ; $M = 0.125$). To compensate for this loss, we would need to increase the sample size by more than 300 students or 12% of the total sample.

Table 3.1. Range of Bias($\hat{\beta}$), Mean and Standard Error of RMSE($\hat{\beta}$), Related Gain per Simulation Condition, and Distribution of Mean Number of Observations

Condition	Bias($\hat{\beta}$) ^a		RMSE($\hat{\beta}$)		Gain over Random		Gain over Targeted ^c		SD	N				
	Min	Max	M	SE ^b	%	N (%)	%	N (%)		M ₁	M ₂	M ₃	M ₄	M ₅
Random	-0.011	0.012	0.119	0.00028	--	--	--	--	0.549	433	--	--	--	--
Full Knowledge														
Targeted	-0.007	0.007	0.115	0.00027	3	159 (6)	--	--	0.222	433	433	433	--	--
TMST A	-0.005	0.005	0.112	0.00026	6	267 (10)	2	122 (5)	6.958	426	433	447	433	426
TMST B	-0.006	0.007	0.110	0.00026	7	336 (13)	4	200 (8)	0.114	433	433	433	433	--
Med. Knowledge														
Targeted	-0.007	0.007	0.117	0.00028	1	64 (2)	--	--	0.175	433	433	433	--	--
TMST A	-0.007	0.007	0.120	0.00029	-1	-67 (-3)	-3	-138 (-5)	60.958	492	433	312	433	496
TMST B	-0.007	0.006	0.116	0.00027	2	114 (4)	1	52 (2)	6.315	442	433	433	424	--
Low Knowledge														
Targeted	-0.006	0.007	0.119	0.00028	0	4 (0)	--	--	0.145	433	433	433	--	--
TMST A	-0.010	0.010	0.125	0.00030	-6	-315 (-12)	-6	-321 (-12)	90.950	519	433	252	433	529
TMST B	-0.008	0.007	0.118	0.00028	1	48 (2)	1	43 (2)	2.925	429	433	433	437	--

Note. The numbers in bold represent efficiency loss compared to the baseline condition; gray numbers represent mean numbers of observations equal to the overall mean of M = 433. N(%) = number of additional students needed to achieve the same efficiency with the baseline design, 100% = 2,600 students; M₁–M₅ = mean numbers of observations per difficulty category (i.e., design-specific categorization of modules' difficulty levels).

^aMean of Bias($\hat{\beta}$) was zero in all conditions.

^bStandard error of RMSE($\hat{\beta}$) based on the first 1,000 simulation runs.

^cGain over targeted design within a specific knowledge condition.

The standard deviations of the mean number of observations displayed in column 10 of Table 3.1 provide some insights into the roots of efficiency differences among the different conditions. The mean numbers of observations were well balanced in the random and the targeted designs for all three knowledge conditions. On average, all items were assigned to approximately 433 simulees, which corresponded to one-sixth of the total sample (i.e., 2,600 simulees). Moreover, variations in TMST B were very small under all simulation conditions ($SD_{full} = 0.114$; $SD_{Medium} = 6.315$; $SD_{Low} = 2.925$), so only the modules in stage 2 differed slightly in their mean number of observations per item (i.e., M_1 and M_4). However, for TMST A, the number of observations per item depended largely on the available knowledge about item difficulty. Low knowledge was associated with a considerable increase in the standard deviation of the mean number of observations per item ($SD = 90.950$) compared to the medium ($SD = 60.958$) and the full knowledge conditions ($SD = 0.114$). Although the modules of stage 1 showed the expected mean numbers of observations ($M_2 = M_4 = 433$), the mean number of observations varied considerably among modules of different difficulty levels in stage 2 (i.e., M_1 , M_3 , and M_5).

In sum, these general findings were consistent with our hypothesis that targeted multistage calibration designs calibrate item difficulty more accurately than targeted designs if item difficulty is known. Furthermore, the results support our hypothesis that limited a priori knowledge about item difficulty impairs the efficiency of all calibration designs. In line with our expectations, the results suggest that limited knowledge does not affect all designs to the same extent. However, counter to our hypothesis, one of the two targeted multistage designs outperformed the targeted design under the limited knowledge conditions.

3.3.3 RMSE($\hat{\beta}$) and Mean Number of Observations per Item

3.3.3.1 Optimal condition with full knowledge

Figure 3.7 shows RMSE($\hat{\beta}$) for each item in relation to its true difficulty β for all four designs under the full knowledge condition. For the random baseline condition, RMSE($\hat{\beta}$) was the lowest for items with a difficulty of $\beta \approx 0.400$, which corresponded to the mean difficulty of the item pool, as well as to the mean ability of the entire sample ($M_\beta = M_\theta = 0.400$). For items with more extreme difficulty, RMSE($\hat{\beta}$) increased, which resulted in a U-shaped relationship between RMSE($\hat{\beta}$) and β over the entire item pool. Given the normal distribution of ability in the two samples, the number of simulees with abilities that matched the difficulties of the extreme items was limited. Therefore, the difficulty of these items was estimated less precisely than that of the items close to the mean abilities of the two samples ($M_{\theta_1} = 0.000$; $M_{\theta_2} = 0.800$).

Figure 3.7. $RMSE(\hat{\beta})$ per item in relation to item difficulty for the four different designs under the full knowledge condition.

The curves related to the other three designs showed a similar tendency to a U-shaped course, but they were characterized by a few kinks (i.e., rapid increase or decrease in $RMSE(\hat{\beta})$ between two items). These kinks divided the curves into multiple sections, which corresponded to the difficulty categories that were part of the design specifications. The gray curve representing the targeted design included two kinks that divided the curve into three sections. The first section included all items from the “easy” category that were only administered to sample 1, the second section corresponded to the linking items (“intermediate” difficulty category), and the third section represented the items from the “difficult” category that were administered only to sample 2. Thus, the three categories of items were administered to three different groups of simulees with different ability distributions, which resulted in differences in the accuracy of item difficulty estimation. The $RMSE(\hat{\beta})$ values of the items in the intermediate difficulty category were comparable to those found in the random condition. For the easy and difficult items, however, targeted assignment of simulees led to lower $RMSE(\hat{\beta})$ values than random assignment. These results suggest that the targeted design outperformed the random condition by improving the calibration of easy and difficult items.

For TMST A, the item pool was divided into five difficulty categories, which resulted in a curve consisting of five sections. Categories 2 and 4 corresponded to the items in stage 1 (i.e., relatively easy and relatively difficult items). The $RMSE(\hat{\beta})$ of these items were comparable to the outcomes of the random and the targeted conditions. However, TMST A provided more accurate item difficulty estimates for the items in stage 2. On the one hand, a reduction in $RMSE(\hat{\beta})$ was achieved for easy and difficult items (i.e., categories 1 and 5). These items were administered to low performing simulees from sample 1 or high performing simulees from sample 2. On the other hand, $RMSE(\hat{\beta})$ was also lower for intermediate items from category 3, which were administered to high performing simulees from sample 1 and to low performing simulees from sample 2. The efficiency gain in this intermediate category could

be explained partly by the enhanced mean numbers of observations per item caused by the slightly unbalanced routing ($M_3 = 447$; $M_1 = M_5 = 426$). Given the composition of the modules in stage 1 and the limited score range of 0–15, this outcome represented the best approximation of a balanced number of observations over the entire design.

Finally, TMST B included four difficulty categories. However, the curve was only divided into three sections such that the two middle categories related to stage 1 were not visually distinguishable. Similar to TMST A, TMST B provided more accurate item difficulty estimates for easy and difficult items (i.e., categories 1 and 4) than the targeted design or the random condition. These two difficulty categories corresponded to stage 2 of TMST B, where the simulees' assignment was determined by their performance in stage 1.

Taken together, the item-level findings provided further insights into the efficiency of the different designs under the full knowledge condition. The targeted design outperformed the random condition owing to greater accuracy in item difficulty estimation for the easy and the difficult items. The two targeted multistage designs further improved the item difficulty estimation for the easy and the difficult items compared to the targeted design by means of the performance-based assignment of simulees to items in stage 2.

3.3.3.2 Conditions with limited knowledge

Figure 3.8 shows six charts. Each chart presents $\text{RMSE}(\hat{\beta})$ in relation to the true difficulty β for one of the three designs in combination with one of the two limited knowledge conditions. The random condition is included as the baseline in all six charts (black line). As an additional orientation, $\text{RMSE}(\hat{\beta})$ under the full knowledge condition is displayed for each design as a gray line.

The two charts in the first row of Figure 3.8 refer to the targeted design, where we distinguished three difficulty categories. Within each difficulty category, the relationship between $\text{RMSE}(\hat{\beta})$ and β was represented by a U-shaped curve. Differences in $\text{RMSE}(\hat{\beta})$ were found mainly for easy and difficult items, but not for the items in the intermediate difficulty range. Items that were assigned correctly to their corresponding difficulty category achieved similar $\text{RMSE}(\hat{\beta})$ values as under the full knowledge condition. However, easy or difficult items that were placed wrongly in the intermediate category showed $\text{RMSE}(\hat{\beta})$ values comparable to those under the random condition. Moreover, the placement of difficult items in the easy category and easy items in the difficult category resulted in $\text{RMSE}(\hat{\beta})$ values that exceeded the values under the random condition. The number of misplaced items and, therefore, the number of items with high $\text{RMSE}(\hat{\beta})$ values were lower under the medium knowledge condition (left chart) than under the low knowledge condition (right chart).

Figure 3.8. $RMSE(\hat{\beta})$ per item in relation to item difficulty for the three different designs under the medium (left) and low knowledge (right) conditions. The solid black line denotes the random condition (baseline) and the solid gray line the optimal condition specific to each chart. The white dots refer to the limited knowledge conditions (LK = low knowledge, MK = medium knowledge), and the related dotted lines cat1 to cat5 indicate the regressed trends of each difficulty category under the limited knowledge conditions.

The two charts in the second row of Figure 3.8 refer to TMST A, where we distinguished five difficulty categories. Within each difficulty category, the relationship between $RMSE(\hat{\beta})$

and β was again represented by a U-shaped curve. However, variations in $\text{RMSE}(\hat{\beta})$ were clearly larger for TMST A under the limited knowledge conditions than for the targeted design. Considerable differences emerged in all five difficulty categories. In particular, the $\text{RMSE}(\hat{\beta})$ values of the items in the intermediate difficulty category (i.e., category 3) were high. Even items of intermediate difficulty, which were assigned correctly to this category, showed $\text{RMSE}(\hat{\beta})$ values greater than those under the random condition.

These findings were related to large variations in the mean number of observations in stage 2 of TMST A (see Table 3.1). Under the limited knowledge conditions, a substantial number of observations shifted away from the intermediate modules in stage 2 toward the easy and the difficult modules, which resulted in considerably lower mean numbers of observations for the items in the intermediate modules (medium knowledge: $M_3 = 312$; low knowledge: $M_3 = 252$). This shift was caused by the shift in the mean module difficulty in stage 1, as shown in Figure 3.6. The relatively easy modules in stage 1 became more difficult due to misplaced items, which resulted in a lower percentage of simulees surpassing the cut-off score and reaching the intermediate modules in stage 2. Simultaneously, the relatively difficult modules became easier, such that a greater number of simulees surpassed the cut-off score and reached the difficult modules. Thus, the impaired overall efficiency of TMST A under the conditions of limited knowledge seems to be triggered mainly by the misplacement of a few items in stage 1 and the related impact on routing.

Finally, the two charts in the third row of Figure 3.8 refer to TMST B, where we distinguished four difficulty categories. In line with previous results, the relationship between $\text{RMSE}(\hat{\beta})$ and β was represented by a U-shaped curve for each difficulty category. Similar to the targeted design, differences in $\text{RMSE}(\hat{\beta})$ values were larger for easy and difficult items than for relatively easy or difficult items. Easy items assigned to the relatively easy category and difficult items assigned to the relatively difficult category were estimated more efficiently with TMST B than with the random condition. However, easy items assigned to one of the two difficult categories and difficult items assigned to one of the two easy categories resulted in $\text{RMSE}(\hat{\beta})$ values higher than those observed under the random condition. For items assigned to the appropriate difficulty category, again, we found similar $\text{RMSE}(\hat{\beta})$ values as under the full knowledge condition, with one exception: slightly higher $\text{RMSE}(\hat{\beta})$ values were reported for items in the difficult category under the medium knowledge condition. This difference was caused by a slightly lower mean number of observations for the items in the difficult category. Nevertheless, we concluded that limited knowledge resulted in similar effects for TMST B and the targeted design, namely, lower accuracy for misplaced items.

3.4 Discussion

In this paper, we introduced the concept of targeted multistage calibration designs for calibrating CAT item pools dedicated to multiple ability groups and, thus, allowing for the

measurement of ability over a broad range. We investigated the efficiency of this design type for calibrating items under the Rasch model by means of simulations. As expected, the two targeted multistage calibration designs were more efficient in calibrating the item pool than the targeted design or the random condition, given that complete knowledge about item difficulty was available during construction of the calibration designs. The reported improvements in overall RMSE corresponded to statistically significant, as well as practically relevant, efficiency gains in terms of the number of students. Namely, additional time and financial resources would be required to recruit up to 200 additional students for calibrating the items with the traditional targeted design in order to compensate for the gain of the targeted multistage calibration designs.

Differences in the accuracy of item difficulty between the targeted design and the two targeted multistage designs were found especially for easy and difficult items. Reliable item difficulty estimates for these groups of items are highly relevant in the context of CAT, where each student is assessed based on a different subset of items from the overall item pool (see also van der Linden & Glas, 2000). The performance-based routing in the targeted multistage designs allowed for identification of low and high performing students including students with extreme and, thus, rare abilities within both samples, and for assigning them specifically to the easy and difficult items. Therefore, the multistage procedure improved the match between item difficulty and ability for those items that clearly differed from the mean abilities of the two samples. In practice, such improved alignment between item difficulty and student ability is not only beneficial for calibrating items, but it also provides more reliable estimates of ability. Moreover, it might prevent low and high ability students from getting discouraged or bored by items that do not fit their ability level (Asseburg & Frey, 2013). Sustaining student motivation during test taking is, in turn, an important basis for producing reliable calibration outcomes (e.g., Finn, 2015; Mittelhaeuser, Béguin, & Sijtsma, 2015; Zwitser & Maris, 2015).

However, the construction of efficient calibration designs requires knowledge not only about the abilities of calibration samples but also about item difficulty. Unfortunately, empirical knowledge about item difficulty is often not available in practical settings prior to calibration. Instead, difficulty ratings from experts are used as approximations of real item difficulty and as a basis for distributing items within a design. Several studies have reported medium-to-high correlations between such ratings and real item difficulties (e.g., Bejar, 1983; Hambleton & Jirka, 2006; Sydorenko, 2011; Wauters et al., 2012). Therefore, we investigated the extent to which misclassification of items in terms of their difficulty could impair the efficiency of the different calibration designs considered in this study. In line with our expectations, limited knowledge about item difficulty was related to a considerable decrease in the efficiency of the traditional targeted design as well as of the two targeted multistage calibration designs.

Furthermore, we hypothesized that targeted multistage calibration designs are more vulnerable to misplaced items than targeted calibration designs because knowledge about item difficulty is crucial for specifying appropriate routing rules. This hypothesis was only partly

supported by the results of our simulation study. In line with our expectations, we found a severe loss of efficiency for one of the two targeted multistage calibration designs under the conditions of limited knowledge and misplaced items. The efficiency loss of TMST A was caused by a mismatch between the specified routing rules based on predicted item difficulties and the real mean difficulties of the routing modules. This mismatch led to inefficient distribution of simulees over the modules of stage 2, considerably low mean numbers of observations for certain items, and, finally, relatively inaccurate item difficulty estimates of the affected items.

However, the second multistage calibration design was more robust for limited knowledge and misplaced items. Contrary to our expectations, TMST B outperformed the targeted design, regardless of the amount of available knowledge, even though the advantage of TMST B over the targeted design was greater under the full knowledge condition. The specific combination of routing module composition and routing rules in TMST B managed to compensate for changes in routing due to misplaced items. Nevertheless, the minor differences in the mean numbers of observations under the medium knowledge condition suggest that overlapping routing paths do not guarantee a stable targeted multistage calibration design. In our study, the mean difficulty of the routing modules shifted toward the mean of the item pool under the limited knowledge conditions. However, experts might also systematically under- or overestimate the difficulty of all items, leading to under- or overrepresentation of the numbers of observations for the easy and the difficult modules in stage 2. In contrast, targeted calibration designs provide full control over the assignment of students to test booklets and, thereby, over the number of observations per item—independent of the amount of knowledge about the difficulty of the items.

As in any simulation study, our study covered a limited set of conditions, which limits generalizability of the findings in some ways (Davey, Nering, & Thompson, 1997; Feinberg & Rubright, 2016). First, we investigated the efficiency of item calibration under the Rasch model. Different results might be found if the response patterns do not fit the Rasch model perfectly as they did in our simulation study. Our results could serve as a starting point for investigating the efficiency of targeted multistage calibration designs in combination with more complex IRT models. Based on previous studies, we hypothesize that targeted multistage designs allow for the efficient estimation of item difficulty parameters independent of the IRT model (Berger, 1991; Stocking, 1988). However, further research is needed for investigating the extent to which the limited ability variation within each subsample could impair efficient estimation of other item parameters, such as discrimination or guessing.

Second, we focused on improving the match between item difficulty and student ability for enhancing calibration efficiency. However, Eggen and Verhelst (2006), as well as Verschoor (2010), found that the efficiency of incomplete calibration designs also depends on the strength of the links between the modules within a design so that designs with a larger number of linking items provide more accurate estimates of item difficulty. However, given the

test length of 30 items and the use of MML estimation procedures, we expect only small improvements through such adaptations.

A third limitation of our study is that the size of the samples was relatively small. In alignment to a practical setting, we included 2,600 simulees in each simulation run. An enhancement of sample size would significantly increase the accuracy of item difficulty estimates under all conditions and decrease the practical relevance of efficiency gain through better alignment of item difficulty and student ability (Wright, 1977, p. 105). On the other hand, recruiting samples that are large enough for item calibration is often challenging, which underlines the practical relevance of this constraint. Moreover, it would be interesting to investigate in further studies whether the current results are replicable with more complex designs, different item pools, and different samples. Based on our findings, we hypothesize that targeted multistage designs will always outperform targeted designs under optimal conditions. However, we assume that the risk of efficiency loss in targeted multistage calibration designs depends not only on the amount of available knowledge about item difficulty, but also on the particular combination of design specifications, item pool, and sample. Thus, further studies should analyze the factors that can contribute to the stability of targeted multistage calibration designs. Furthermore, it would be interesting to examine whether step-by-step adaption of the routing rules or of the composition of modules during calibration (cf. Ali & Chang, 2014; Kingsbury, 2009) could reduce the risk of efficiency loss due to imbalanced distribution of the number of observations per item.

Efficient calibration designs are crucial in practice because it is often difficult to recruit sufficient numbers of students for calibrating CAT item pools. With our simulation study, we extended previous research on the efficiency of calibration designs by introducing targeted multistage calibration design as a new design type and by investigating a practically relevant constraint, namely, the influence of limited knowledge during test development. We conclude from our findings that targeted multistage calibration designs are an option for enhancing the efficiency of calibration under the condition that reliable knowledge about item difficulty is available. However, these theoretically superior designs come at a certain price in practice if a priori knowledge about item difficulty is limited. Therefore, we recommend practitioners to use targeted designs instead of targeted multistage designs for calibration whenever there are doubts about the accuracy of predicted item difficulty.

3.5 References

- Ali, U. S., & Chang, H.-H. (2014). *An item-driven adaptive design for calibrating pretest items* (ETS Research Reports Series No. RR-14-38). Princeton, NJ: Educational Testing Service.
- Angelone, D., Keller, F., & Moser, U. (2013). *Entwicklung schulischer Leistungen während der obligatorischen Schulzeit: Bericht zur vierten Zürcher Lernstandserhebung zuhanden der Bildungsdirektion des Kantons Zürich* [Development of school performance during compulsory school: Report on the fourth assessment for the attention of the Zurich department of education]. Zürich: Institut für Bildungsevaluation. Retrieved from Institut für Bildungsevaluation website: http://www.bi.zh.ch/internet/bildungsdirektion/de/unsere_direktion/bildungsplanung/arbeitenundprojekte/lernstand/_jcr_content/contentPar/downloadlist_3/downloaditems/347_1408695301431.spooler.download.1408694967389.pdf/Lernstand9Klassen_Wissenschaftlicher+Bericht+2013.pdf
- Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, 55, 92–104.
- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303–310. doi:10.1177/014662168300700306
- Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, 15, 293–306. doi:10.1177/014662169101500310
- Berger, S., Moser, U., Verschoor, A. J., & Eggen, T. J. H. M. (2015). *Development of an online item bank for adaptive formative assessment*. Paper presented at the AEA-Europe Conference. AEA-Europe, Glasgow, Scotland.
- Berger, S., Oostlander, J., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2015). *Development of multistage tests based on teacher ratings*. Paper presented at the conference of the International Association for Computerized Adaptive Testing, Cambridge, England.
- Berger, S., Verschoor, A. J., Moser, U., & Oostlander, J. (2014). *Educational assessment in Northwestern Switzerland: Psychometrical challenges in linking standardized tests to an online item bank for formative assessment*. Paper presented at the AEA-Europe Conference, Tallinn, Estonia.
- Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series No. 97-4). Iowa City, IA: ACT. Retrieved from ACT website: http://www.act.org/content/dam/act/unsecured/documents/ACT_RR97-04.pdf
- Eggen, T. J. H. M., & Verhelst, N. D. (2006). Loss of information in estimating item parameters in incomplete designs. *Psychometrika*, 71, 303–322. doi:10.1007/s11336-004-1205-6
- Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica*, 32, 107–132.

- Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35, 36–49. doi:10.1111/emip.12111
- Finn, B. (2015). *Measuring motivation in low-stakes assessments* (ETS Research Reports Series No. RR-15-19). Princeton, NJ: Educational Testing Service.
- Glas, C. A. W. (2010). MIRT: Multidimensional item response theory. Enschede: University of Twente. Retrieved from https://www.utwente.nl/nl/bms/omd/Medewerkers/temp_test/mirt-manual.pdf
- Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, 35, 83–88. doi:10.1016/j.stueduc.2009.10.006
- Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Lawrence Erlbaum Associates.
- Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from <http://www.iacat.org/sites/default/files/biblio/cat09kingsbury.pdf>
- Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York, NY: Springer.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
- Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, 11, 1–20.
- Mislevy, R. J., & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Reports Series No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
- Mittelhaeuser, M.-A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, 52, 339–358. doi:10.1111/jedm.12080
- Northwest Evaluation Association. (2011). *Technical manual for Measure of Academic Progress & Measure of Academic Progress for Primary Grades*. Portland, Oregon. Retrieved from <https://www.richland2.org/RichlandDistrict/media/Richland-District/AdvancED/Standard%205/5.1/5-1-NWEA-Technical-Manual-for-MAP-and-MPG.pdf>
- O'Neill, T. R., Marks, C. M., & Reynolds, M. (2005). Re-evaluating the NCLEX-RN passing standard. *Journal of Nursing Measurement*, 13, 147–165. doi:10.1891/jnum.2005.13.2.147

- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.
- Renaissance Learning. (2015). *STAR Math: Technical manual*. Wisconsin Rapids, WI: Renaissance Learning.
- Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion* [Textbook test theory, test construction]. Bern: Verlag Hans Huber.
- Stocking, M. L. (1988). *Specifying optimum examinees for item parameter estimation in item response theory* (ETS Research Reports Series No. RR-88-57-ONR). Princeton, NJ: Educational Testing Service.
- Strobl, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* [The Rasch model: A coherent introduction for students and practitioners] (2nd ed.). Mering: Rainer Hampp Verlag.
- Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, 8, 34–52. doi:10.1080/15434303.2010.536924
- Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement*, 12, 53–67. doi:10.1177/014662168801200106
- van der Linden, W. J. (1988). *Optimizing incomplete sample designs for item response model parameters* (Research Report No. 88-5). Enschede: University of Twente.
- van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, 13, 35–53. doi:10.1207/s15324818ame1301_2
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.
- Verschoor, A. J. (2010). *Optimal calibration designs for computerized adaptive testing*. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, the Netherlands.
- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wauters, K., Desmet, P., & van den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58, 1183–1193. doi:10.1016/j.compedu.2011.11.020
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492. doi:10.1177/014662168200600408
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361–375. doi:10.1111/j.1745-3984.1984.tb01040.x
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97–116. doi:10.1111/j.1745-3984.1977.tb00031.x

- Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.
- Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer.
- Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, 80, 65–84. doi:10.1007/s11336-013-9369-6

Chapter 4. Improvement of Measurement Efficiency in Multistage Tests by Targeted Assignment⁴

Abstract

A good match between item difficulty and student ability ensures efficient measurement and prevents students from becoming discouraged or bored by test items that are too easy or too difficult. Targeted test designs consider ability-related background variables to assign students to matching test forms. However, these designs do not consider that students might significantly differ in ability within the resulting groups. In contrast, multistage test designs consider students' performance during test taking to route them to the most informative modules. Yet, multistage test designs usually include one starting module of moderate difficulty in the first stage, which does not account for differences in ability. In this paper, we investigated whether measurement efficiency can be improved by targeted multistage test designs that consider ability-related background information for a targeted assignment at the beginning of the test and performance during test taking for selecting matching test modules. By means of simulations, we compared the efficiency of the traditional targeted test design, the multistage test (MST) design, and the targeted multistage test (TMST) design for estimating student ability. Furthermore, we analyzed the extent to which the efficiency of the different designs depends on the correlation between the ability-related background variable and the true ability, students' ability level and their categorization into an ability group, and the length of the starting module. The results indicated that TMST designs were generally more efficient for estimating student ability than targeted test designs and MST designs, especially if the ability-related background variable correlated high with and, thus, was a good indicator of, students' true ability. Furthermore, TMST designs were particularly efficient in estimating abilities for low- and high-ability students within a given population. Finally, very long starting modules resulted in less efficient estimation of low and high abilities than shorter starting modules. However, this finding was more prominent for MST than for TMST designs. In conclusion, TMST designs are recommended for assessing students from a wide ability distribution if a reliable ability-related background variable is available.

⁴ This chapter is a minor revision of Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education*, 4, 1. doi: 10.3389/feduc.2019.00001

4.1 Introduction

Understanding the purpose of a test and its target population is essential for determining an appropriate test design. A test for classifying students or a test targeted to a very specific population needs to measure ability most accurately close to specific points along the ability scale. Therefore, an appropriate test design includes mainly items that provide a large amount of information at these specific points. In contrast, tests aiming to assess the ability level of each student of a diverse target population—such as formative assessments—require test designs that provide accurate results over a wide ability range. Often, a single linear test, including items of varying difficulties, is not appropriate in such a context because each student is confronted with a relatively high number of items that do not match his or her ability level. This in turn results in a low measurement efficiency, and it could also impair students' motivation during test taking (Lord, 1980).

In general, two different approaches are used to account for a broad variation in ability via targeted assignment of items of varying difficulty (e.g., Mislevy & Wu, 1996). On one hand, we can consider preliminary information about students' ability in order to assign students to matching test forms. In a school context, preliminary information about the ability level of a student is often available prior to testing. Students are assessed in various tests by their teachers and are marked accordingly. In addition, students are assigned to different grades in school and—especially in the secondary school level—even to different school types or performances groups. This information can be used to divide the target population into ability groups and to determine matching test forms for each group. However, the disadvantage of this approach is that ability-related background variables are only an approximation of the students' true ability. As a result, some students might differ greatly from the group mean and, hence, from the target ability level of the test.

On the other hand, we can assign students to targeted items or item sets step by step based on their performance during test taking. In other words, students with a good performance automatically receive more difficult items, which allows them to show their full potential, whereas students with a lower performance automatically receive easier items. This is the basic idea of computer adaptive tests (CATs; van der Linden & Glas, 2010; Wainer, 2000a) or multistage tests (MSTs; Yan, von Davier, & Lewis, 2014; see also Luecht & Nungester, 1998). Whereas CATs select each item based on the students' performance, MSTs select predefined sets of items (i.e., test modules). MSTs have become more and more popular in recent years because they combine the advantages of adaptivity with the advantages of traditional linear tests. Compared to CATs, MSTs are easier to develop and implement, and they allow test developers as well as test takers to review items within a given test module (Yan, Lewis, & von Davier, 2014). Therefore, in our paper, we focus on the MST design as representative of a performance-based item assignment. A disadvantage of the MST design is that performance-based assignment is only possible after students have answered an initial untargeted set of items. This routing or starting module usually consists of items of moderate difficulty, and the module

is administered to all students independent of their true ability (Hendrickson, 2007; Yan, Lewis et al., 2014; Zenisky & Hambleton, 2014). Hence, the starting module is most likely too easy for some of the high-ability students or too difficult for some of the low-ability students within the target population, that is, for students whose ability largely differs from the population's mean ability.

Taken together, neither targeted assignments based on ability-related background variables nor assignments based on performance ensure that all students receive items that completely match their true ability. An inappropriate test form or module measures the ability of the concerned students less efficiently (Lord, 1980) and might even impair the students' performance owing to decreased motivation or excessive demand (e.g., Asseburg & Frey, 2013; Wise, 2014). Therefore, we extended previous research on efficient test designs by investigating whether measurement efficiency can be improved by combining targeted assignment based on ability-related background variables with performance-based assignments of test modules of varying difficulty. To this end, we introduced the targeted multistage test (TMST) design as a new design type that considers both preliminary ability-related background information for a targeted assignment at the beginning of the test and performance during test taking for selecting matching test modules. By means of simulations, we compared the efficiency of traditional targeted test designs, MST designs, and TMST designs for estimating student ability under the Rasch model (Rasch, 1960). In this way, we not only provided new insights on the measurement efficiency of the combination of targeted and performance-based testing (i.e., the TMST design) but also allowed for comparing the efficiency of targeted test designs with those of MST designs, and, thus, for contrasting test assignment based on ability-related background variables versus test assignment based on performance. In addition, we explored how different groups of students—namely, students whose abilities differ from the mean ability of the total population or of their performance group—benefit from or are disadvantaged by the three different design types. Furthermore, we investigated the extent to which the efficiency of the different designs depends on the degree of ability distribution overlap of the groups and the starting module length.

4.1.1 Efficient Measurement Based on Item Response Theory

As stated by Lord (1980), “an examinee is measured most effectively when the test items are neither too difficult nor too easy for him” (p. 150), implying that, ideally, students differing in ability should be assessed with different test booklets or item sets of varying difficulty to efficiently measure each student's ability (Weiss, 1982). Nonetheless, the resulting test scores should be directly comparable, even though students worked on different items included in easy, moderate, or difficult test booklets. Item response theory (IRT) is a powerful measurement approach that fulfills this requirement (Kolen & Brennan, 2014). IRT refers to a family of models that express the probability of a student solving an item correctly as a function of student

ability and item difficulty (Lord, 1980). For the Rasch model—the simplest unidimensional IRT model—the probability of a student solving a specific item correctly is given by

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} = p_{ij}, \quad (4.1)$$

where θ_i represents the ability of student i , and β_j represents the difficulty of item j (Rasch, 1960; Rost, 2004). Given this relationship, student ability and the related standard error can be estimated by various maximum likelihood procedures.

Generally, the standard error of the estimated ability of student i can be approximated as (Rost, 2004)

$$SE(\hat{\theta}_i) \sim \sqrt{\frac{1}{\sum_{j=1}^k p_{ij}(1 - p_{ij})}}, \quad (4.2)$$

where p_{ij} corresponds to the probability that student i answers item j correctly, as defined in equation 4.1. On one hand, we can infer from equation 4.2 that k , which refers to the total number of items answered by a student, is a crucial factor for estimating ability accurately according to the Rasch model. On the other hand, we can deduce from equation 4.2 that the accuracy of the ability estimation depends on the relationship of the student's ability and the difficulty of the items in a given test, as discussed earlier. Student ability can be estimated most accurately under the Rasch model if the difficulty of the items in the test corresponds to the ability of the student (Lord, 1980; Rost, 2004).

In practical settings, test length (i.e., the number of items) is often predefined according to the available testing time. This leaves optimization of the relationship between student ability and item difficulty as the only option for improving the accuracy of ability estimation and thus for enhancing the efficiency of a test. As a consequence, the idea of targeted testing arose to enhance measurement efficiency within a given testing time (e.g., Lord, 1971a, 1971b, 1971c, 1980; for a historical overview, see, e.g., Bejar, 2014; Wainer, 2000b).

4.1.2 Designs for Targeted Testing

As discussed earlier, two different approaches are used for assigning targeted items to students of varying ability levels: (1) assign students by means of ability-related background variables to test booklets of varying mean difficulties as in traditional targeted test designs (Mislevy & Wu, 1996), or (2) adaptively assign students based on their performance in a first test part to easy, moderate, or difficult subsequent test parts as in MSTs (Yan, von Davier et al., 2014). In the following, we will describe both test designs in more detail. In addition, we will elaborate on how to combine these two approaches to further enhance the targeting and thus measurement efficiency.

4.1.2.1 *Traditional targeted test designs*

Targeted test designs consist of several test booklets of varying mean difficulties, and students are assigned to the most informative booklet via ability-related background variables, such as school grade, exam grade, courses taken, or performance-related school type (Mislevy & Wu, 1996). Students can be classified based on such background variables before test administration. The resulting ability groups are more homogeneous in terms of abilities than the overall sample. Subsequently, low-ability groups are assigned to easy test booklets, whereas high-ability groups are assigned to difficult test booklets. In an IRT framework, common items (i.e., linking items), or a joint underlying calibrated item bank, ensure that the outcomes from the different test booklets can be represented on a single reporting scale for direct comparison (Kolen & Brennan, 2014). Unfortunately, to our knowledge, there are no empirical studies investigating the efficiency of targeted test designs.

A disadvantage of targeted test designs is that ability-related background variables are only an approximation of the student's true ability. Depending on the reliability of the ability-related background variable, students might still substantially differ in ability within one ability group. For performance groups at the secondary school level as a potential ability-related background variable, several studies found a large overlap between the abilities of students assigned to different secondary school types (e.g., Angelone, Keller, & Moser, 2013; Baumert, Stanat, & Watermann, 2006). For average exam grades, Moser, Buff, Angelone, and Hollenweger (2011) reported correlations of $r = .69$ and $r = .70$ between exam grade averages at the end of primary school in language and mathematics and achievement tests measuring abilities within the same subjects. For high school, the technical report of the Scholastic Assessment Test (SAT) referred to correlations around $r = .50$ between SAT scores and high school grade point averages for both language and mathematics (College Board, 2017). These findings support the assumption that an overlap occurs between the ability distributions of different groups, such that some students from the lower ability group outperform the students with the lowest abilities of the higher ability group. Students who largely differ from the mean ability of the group get an inappropriate test form, which measures their abilities less efficiently and might impair their performance due to decreased motivation or excessive demand (e.g., Asseburg & Frey, 2013; Wise, 2014). Pohl (2013) raised similar concerns for longitudinal MSTs, where pretest results serve as a basis for routing to different test forms at later measurement occasions. In a simulation study, she showed that correlations below $r = .70$ between the pretest results and the ability at the later measurement occasion can result in a significant number of heavily misallocated test forms. Furthermore, the study showed that the efficiency gains of longitudinal MSTs increased as the correlations increased when compared to conventional tests, especially for low-ability and high-ability students.

4.1.2.2 *MST designs*

MST designs route students based on their performance to the most informative modules. Usually, they rely on IRT as a methodological framework for routing as well as for reporting

standardized scores for all possible paths within the MST design (Weissman, 2014; Yan, Lewis et al., 2014). MSTs consist of two or more test parts (i.e., stages), including multiple modules of varying difficulty (Hendrickson, 2007; Yan, von Davier et al., 2014; Zenisky, Hambleton, & Luecht, 2010). At the end of each stage, predefined routing rules determine—based on the student’s preliminary performance (i.e., based on preliminary ability estimates)—whether a student is assigned to an easier or more difficult module in the subsequent stage. Hence, the module assignment is based on objective cutoff scores and, more importantly, on a preliminary estimate of the same latent construct as that measured by the test as a whole. Several studies have highlighted the improved measurement efficiency of MSTs compared to linear tests for measuring ability over a wide range (for a general overview, see Hendrickson, 2007; Yan, Lewis et al., 2014; Zenisky & Hambleton, 2014). However, studies that compare the efficiency of MSTs with the efficiency of traditional targeted tests are missing to our knowledge.

From their simulation study, Kim and Plake (1993, p. 17) concluded that the statistical characteristics of the starting module are significant determinants of an accurate ability estimation in MSTs. This results from the fact that MST designs usually do not differ between ability groups in the first stage but start with a single general starting module for the entire population, which mostly includes items of moderate difficulty (Hendrickson, 2007; Yan, Lewis et al., 2014; Zenisky & Hambleton, 2014). The purpose of the starting module is twofold (Lord, 1971b; Verschoor & Eggen, 2014): (1) it collects information for assigning students to the most informative module in the subsequent stage, and (2) it is part of the test itself with the aim to measure student ability as accurately as possible. A general moderately difficult starting module well fulfills the first but not necessarily the second purpose because it measures low ability and high ability in students less efficiently than the subsequent targeted modules. As already elaborated on for traditional targeted designs, low-ability and high-ability students might also be intimidated, frustrated, or bored due to the overload or underload, respectively, through the suboptimal MST modules (e.g., Asseburg & Frey, 2013; Wise, 2014).

The degree of discrimination through a suboptimal starting module depends on the length of the starting module in relation to the total test length (Verschoor & Eggen, 2014). The shorter the starting module is, the fewer mismatched items are administered to low-ability and high-ability students in the first stage and the more items remain for targeted testing in subsequent stages. At the same time, a short starting module involves the risk of routing errors due to low measurement precision, which, in turn, results in higher estimation errors of the overall ability estimates (Kim & Plake, 1993). Thus, the length of the starting module is an important factor that needs to be considered when analyzing the efficiency of MST designs.

4.1.2.3 *TMST designs*

TMST designs refer to MST designs with more than one starting module, meaning they are a hybrid of targeted and MST designs. As in traditional targeted tests, based on ability-related background variables, students are assigned in stage 1 to the most informative of multiple starting modules of varying difficulty. Due to the lack of performance-based ability estimates,

ability-related background variables are the best approximation of student ability at this stage of the test. In stage 2 and all subsequent stages, students are routed based on preliminary ability estimates to the most informative module as in traditional MSTs. By combining ability-related background variables and preliminary ability estimates for module selection, the TMST design aims to improve the match between student ability and item or module difficulty in the first stage compared to the MST design for all students whose abilities are well represented by the ability-related background variable. The performance-based routing after stage 1 aims to homogenize the ability groups further and correct possible misallocations of students whose abilities are poorly represented by the ability-related background variable.

TMST designs or MST designs with more than one starting module are very rare in practice (Hendrickson, 2007; Zenisky & Hambleton, 2014). Practical examples include the Massachusetts Adult Proficiency Tests (MAPT) for reading and mathematics (Sireci et al., 2008; Zenisky, Sireci, Martone, Baldwin, & Lam, 2009). In both tests, students are assigned based on the teachers' judgement of their so-called "Educational Functioning Levels" or based on previous test outcomes to the most informative module in stage 1. The MAPT for reading has also been evaluated in a research paper by Crotts, Zenisky, Sireci, and Li (2013). However, neither the technical manuals nor the related research paper have explicitly discussed or empirically analyzed the added value of having more than one starting module. In addition, we are not aware of any previous research that investigated the efficiency of TMSTs based on simulations.

4.1.3 The Present Study

The purpose of this study is to investigate whether TMST designs achieve more accurate and, therefore, more efficient ability estimates than traditional targeted test designs or MST designs with one starting module. We hypothesize that TMST designs enhance the accuracy of ability estimates compared to targeted test designs—similar to MST designs—through performance-based routing after stage 1 and that TMST designs outperform MST designs due to the targeted assignment of modules to different ability groups in stage 1.

In addition to this general research question, we explore the efficiency of TMST designs in more detail from three different perspectives. First, we investigate the extent to which the efficiency gain through TMST designs depends on the correlation between the ability-related background variable and students' true ability. Following Pohl (2013), we hypothesize that the efficiency gain is more prominent if the correlation between the ability-related background variable and the true ability is high, or, in other words, if the distance in mean ability between the resulting groups is large, and the overlap between their ability distributions is small.

Second, we examine the extent to which different ability groups profit or are disadvantaged by TMST designs compared to targeted and MST designs. In comparison with targeted designs, we expect an efficiency gain through TMST designs for students whose

abilities are poorly described by the ability-related background variable (i.e., students with abilities that are deviant from the target ability group mean) because the performance-based routing after stage 1 allows for correcting the misallocation of these students to suboptimal starting modules. In comparison with MST designs, we expect an efficiency gain through TMST designs for students whose abilities differ from the mean of the total population (i.e., low-ability and high-ability students) and whose abilities are well described by the ability-related background variable (i.e., students with abilities close to the target ability group mean). In contrast, we expect an efficiency loss through TMST designs compared to MST designs for students with medium abilities who are classified by mistake into a low-ability or high-ability group because their ability is poorly described by the background variable.

Third, we explore the extent to which the efficiency gain through TMST designs depends on the length of the starting module compared to the total test length. As elaborated by Verschoor and Eggen (2014), the length of the starting module is an important factor in optimizing MST designs. However, clear general recommendations for distributing items between the starting and follow-up modules are missing for MST designs. Furthermore, the relationship between efficiency gain and starting module length becomes even more complex in the TMST design due to module assignment based on a combination of ability-related background variables and performance. Therefore, we explore this relationship in our study without stating a priori hypotheses.

To address these research questions and hypotheses, we conducted a simulation study, in which we varied the test design, the correlation between ability and the ability-related background variable, and the length of the starting module in relation to the total test length.

4.2 Method

4.2.1 Ability Distributions and Population Distribution Conditions

For each simulation run, we drew three samples of 10,000 simulees randomly from three normal distributions to simulate students from three overlapping ability groups, where each group represented one of three levels of the ability-related background variable. To vary the correlation between the ability-related background variable and students' true ability, we manipulated the degree of overlap between the three groups through three population distribution conditions, as described in Table 4.1. In particular, we varied the difference d between the mean abilities of the groups. For the narrow condition, which reflected a low correlation between the ability-related background variable and students' true ability, the difference between the mean abilities was set to $d = 0.5$; for the medium condition to $d = 1.0$; and for the wide condition, which reflected a high correlation between the ability-related background variable and students' true ability, to $d = 1.5$. For all three distribution conditions, we assumed a standard deviation (SD) of 1 for each group. Besides the distribution parameters

of the ability groups, Table 4.1 also includes the distribution parameters μ and σ of the related mixture populations given by

$$\mu = \sum_i^n w_i \mu_i \quad \text{and} \quad \sigma = \sqrt{\sum_i^n w_i (\mu_i^2 + \sigma_i^2) - \mu^2} \quad (4.3)$$

where μ_i and σ_i refer to the mean and the *SD* of the $n = 3$ ability groups, respectively, and w_i refers to the relative weight of the distributions with $\sum_i^n w_i = 1$ (e.g., Frühwirth-Schnatter, 2006, p. 11). Furthermore, Table 4.1 displays Spearman's rank correlation ρ between the students' group classification (i.e., the grouping based on the ability-related background variable) and their true ability θ for each distribution condition. Following Pohl (2013), the narrow condition represented a low correlation between the ability-related background variable and the true ability ($\rho = .38$) that is clearly below the recommended minimal correlation of $r = .70$, the medium condition represented a medium correlation ($\rho = .65$) similar to the recommended minimal correlation, and the wide condition represented a high correlation ($\rho = .79$).

Table 4.1. Parameters of the Normal Distributions per Ability Distribution Condition

Distribution	Low AG		Medium AG		High AG		Mixture Population		Correlation AG and θ
	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3	μ	σ	ρ
Narrow	−0.5	1.0	0.0	1.0	0.5	1.0	0.0	1.08	.38
Medium	−1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.29	.65
Wide	−1.5	1.0	0.0	1.0	1.5	1.0	0.0	1.58	.79

Note. AG = ability group.

4.2.2 Test Designs

The simulation study included four different test designs: (1) a linear design, which served as the baseline condition; (2) a targeted design; (3) an MST design; and (4) a TMST design. For all four design conditions, test length was constrained to 30 items, which refers to the number of items that we expect students to answer within one school lesson (i.e., 45 minutes). We treated item difficulty parameters as known and did not specify any overlap between the different modules within a test design.

4.2.2.1 Linear design

In the linear design condition, all simulees were assigned to the same 30 items. To accommodate for the different ability groups, we combined items targeted to the mean ability of all three groups in this test (i.e., 10 easy, 10 moderate, and 10 difficult items).

4.2.2.2 Targeted design

In the targeted test design condition, we distinguished three different linear modules with 30 items each, which were targeted to the mean ability of the three ability groups. Thus, simulees from each group were assigned to their dedicated test module as indicated on the left in Figure 4.1.

Figure 4.1. Illustration of the targeted, MST, and TMST designs. y = ability-related background variable (e.g., grades in school); E = easy module, M = moderate module, D = difficult module; E_1 , M_1 , and D_1 = starting modules in stage 1; E_{23} , M_{23} , and D_{23} = modules of stages 2 and 3.

4.2.2.3 MST design

The MST was a 1-3-3 design consisting of three stages and seven modules as indicated in the middle of Figure 4.1. Simulees from all three ability groups were assigned to a single starting module of moderate difficulty in stage 1. In stage 2 and stage 3, simulees were assigned to an easy, moderate, or difficult module depending on their performance in the preceding stage. The two routing cutoff scores, c_1 and c_2 , as indicated in Table 4.2, were defined at the percentiles P_{33} and P_{66} of the mixture population to guide an equal number of simulees to each module of stages 2 and 3. The difficulties of the items within the different modules were varied depending on the target difficulty of the module (i.e., easy, moderate, difficult) and the ability distribution condition (i.e., narrow, medium, wide), as described in more detail in the next section of this paper.

4.2.2.4 TMST design

The TMST was a 3-3-3 design consisting of three stages with three modules each. In the first stage, simulees from each ability group were assigned to their dedicated starting module, as shown on the right in Figure 4.1. Simulees from the low-ability group were assigned to the easy module, simulees from the medium-ability group were assigned to the moderate module, and simulees from the high-ability group were assigned to the difficult module. As in the MST

design, the routing cutoff scores in the TMST design were defined based on the percentiles P_{33} and P_{66} of the mixture population (see Table 4.2), and simulees were guided to the easy, moderate, or difficult modules in stages 2 and 3 based on their performance in the preceding stage. The difficulties of the items within the different modules were varied depending on the target ability group (i.e., low-, medium-, high-ability) for modules in stage 1, the target difficulty of the module (i.e., easy, moderate, difficult) for modules in stages 2 and 3, and the ability distribution condition (i.e., narrow, medium, wide).

Table 4.2. Routing Cutoff Scores and Item Difficulty Parameters of Easy and Difficult Modules in Stage 1 and in Stage 2/3 per Ability Distribution Condition

Distributions	Routing		Item Difficulty S1 ^a		Item Difficulty S2/S3	
	c_1	c_2	β_{E1}	β_{D1}	β_{E23}	β_{D23}
Narrow	-0.47	0.47	-0.50	0.50	-1.20	1.20
Medium	-0.56	0.56	-1.00	1.00	-1.40	1.40
Wide	-0.68	0.68	-1.50	1.50	-1.70	1.70

Note. c_1 = cutoff score 1 for routing between the easy and moderate modules, c_2 = cutoff score 2 for routing between the moderate and difficult modules, β_{E1} = difficulty parameters in the easy modules in stage 1, β_{D1} = difficulty parameters in the difficult modules in stage 1, β_{E23} = difficulty parameters in the easy modules in stages 2 and 3, β_{D23} = difficulty parameters in the difficult modules in stages 2 and 3.

^aOnly applicable to the targeted and TMST designs.

4.2.2.5 Starting module length

For the MST and TMST conditions, we distinguished four different variations of starting module lengths corresponding to $\frac{1}{5}$, $\frac{1}{4}$, $\frac{1}{3}$, and $\frac{1}{2}$ of the total test length. In all four conditions, the starting modules contained an even number of items to allow for dividing the remaining items equally over the two subsequent stages. The shortest starting modules consisted of 6 items with a remaining 12 items each for stages 2 and 3, the second condition consisted of 8 items in stage 1 with a remaining 11 items each for stages 2 and 3, the third condition consisted of 10 items in each of the three stages and represented the main condition regarding starting module length, and the fourth condition consisted of 16 items in stage 1 with a remaining 7 items each for stages 2 and 3. The length of the starting modules was only varied in combination with the medium ability distribution condition. The narrow and the wide ability distribution conditions were only combined with the main length condition, which included 10 items in each of the three stages (i.e., $\frac{1}{3}$ of the total test length).

4.2.3 Item Pools

For each simulation condition, we generated a dedicated set of dichotomous Rasch items targeted to the specific characteristics of the test design and the ability distributions. To simplify matters and to facilitate comparing different conditions, we specified homogeneous item difficulty parameters within each module according to the following rules: The difficulty parameters of the items of all modules of moderate difficulty were set to $\beta = 0$ for all stages and all conditions. For stage 1 and for the modules of the targeted design, we specified the item difficulty parameters of the easy and difficult modules based on the mean of the target population. For the easy and difficult modules in stages 2 and 3, we specified the item difficulty parameters based on the expected mean ability of the assigned subgroup created by the routing under the assumption of no routing errors, that is, based on the rounded mean of the three truncated normal distributions resulting from the two cutoff scores (Barr & Sherrill, 1999). Table 4.2 provides an overview of the resulting item difficulties for the three stages combined with the three ability distribution conditions.

4.2.4 Item Response Generation and Ability Estimation

Altogether, the simulation study included 18 different conditions, which are summarized in Table 4.3. Of these conditions, 12 resulted from the combination of the three population distributions (i.e., narrow, medium, and wide) with the four test designs (i.e., linear, targeted, MST, and TMST). Furthermore, for the MST and TMST designs under the medium distribution condition, we investigated three additional variations of module length (i.e., starting module length equal to $\frac{1}{5}$, $\frac{1}{4}$, and $\frac{1}{2}$ of the total test length), which resulted in six additional conditions. For each condition, we generated 1,000 data sets with data related to 30,000 simulees.⁵ For each simulee i , each data set included the true ability θ_i , the assigned items, the estimated ability $\hat{\theta}_i$, and its standard error $SE(\hat{\theta}_i)$. To estimate student ability, we used the weighted maximum likelihood (WML) method proposed by Warm (1989).

Table 4.3. Overview of the 18 Simulation Conditions

	Linear	Targeted	MST				TMST			
Distribution			1/5	1/4	1/3	1/2	1/5	1/4	1/3	1/2
Narrow	x	x			x				x	
Medium	x	x	x	x	x	x	x	x	x	x
Wide	x	x			x				x	

⁵ For producing smooth graphs at the extremes, we generated additional data sets, such that each data point within the graphs was based on 1,000,000 observations.

4.2.5 Evaluation Criteria

4.2.5.1 Root mean square error

To determine the accuracy of ability estimation based on the different test designs, the root mean square error (RMSE) of the ability estimates was computed for each simulation run l as

$$\text{RMSE}(\hat{\theta})_l = \sqrt{\frac{\sum_{i_l=1}^{m_l} (\hat{\theta}_{i_l} - \theta_{i_l})^2}{m_l}}, \quad (4.4)$$

where i represents the simulees within one run l ; m represents the total number of aggregated simulees within this run; and $\hat{\theta}$ denotes the estimate of the student ability θ . For each simulation condition, $\text{RMSE}(\hat{\theta})$ was calculated on three different levels: (1) over the total mixture population for each simulation run ($m = 30,000$), (2) for each ability group within each simulation run ($m = 10,000$), and (3) for intervals of 0.1 on the theta scale to investigate measurement accuracy in relation to student ability. On this level, the number of aggregated simulees m depended on the frequency distribution of the simulees over the theta scale as defined for the three distribution conditions. In addition, the overall mean $\text{RMSE}(\hat{\theta})$ over the 1,000 simulation runs was calculated for each simulation condition for the mixture population as well as for each ability group as

$$M_{\text{RMSE}(\hat{\theta})} = \frac{\sum_{l=1}^{1000} \text{RMSE}(\hat{\theta})_l}{1000}. \quad (4.5)$$

4.2.5.2 Efficiency gain over the targeted and the MST designs

To facilitate comparing the efficiency of the different designs with the efficiency of the targeted and the MST designs, we translated the differences in mean $\text{RMSE}(\hat{\theta})$ between each design D and the related targeted or MST designs into numbers of items required to compensate for these differences. Under the constraint that the properties of the sample and of the item pool remain constant, we can conclude from equation 4.2 that the standard error $SE(\hat{\theta}_i)$ is inversely proportional to the square root of the number of items k :

$$SE(\hat{\theta}_i) \sim \frac{1}{\sqrt{k}}. \quad (4.6)$$

Based on this relationship and following the definition of relative efficiency by Lord (1980), we calculated the efficiency gain of design D over the targeted design, Gain_T , and over the MST design, Gain_{MST} , as

$$\text{Gain}_T = 30 - 30 \cdot \left(\frac{M_{\text{RMSE}(\hat{\theta})_T}}{M_{\text{RMSE}(\hat{\theta})_D}} \right)^2 \quad \text{and} \quad \text{Gain}_{\text{MST}} = 30 - 30 \cdot \left(\frac{M_{\text{RMSE}(\hat{\theta})_{\text{MST}}}}{M_{\text{RMSE}(\hat{\theta})_D}} \right)^2. \quad (4.7)$$

Gain_T and Gain_{MST} refer to the relative efficiency of the designs compared to the targeted and the MST designs expressed in numbers of items, and, therefore, serve as indicators of the practical meaning of the differences in mean RMSE($\hat{\theta}$) between the designs (i.e., “how many extra items do we need?”).

4.2.5.3 Analysis of variance and effect sizes

Variation of RMSE($\hat{\theta}$) over the simulation runs between the different simulation conditions was further analyzed by two-way analyses of variance (ANOVAs). In the first set of ANOVAs, we compared RMSE($\hat{\theta}$) over the four design conditions combined with the three ability distribution conditions within the mixture population as well as within the three ability groups. In the second set of ANOVAs, we compared RMSE($\hat{\theta}$) over the MST and the TMST design conditions combined with the four starting module length conditions within the mixture population as well as within each ability group. To facilitate comparing the effects and interactions among the manipulated factors (i.e., population distribution, test design, and starting module length), effect size η^2 was calculated as

$$\eta^2 = \frac{SS_{between}}{SS_{total}}, \quad (4.8)$$

where $SS_{between}$ is the sum of squares between effects, and SS_{total} is the total sum of squares of the model (Richardson, 2011).

4.2.5.4 Match between true ability and module difficulty

We investigated the match between students' true ability θ and module difficulty as a potential source for differences in the efficiency of the different designs, the different distribution conditions, and the different student abilities. To this end, we calculated the percentage of correctly allocated, slightly misallocated, and heavily misallocated simulees per ability group and stage under the three distribution conditions by following Pohl (2013). In stage 1, simulees were classified as correctly allocated to the easy, moderate, and difficult modules if their ability θ was below the mean of μ_1 and μ_2 , between the mean of μ_1 and μ_2 and the mean of μ_2 and μ_3 , and above the mean of μ_2 and μ_3 , respectively (see Table 4.1 for the mean ability of each group under the three distribution conditions). In stages 2 and 3, simulees were classified as correctly allocated to the easy, moderate, and difficult modules, if their ability θ was below the routing cutoff score c_1 , between the two cutoff scores c_1 and c_2 , and above the cutoff score c_2 , respectively (see Table 4.2 for the routing cutoff scores under each distribution condition). Independent of the stage, simulees were classified as slightly misallocated if they were assigned to a module that was either one level too high or one level too low; they were classified as heavily misallocated if they were assigned to a difficult instead of an easy module or vice versa.

4.3 Results

4.3.1 RMSE($\hat{\theta}$) of the Mixture Population per Design and Distribution Condition

4.3.1.1 Mean RMSE($\hat{\theta}$) and efficiency gain over the targeted and MST designs

As an indicator of overall efficiency, Table 4.4 provides an overview of the mean RMSE($\hat{\theta}$) over the 1,000 simulation runs under each of the different distribution-by-design combinations for the mixture population and the three ability groups. Results for the low-ability and high-ability groups are displayed together because they were identical given the symmetrical distributions of student ability and module difficulty. To indicate the relative overall efficiency of the different designs, Table 4.4 reports on the relative efficiency gain (or loss) compared to the targeted design and the MST design within the corresponding distribution condition. Furthermore, Table 4.5 provides information about the relative effects of the different factors of the simulations on RMSE($\hat{\theta}$).

With regard to the total (i.e., mixture) population, the TMST design reached—as hypothesized—the lowest mean RMSE($\hat{\theta}$) of all four design conditions within all three distribution conditions. Furthermore, the MST design outperformed the targeted design, which, in turn, outperformed the linear design. The overall efficiency gain of the TMST design over the targeted design was larger within the narrow design condition (i.e., when the correlation between the ability-related background variable and the ability was low) than within the medium and wide distribution conditions. While the mean RMSE($\hat{\theta}$) of the targeted design was equal for all distribution conditions ($M_{\text{RMSE}(\hat{\theta})} = 0.425$), the mean RMSE($\hat{\theta}$) of the TMST design was lowest in the narrow distribution condition and highest in the wide distribution condition ($M_{\text{RMSE}(\hat{\theta})} \text{ Narrow} = 0.394$; $M_{\text{RMSE}(\hat{\theta})} \text{ Medium} = 0.398$; $M_{\text{RMSE}(\hat{\theta})} \text{ Wide} = 0.403$). In terms of numbers of items, the efficiency gain within the narrow condition corresponded to $\text{Gain}_T = 4$ additional items, or an increase of the total test length by 13%, to achieve the same accuracy with the targeted design as with the TMST design. Within the medium and the wide conditions, the efficiency gain corresponded to $\text{Gain}_T = 3$ additional items (i.e., 10% of the total test length).

The results of the overall efficiency gain of the TMST design compared to the MST design were also in line with our hypothesis: the highest efficiency gain was achieved in the wide condition, whereas no observable efficiency gain occurred in the narrow condition. As for the TMST design, the mean RMSE($\hat{\theta}$) of the MST design was lowest in the narrow and highest in the wide distribution condition ($M_{\text{RMSE}(\hat{\theta})} \text{ Narrow} = 0.395$; $M_{\text{RMSE}(\hat{\theta})} \text{ Medium} = 0.406$; $M_{\text{RMSE}(\hat{\theta})} \text{ Wide} = 0.419$). However, differences between the distribution conditions were larger for the MST than for the TMST design. In the narrow condition, the MST and TMST designs resulted in similar mean RMSE($\hat{\theta}$) such that no efficiency gain in terms of number of items was found. In the medium condition, the difference in mean RMSE($\hat{\theta}$) between the MST and TMST designs corresponded to an efficiency gain of $\text{Gain}_{\text{MST}} = 1$ additional item (i.e., 3% of the total

test length), and the difference in the wide condition corresponded to an efficiency gain of $\text{Gain}_{\text{MST}} = 2$ additional items (i.e., 7% of the total test length).

Table 4.4. Mean $\text{RMSE}(\hat{\theta})$ and Relative Gain over Targeted and MST Designs per Distribution Condition and Ability Group

Condition	Mixture Population			Medium-Ability Group			Low-Ability/High-Ability Group		
	$M_{\text{RMSE}(\theta)}$	Gain_T (%)	Gain_{MST} (%)	$M_{\text{RMSE}(\theta)}$	Gain_T (%)	Gain_{MST} (%)	$M_{\text{RMSE}(\theta)}$	Gain_T (%)	Gain_{MST} (%)
Narrow									
L	0.436	-2(-7)	-7(-23)	0.427	0(0)	-6(-20)	0.441	-2(-7)	-8(-27)
T	0.425		-5(-17)	0.425	0(0)	-6(-20)	0.425		-5(-17)
MST	0.395	4(13)		0.391	4(13)		0.397	4(13)	
TMST	0.394	4(13)	0(0)	0.391	4(13)	0(0)	0.395	4(13)	0(0)
Medium									
L	0.465	-7(-23)	-11(-37)	0.435	-1(-3)	-8(-27)	0.479	-9(-30)	-13(-43)
T	0.425		-3(-10)	0.425		-6(-20)	0.425		-2(-7)
MST	0.406	3(10)		0.393	4(13)		0.412	2(7)	
TMST	0.398	3(10)	1(3)	0.393	4(13)	0(0)	0.401	3(10)	2(7)
Wide									
L	0.501	-14(-47)	-16(-53)	0.452	-4(-13)	-11(-37)	0.524	-21(-70)	-21(-70)
T	0.425		-1(-3)	0.425		-5(-17)	0.425		0(0)
MST	0.419	1(3)		0.396	4(13)		0.429	-1(-3)	
TMST	0.403	3(10)	2(7)	0.396	4(13)	0(0)	0.406	3(10)	3(10)

Note. $\text{SE}_{\text{RMSE}(\theta)} < 0.0002$ for all conditions and ability groups. $\text{Gain}_T(\%)$ = relative gain over targeted design in numbers of items and percent (100% = 30 items); $\text{Gain}_{\text{MST}}(\%)$ = relative gain over MST design in numbers of items and percent (100% = 30 items); L = linear design (i.e., baseline); T = targeted design.

These results were also reflected by effect sizes calculated based on the two-way ANOVA for the mixture population for the two factors distribution and design (see the second column of Table 4.5). The design showed by far the largest effect on $\text{RMSE}(\hat{\theta})$ with $\eta^2 = .776$, meaning that the design explained 77.6% of the variance in $\text{RMSE}(\hat{\theta})$ between the different simulation conditions. The main effect of the distribution was $\eta^2 = .108$, and the interaction of the distribution and the design was $\eta^2 = .111$. All effects were statistically significant.

Table 4.5. Effect Sizes (η^2) for Main Effects and Interactions of the Factors Distribution and Design on $RMSE(\hat{\theta})$

Source of Variation	Mixture Population	Medium-Ability Group	Low-Ability/High-Ability Group
Distribution	.108	.031	.128
Design	.776	.902	.732
Distribution $\times$ Design	.111	.038	.128
Residual	.005	.028	.011

4.3.1.2 $RMSE(\hat{\theta})$ in relation to ability

Figure 4.2 shows the distribution of $RMSE(\hat{\theta})$ in relation to the true ability θ over a range from -2.6 to 2.6 (i.e., $\pm$ two SDs of the mixture population under the medium condition) for the different designs and the different distribution conditions. Furthermore, Figure 4.2 illustrates the number of students in relation to θ for each ability group and for the mixture population for all three distribution conditions. In general, $RMSE(\hat{\theta})$ was similar for all four designs close to the mean of the mixture population (i.e., $\theta = 0$), and differences in $RMSE(\hat{\theta})$ increased as expected toward more extreme abilities. Independent of the design condition, $RMSE(\hat{\theta})$ was generally higher for more extreme abilities because all designs included items targeted to the mixture population mean, the ability groups, or the routing groups.

As hypothesized, we found the smallest differences in $RMSE(\hat{\theta})$ between the different designs under the narrow condition, where the correlation between the group categorization (i.e., the ability-related background variable) and the true ability was small (see the top-left graph in Figure 4.2). Close to $\theta = 0$, $RMSE(\hat{\theta})$ was slightly lower for the linear and targeted designs than for the MST and TMST designs. The minimum of the $RMSE(\hat{\theta})$ distribution was at $\theta = 0$ for the linear and targeted designs, whereas $RMSE(\hat{\theta})$ increased toward $\theta = 0$ for the MST and TMST designs. However, $RMSE(\hat{\theta})$ was clearly lower for the MST and TMST designs than for the other two designs for low and high abilities. In line with the mean $RMSE(\hat{\theta})$ reported in Table 4.4, differences in $RMSE(\hat{\theta})$ between the MST and TMST designs were very small, whereas the TMST design reached slightly lower $RMSE(\hat{\theta})$ for low and high abilities.

Under the medium distribution condition displayed in the center graph of Figure 4.2, differences in $RMSE(\hat{\theta})$ between the designs were again small for average abilities close to $\theta = 0$. $RMSE(\hat{\theta})$ decreased from $\theta = 0$ toward low and high abilities for the MST and TMST designs before it increased toward the extremes. For the linear and the targeted designs, $RMSE(\hat{\theta})$ directly increased from average abilities toward the extremes. The TMST design provided the most accurate ability estimates for low and high abilities. The MST design, in turn, provided

more accurate estimates than the targeted design, and the linear design provided the least accurate estimates for low and high abilities.

Figure 4.2. $RMSE(\hat{\theta})$ in relation to ability for the different designs under the three distribution conditions and related frequency distributions of the ability groups.

Under the wide distribution conditions, where the correlation between the group categorization and the true ability was large, the MST design achieved lower $RMSE(\hat{\theta})$ for abilities close to $\theta = 0$ than the other three designs (see the top-right graph in Figure 4.2). From $\theta = 0$ toward low and high abilities, $RMSE(\hat{\theta})$ decreased for the targeted, the MST, and TMST designs before it increased toward the extremes. Only the linear design reached its lowest $RMSE(\hat{\theta})$ at $\theta = 0$. The decrease in $RMSE(\hat{\theta})$ was more prominent for the TMST design than for the targeted and MST designs, and the MST design reached its lowest $RMSE(\hat{\theta})$ closer to $\theta = 0$ than the two targeted designs. Consequently, the TMST design achieved lower $RMSE(\hat{\theta})$ values for low and high abilities than the other three designs. Similarly, the targeted design resulted in low $RMSE(\hat{\theta})$ for low and high abilities, and it even slightly outperformed the MST design at the extremes.

In sum, the results were consistent with our hypotheses. The combination of targeted and performance-based module assignment in the TMST design resulted in low $RMSE(\hat{\theta})$

values, especially for low and high abilities. By far, the variation of the design showed a larger effect on $RMSE(\hat{\theta})$ than the variation of the population distribution. Nevertheless, the efficiency gain of the TMST design over the other designs was larger under the wide condition than under the narrow condition. Under the narrow condition, differences were small not only between the mean abilities of the groups but also between the mean difficulties of the starting modules targeted to the different group means. Thus, all starting modules provided a similar amount of information close to $\theta = 0$ under the narrow condition but provided limited information for more extreme abilities. In contrast, differences were large between the groups' mean abilities and the modules' mean difficulties under the wide condition. As a consequence, targeted module assignment in stage 1 was much more crucial and effective under the wide condition than under the narrow condition, especially for low and high abilities. For medium abilities, however, the MST design was the safest option because it prevented misallocation to a too-easy or too-difficult starting module. The increase of $RMSE(\hat{\theta})$ of the MST and TMST designs close to $\theta = 0$, which was also observed for the targeted design under the wide condition, indicates that some simulees were either assigned or routed to suboptimal modules, which resulted in less efficient ability estimates. As the distributions widen, the modules differ more, and the consequences of possible misallocations increase.

4.3.2 $RMSE(\hat{\theta})$ of the Ability Groups per Design and Distribution Condition

4.3.2.1 *Mean $RMSE(\hat{\theta})$ and efficiency gain over the targeted and MST designs*

Results for the overall efficiency of the different designs per ability group are displayed in Table 4.4. In addition, Table 4.5 includes the η^2 -values for the main effects and interactions of the factors population distribution and design on $RMSE(\hat{\theta})$ for the different ability groups. For the medium-ability group, the MST and TMST designs resulted in identical mean $RMSE(\hat{\theta})$ values independent of the distribution condition. This result is given by the fact that the medium starting module of the TMST design corresponded to the general starting module of the MST design. The MST and TMST designs (i.e., the two designs with performance-based module assignments) clearly outperformed the targeted and the linear design under all three distribution conditions. Four additional items, or an increase of the total test length by 13%, would be required to achieve the same accuracy with the targeted design as with the MST and TMST designs. According to the ANOVA for the medium-ability group, the factor design showed the largest effect on $RMSE(\hat{\theta})$ with $\eta^2 = .902$. Thus, even though the MST and the TMST designs did not differ, the design explained 90.2% of the variance in $RMSE(\hat{\theta})$ between the different simulation conditions in the medium-ability group (distribution: $\eta^2 = .031$; interaction: $\eta^2 = .038$), which underlines the advantage of performance-based routing for the medium-ability group.

For the low-ability and high-ability groups, the TMST design provided—as expected—the lowest mean $RMSE(\hat{\theta})$ values. As reported for the mixture population, differences in mean

$RMSE(\hat{\theta})$ between the different designs increased from the narrow to the wide distribution condition. Under the narrow condition, the MST and TMST designs achieved comparable mean $RMSE(\hat{\theta})$ values for the low-ability and high-ability groups ($M_{RMSE(\theta)}^{MST} = 0.397$; $M_{RMSE(\theta)}^{TMST} = 0.395$). These values were clearly lower than the mean $RMSE(\hat{\theta})$ of the targeted ($M_{RMSE(\theta)} = 0.425$) and the linear designs ($M_{RMSE(\theta)} = 0.441$). In contrast, the targeted design slightly outperformed the MST design for the low-ability and high-ability groups under the wide condition ($M_{RMSE(\theta)}^{Targeted} = 0.425$; $M_{RMSE(\theta)}^{MST} = 0.429$), whereas the TMST design reached the lowest mean ($M_{RMSE(\theta)} = 0.406$) again. The efficiency gain of the TMST design over the MST and targeted designs corresponded to $Gain_T = Gain_{MST} = 3$ additional items or an increase of the total test length by 10%. In line with these results, the main effect of the population distribution on $RMSE(\hat{\theta})$ was $\eta^2 = .128$ and, therefore, was higher for the low-ability and high-ability groups than the effect reported above for the medium-ability group. The main effect of the design and the effect of the interaction were $\eta^2 = .732$ and $\eta^2 = .128$, respectively. Thus, the design also explained the largest percentage of the variance in $RMSE(\hat{\theta})$ between the different simulation conditions for the low-ability and high-ability groups.

In sum, the overall results for the different ability groups demonstrate that the reported difference in mean $RMSE(\hat{\theta})$ between the MST and TMST designs for the mixture population originates from the differing efficiency of these two designs with regard to the low-ability and high-ability groups. In line with our hypothesis, simulees of the low-ability and high-ability groups profited from the targeted module assignment in the first stage of the TMST, especially under the wide distribution condition. For all simulees of the medium-ability group, it made no difference whether their ability was estimated based on an MST design or a TMST design. In both designs, the medium-ability group clearly profited from the performance-based routing.

4.3.2.2 $RMSE(\hat{\theta})$ in relation to ability

Figure 4.3 shows the distribution of $RMSE(\hat{\theta})$ in relation to the true ability θ per ability group for the targeted and TMST designs, as well as for the different distribution conditions. In addition, it includes the number of students in relation to θ for each ability group and for the mixture population for each distribution condition. For the MST design, $RMSE(\hat{\theta})$ did not differ for the three ability groups within a distribution condition because the ability-related background variable had no influence on the module selection in this design condition. The results for the MST design corresponded again to the results for the medium-ability group of the TMST design because both conditions included identical starting modules.

In general, Figure 4.3 displays similar results patterns for all three distribution conditions. Differences between the designs and the ability groups were again more prominent for the wide than for the narrow distribution condition. As hypothesized, the targeted and TMST designs resulted in the lowest $RMSE(\hat{\theta})$ values for abilities close to the mean ability of the low-ability and high-ability groups. Both targeted designs outperformed the MST design for these ability ranges. The TMST design outperformed the targeted design for abilities that differed

from the mean ability of their group and thus were poorly represented by the ability-related background variable (i.e., high-ability simulees in the low-ability group and vice versa). Hence, the TMST design allowed for compensating for possible suboptimal module assignment in the first stage through performance-based routing to the modules of the second and third stages.

Figure 4.3. $\text{RMSE}(\hat{\theta})$ in relation to ability per ability group for the different designs under the three distribution conditions, and related frequency distributions of the ability groups. Targ. = targeted test design; LA = low-ability group; MA = medium-ability group; HA = high-ability group.

However, the MST design again provided lower $\text{RMSE}(\hat{\theta})$ values than the TMST design for abilities that clearly differed from the group mean, such as high-ability simulees in the low-ability group and low-ability simulees in the high-ability group. Nevertheless, it is important to contrast these results with the frequency distributions of the different ability groups and the mixture population displayed at the bottom of Figure 4.3. The number of simulees with abilities close to the group mean was high by design, meaning that, overall, the number of simulees who benefited from the targeted assignment in the first stage of the TMST was larger than the number of simulees who were disadvantaged because the ability-related background variable did not correspond to their abilities. As a consequence, mean $\text{RMSE}(\hat{\theta})$ was generally lower for the TMST design than for the MST design (see Table 4.4).

For the medium-ability group, results of the TMST design differed from our expectations. As shown by the central row of graphs in Figure 4.3, the minimum $\text{RMSE}(\hat{\theta})$ distribution did not coincide with the mean ability of the medium-ability group. Instead, $\text{RMSE}(\hat{\theta})$ was lowest between the mean of the low-ability group and the mean of the medium-ability group or between the mean of the medium-ability group and the mean of the high-ability group, and it slightly increased toward $\theta = 0$. This result corresponds to the $\text{RMSE}(\hat{\theta})$ distributions of the MST and TMST designs for the mixture population as displayed in Figure 4.2. Given that the starting module was targeted to the mean ability of the medium-ability group, the increase of $\text{RMSE}(\hat{\theta})$ within the group cannot be explained by assignment errors but must be a result of routing errors based on inaccurate preliminary ability estimates after stage 1. The risk of routing errors increases the closer the abilities are to the two routing cutoff scores (Weissman, 2014). The medium-ability group was located in between the two routing cutoff scores, such that the abilities of a vast majority of simulees within this group were relatively close to one of the two cutoff scores. Hence, they were more likely to be misallocated to a too-easy or too-difficult module than simulees from the low-ability or high-ability groups. Furthermore, routing errors had a larger impact on intermediate abilities, which clearly differed from the target difficulty of the easy and difficult modules. As a consequence, the abilities of average simulees within the medium group were estimated most efficiently with the traditional targeted design, which does not include any performance-based routing, making it resistant to routing errors.

4.3.2.3 Match between true ability and module difficulty per ability group

Figure 4.4 displays the percentage of correctly allocated, slightly misallocated, and heavily misallocated simulees for the MST and TMST designs⁶ for the medium-ability and low-ability or high-ability groups and for each stage under the different distribution conditions. In general, the percentage of correctly allocated simulees was higher under the wide distribution condition than under the narrow condition. This result confirmed our hypothesis that TMST designs are

⁶ Results for the targeted design correspond to results of the TMST design of stage 1.

more efficient if the correlation between the ability-related background variable and the true ability is high. A clearer distinction between the three ability groups as under the wide condition not only resulted in better targeting in stage 1 of the TMST design but also seemed to result in fewer routing errors. Furthermore, the percentage of correctly allocated simulees strongly increased from stage 1 to stage 2 and slightly increased from stage 2 to stage 3. Thus, the performance-based routing after stage 1 considerably increased the match between ability and module difficulty independent of the design and the ability group. In stages 2 and 3, the percentage of correctly allocated simulees was slightly higher in the low-ability and high-ability groups than in the medium-ability group. This finding strengthens our argument that simulees from the medium-ability group are more likely to be misrouted to a too-easy or too-difficult module than simulees from the low-ability or high-ability groups because a large proportion of simulees in the medium-ability group have abilities close to one of the two routing cutoff scores.

Figure 4.4. Match between θ and module assignment per ability group and stage for the MST and TMST designs under the three distribution conditions.

Results for the MST and TMST designs were comparable for the medium-ability group in stage 1 as well as for all comparisons within stages 2 and 3. However, the targeted module assignment in stage 1 of the TMST design allowed for increasing the percentage of correctly allocated simulees within the low-ability and high-ability groups from approximately 20% to over 60% compared to the MST design. This finding is consistent with our hypothesis that the TMST design is especially efficient for students whose abilities differ from the mean ability of

the total population. On the downside, the TMST design resulted in 22% of heavily misallocated simulees under the narrow distribution condition, whereas heavy misallocation was not possible in the MST design given one single starting module of moderate difficulty. Nevertheless, the percentage of heavily misallocated simulees in stage 1 of the TMST design considerably decreased as the distance between the ability groups increased, such that only 1% of the simulees of the low-ability or high-ability groups were heavily misallocated under the wide condition. Because the MST and TMST designs showed very similar patterns for the match of ability and module difficulty for stages 2 and 3, we combined the results of both designs for these stages in Figure 4.4. The corresponding percentages of correctly and incorrectly allocated simulees suggest that the accuracy of the routing to stages 2 and 3 was not impaired by either differences in the number and difficulty of starting modules or by differences in the match between ability and module difficulty between the MST design and the TMST design.

4.3.3 RMSE($\hat{\theta}$) in Relation to the Starting Module Length

4.3.3.1 Mean RMSE($\hat{\theta}$) and efficiency gain over the MST design

The length of the starting modules of the MST and TMST designs (i.e., the number of items in stage 1 compared to stages 2 and 3) was varied under the medium distribution condition. Table 4.6 displays the mean RMSE($\hat{\theta}$) of the MST and TMST designs as well as the efficiency gain of the TMST design over the MST design (Gain_{MST}) for starting modules with a length that corresponded to $\frac{1}{5}$, $\frac{1}{4}$, $\frac{1}{3}$, and $\frac{1}{2}$ of the total test length. In addition, Table 4.7 displays the effect sizes for the main effects and the interaction of the factors design and starting module length on RMSE($\hat{\theta}$) for the mixture population as well as for the ability groups.

Table 4.6. Mean RMSE($\hat{\theta}$) and Gain of TMST over MST Design for MST and TMST Designs with Different Starting Module Lengths per Ability Group

Starting Module Length	Mixture Population			Medium-Ability Group			Low-Ability/High- Ability Group		
	MST	TMST	Gain_{MST}	MST	TMST	Gain_{MST}	MST	TMST	Gain_{MST}
	$M_{\text{RMSE}(\hat{\theta})}$	$M_{\text{RMSE}(\hat{\theta})}$	(%)	$M_{\text{RMSE}(\hat{\theta})}$	$M_{\text{RMSE}(\hat{\theta})}$	(%)	$M_{\text{RMSE}(\hat{\theta})}$	$M_{\text{RMSE}(\hat{\theta})}$	(%)
1/5 TTL	0.403	0.398	1 (3)	0.394	0.394	0 (0)	0.407	0.399	1 (3)
1/4 TTL	0.402	0.398	1 (3)	0.390	0.390	0 (0)	0.408	0.402	1 (3)
1/3 TTL ^a	0.406	0.398	1 (3)	0.393	0.393	0 (0)	0.412	0.401	2 (7)
1/2 TTL	0.417	0.402	2 (7)	0.397	0.398	0 (0)	0.426	0.405	3 (10)

Note. $\text{SE}_{\text{RMSE}(\hat{\theta})} < 0.0002$ for all conditions and ability groups. $\text{Gain}_{\text{MST}}(\%)$ = relative gain over MST design condition in numbers of items and percent (100% = 30 items); TTL = total test length; 1/5, 1/4, 1/3, and 1/2 TTL corresponds to 6, 8, 10, and 16 items, respectively.

^aMain condition.

For the mixture population, we found that longer starting modules resulted in higher mean $\text{RMSE}(\hat{\theta})$ values for both designs. If the starting module length corresponded to $\frac{1}{5}$, $\frac{1}{4}$, or $\frac{1}{3}$ of the total test length, the TMST design was slightly more efficient than the MST design. In all three conditions, we would need to extend the MST design by one item, or 3% of the total test length, to achieve the same accuracy as with the TMST design. When starting modules of $\frac{1}{2}$ of the total test length were used, differences in mean $\text{RMSE}(\hat{\theta})$ between the two designs increased. For this condition, the mean $\text{RMSE}(\hat{\theta})$ of the MST design was $M_{\text{RMSE}(\theta)} = 0.417$, and the $\text{RMSE}(\hat{\theta})$ of the TMST design was $M_{\text{RMSE}(\theta)} = 0.402$. The difference in $\text{RMSE}(\hat{\theta})$ between the two designs, or, rather, the efficiency gain of the TMST design over the MST design, corresponded to $\text{Gain}_{\text{MST}} = 2$ items or an increase of the total test length by 7%. The effect of the starting module length on $\text{RMSE}(\hat{\theta})$ was $\eta^2 = .389$; thus, it was almost as large as the effect of the design on $\text{RMSE}(\hat{\theta})$ with $\eta^2 = .409$ (see Table 4.7). The interaction of the design and the starting module length explained the additional 10.1% of the total variance of $\text{RMSE}(\hat{\theta})$ between the different simulation conditions.

Table 4.7. Effect Sizes (η^2) for Main Effects and Interactions of the Factors Design (MST vs. TMST) and Starting Module Length on $\text{RMSE}(\hat{\theta})$

Source of Variation	Mixture Population	Medium-Ability Group	Low-Ability/High-Ability Group
Design	.409	.000	.445
Starting Module Length	.389	.424	.279
Design $\times$ Starting Module Length	.101	.000	.112
Residual	.101	.576	.164

For the medium-ability group, the MST and TMST designs resulted again in identical mean $\text{RMSE}(\hat{\theta})$ values. Accordingly, we found no main effect of the design and no interaction effect for the medium-ability group (i.e., $\eta^2 = .000$). Mean $\text{RMSE}(\hat{\theta})$ was lowest for this group if the starting module length corresponded to $\frac{1}{4}$ of the total test length ($M_{\text{RMSE}(\theta)} = 0.390$), whereas it increased for shorter as well as for longer starting modules. The starting module length accounted for 40.4% of the total variance in $\text{RMSE}(\hat{\theta})$ between the different simulation conditions for the medium-ability group.

For the low-ability and high-ability groups, mean $\text{RMSE}(\hat{\theta})$ was lowest for the TMST design with starting modules of $\frac{1}{5}$ of the total test length ($M_{\text{RMSE}(\theta)} = 0.399$) and highest for the MST design with a starting module of $\frac{1}{2}$ of the total test length ($M_{\text{RMSE}(\theta)} = 0.426$). The main effect of the design ($\eta^2 = .445$) was larger than that of the starting module length ($\eta^2 = .279$).

Differences in mean $RMSE(\hat{\theta})$ between the two designs were larger for longer starting modules than for shorter ones. The related interaction effect was $\eta^2 = .112$. If the starting module length corresponded to $\frac{1}{3}$ of the total test length, we would need to extend the MST design by two items, or 7%, to achieve the same accuracy as with the TMST design. If the starting module length corresponded to $\frac{1}{2}$ of the total test length, we would even need to extend the MST design by three items, or 10%, to achieve the same accuracy as with the TMST design.

In summary, the starting module length was much more relevant for the MST design than for the TMST design, and it was especially relevant for the medium-ability group. The longer the starting module of the MST design, the more disadvantaged were simulees of the high-ability and low-ability groups by the general starting module of moderate difficulty. In contrast, the starting modules of the TMST design were generally more efficient thanks to their targeting to each ability group, causing the length of the starting module to be less relevant under this design condition.

4.3.3.2 $RMSE(\hat{\theta})$ in relation to ability

Figure 4.5 shows the distribution of $RMSE(\hat{\theta})$ in relation to the true ability θ for the MST and TMST designs and the different starting module length conditions. As a reference, the two graphs include the distribution of $RMSE(\hat{\theta})$ for the linear and traditional targeted designs. In line with the reported differences in mean $RMSE(\hat{\theta})$, we found larger differences between the length conditions for the MST design than for the TMST design. Especially noteworthy were the results related to the longest starting modules (i.e., $\frac{1}{2}$ of the total test length). The graph on the right in Figure 4.5 indicates that the distribution of $RMSE(\hat{\theta})$ of the TMST design with such long starting modules was similar to those of the shorter conditions, and it achieved clearly lower $RMSE(\hat{\theta})$ values for high and low abilities than the traditional targeted design. However, as displayed on the left in Figure 4.5, the distribution of $RMSE(\hat{\theta})$ of the MST design with the longest starting module showed a stronger increase of $RMSE(\hat{\theta})$ toward low and high abilities than the other three length conditions and, thereby, reached $RMSE(\hat{\theta})$ values comparable to that of the targeted design. Thus, the performance-based routing of the MST design was comparably efficient as the sole a priori assignment of modules based on an ability-related background variable in the targeted design if the first stage included half of the total test.

Finally, Figure 4.6 shows $RMSE(\hat{\theta})$ in relation to ability for the different designs by ability group for each length condition. The general curve for the MST design was again overlapping with the curve of the TMST design for the medium-ability group. For the TMST design, Figure 4.6 indicates that as the starting module length increased, so did the differences between the ability groups at a given ability. Nevertheless, $RMSE(\hat{\theta})$ was low for simulees close to the mean of the target ability group and, thus, for the broad majority of simulees under all four length conditions (cf. the distributions displayed in Figure 4.2 and Figure 4.3). Only a

limited number of simulees with abilities deviant from the mean ability of their ability group were disadvantaged by long, misallocated starting modules.

Figure 4.5. $\text{RMSE}(\hat{\theta})$ in relation to ability for the MST and TMST designs with different lengths of starting modules.

Figure 4.6. $\text{RMSE}(\hat{\theta})$ in relation to ability for the MST and TMST designs with different lengths of starting modules per ability group. LA = low-ability group; MA = medium-ability group; HA = high-ability group.

4.4 Discussion

A good match between item difficulty and student ability is crucial from both measurement and motivational perspectives (Lord, 1980). In this paper, we investigated whether a combination of targeted and performance-based module assignments could increase the efficiency for estimating student ability under the Rasch model over a wide ability range. By means of simulations, we compared the efficiency of TMST designs—an extension of MST designs by targeted starting modules—with that of linear, traditional targeted, and MST designs. As hypothesized, the TMST design achieved the highest overall efficiency of all four designs, independent of the strength of the correlation between ability and the ability-related background variable. The TMST design also achieved higher overall efficiency than the MST design under all four starting module length conditions. The efficiency gain of the TMST design over the targeted design corresponded to up to 13% of the total test length, and the efficiency gain of the TMST design over the MST design corresponded to up to 7% of the total test length. Furthermore, our study allowed for comparing the efficiency of the MST and traditional targeted designs. Results showed that the MST design outperformed the targeted design in overall efficiency. These findings indicate that step-by-step module assignment based on performance is generally more efficient than one-time module assignment based on ability-related background variables. Finally, our results were in line with previous research by showing that the MST design and, therefore, performance-based module assignment, considerably increased measurement efficiency compared to a simple linear test (e.g., Lord, 1971b; Patsula, 1999; Pohl, 2013; Reese, Schnipke, & Luebke, 1999).

To get further insights into the efficiency of TMST designs, we varied the correlation between ability and background variable by manipulating the distance in mean ability between three ability groups. In line with our expectations and with previous results from Pohl (2013), the TMST design achieved the highest efficiency gain over the MST design when the background variable was a strong indicator of students' true ability. Under this condition, the TMST design achieved a considerably better match between ability and module difficulty within the starting modules than the MST design with its general, untargeted starting module. At the same time, the MST and TMST designs were comparably efficient under the condition of strongly overlapping ability distributions of the groups.

The overall efficiency gain of the TMST design over the targeted design was smaller under the wide distribution than under the medium and narrow distribution conditions. Furthermore, the efficiency gain of the MST design over the targeted design increased with decreasing distance between the ability groups. Analyses of the match between ability and module difficulty illustrated that as the strength of the overlap between the ability groups increased, so did the percentage of misallocated simulees in stage 1 of the TMST and, thus, also in the targeted design as a whole. This finding was in line with results reported by Pohl (2013) for the second wave of a longitudinal MST. However, the TMST design allowed for considerably decreasing the percentage of misallocations through performance-based routing

to stages 2 and 3 under all three distribution conditions. Such a compensation was not available in the targeted design. In contrast, a good indicator of ability could partly compensate for the lack of performance-based routing in the targeted design compared to the MST design.

The difference in efficiency between the designs varied considerably along the ability scale. As hypothesized and in line with previous research (Pohl, 2013), the TMST design achieved a higher efficiency than the other designs for low-ability and high-ability students, particularly if the ability-related variable was a strong indicator of ability. For medium abilities, the MST design was the most efficient because the single starting module was targeted to this ability range and prevented misallocations to easy or difficult starting modules. The analysis of measurement efficiency by ability group showed comparable results. The TMST design was generally more efficient than the other designs for the low-ability and high-ability groups. For the medium-ability group, the TMST design corresponded to the MST design, and they both outperformed the targeted design, which does not compensate for possible misallocations, in overall efficiency. Furthermore, results confirmed our hypothesis that the TMST design outperformed the targeted design for students whose abilities are poorly described by the background variable (i.e., high-ability students in the low-ability group and vice versa). In turn, the MST design outperformed the TMST design for these students because the module assignment in the MST design solely depended on performance and, therefore, was independent of the background variable. Nonetheless, the TMST design achieved the highest overall efficiency for the mixture population as well as within each ability group because the number of students who profited from the targeted assignment in the first stage was larger than the number of disadvantaged students. Interestingly, the targeted design achieved the highest efficiency of all designs for medium abilities within the medium group, which were partly affected by routing errors in the MST and TMST designs. However, this strength of the targeted design was clearly overshadowed by its low efficiency for students with abilities deviant from the group mean.

Finally, we explored the extent to which the efficiency of the MST and TMST designs depended on the length of the starting module. Based on their simulations, Kim and Plake (1993) suggested that longer starting modules should be preferred to shorter ones. However, the length of the starting module was confounded with the total test length in their study. We found that the MST design was most efficient if the starting module length corresponded from $\frac{1}{5}$ up to $\frac{1}{4}$ of the total test length. These results corresponded to those of previous simulations by Verschoor and Eggen (2014). Longer starting modules resulted in an efficiency loss (see also Lord, 1971b). A decrease in efficiency for longer starting modules was also found for the TMST design. However, the efficiency loss was much stronger for the MST design, which included only one general starting module, than for the TMST design, which included three targeted starting modules. The two TMST designs with long starting modules were disadvantageous particularly for students with abilities that strongly deviated from the mean of their ability group. Nevertheless, the vast majority of students, especially those in the low-ability and high-

ability groups, profited from the targeted modules in stage 1, independent of the length of the starting module. Hence, the length of the starting modules seems to be a negligible factor when developing TMST designs.

4.4.1 Limitations and Future Research

As in any simulation study, our study included a restricted set of conditions that constrains the generalizability of the results to some extent (Davey, Nering, & Thompson, 1997; Feinberg & Rubright, 2016). First, we included only a limited item pool with Rasch items targeted to the mean abilities of the ability groups or the subpopulations resulting from the routing. A strong variation of item difficulty within a module as well as the use of more complex IRT models could result in different module or test information. This in turn would affect both the accuracy and efficiency of ability estimation (e.g., Lord, 1980; Luecht, 2014). For example, we expect that more peaked module information would further increase the efficiency for abilities close to the target ability, whereas it would decrease the efficiency for abilities deviant from the target ability. At the same time, peaked module information could increase the number of routing errors due to lower information close to the routing cutoff scores (Weissman, 2014). Hence, differences in module information could enlarge or reduce the differences in efficiency between the different designs. It would be interesting to investigate the relationship between module information and efficiency of TMST designs compared to MST designs in more detail to identify the optimal TMST design for a given target population (cf. Verschoor & Eggen, 2014).

Second, we used a fixed test length as well as a fixed number of stages and modules in our simulation study. For longer tests, we hypothesize that the efficiency gain of the TMST design compared to the MST design would be reduced due to the increase in overall measurement accuracy given by the higher number of items, as stated in equation 4.2 (Rost, 2004; see also Stark & Chernyshenko, 2006). In addition, we expect that the efficiency gain of the TMST design compared to the MST design would decrease with increasing adaptivity within the two designs (i.e., more stages or more modules per stage for a given test length). However, increasing the adaptivity of the designs would also increase the test complexity and the effort required to assemble the modules. Previous studies indicated that the limited increase of measurement accuracy does not necessarily justify the increase in complexity for assembling more complex MST designs (e.g., Jodoin, Zenisky, & Hambleton, 2006; for an overview, see Luecht, 2014; Yan, von Davier et al., 2014). Based on these findings and our simulations, we expect a similar relationship for TMST designs. However, the hypothesized relationship between the efficiency of the TMST design and test length, as well as between efficiency and design complexity, should be verified in further studies (cf. Dallas, 2014).

4.4.2 Conclusion and Practical Implications

In conclusion, TMST designs refer to an innovative and efficient design type that combines traditional targeted testing with modern computer-based adaptive testing in the form of an MST.

With our simulation study, we extended previous research on the efficiency of different test designs in various ways. We not only introduced the TMST design as a new design type and analyzed its efficiency, but we also provided insights on the relative efficiency of targeted and MST designs. In particular, the efficiency of targeted test designs was not systematically studied in the past. As a consequence, our study allows for comparing the efficiency of module assignment based on ability-related background variables to those of performance-based module assignment. Our results indicated that the performance-based module assignment in the MST and TMST designs could substantially increase measurement efficiency compared to pure targeted module assignment based on ability-related background variables. When the target population spanned a narrow ability range, and the ability-related background variable was a poor indicator of students' true ability, the MST and TMST designs achieved comparable measurement efficiency. Hence, MST designs might be the better choice under this condition because they require fewer items in the starting modules than TMST designs, making them easier and cheaper to implement. However, TMST designs are a good option if the target population spans a wide ability range and a reliable ability-related background variable is available. Thanks to the targeted starting modules, TMST designs allow for taking low and high abilities into account from the first stage onward. As a result, TMST designs not only ensured efficient measurement of high-ability and low-ability students but also prevented underload and overload due to too-easy or too-difficult items.

In practice, the development and application of a TMST design brings about similar challenges and requirements as the development of an MST design. Given the limited research on TMST, we suggest practitioners follow the literature on MST (Hendrickson, 2007; Yan, von Davier et al., 2014; Zenisky et al., 2010) to clarify questions on the structure of the design (e.g., number of stages, number of segments per stage, routing rules, etc.). A crucial additional complexity of TMST compared to MST is the selection of a suitable ability-related background variable for the targeted assignment in the first stage. This variable must both provide reliable information about the students' true ability and be perceived as a fair criterion by the test takers, especially in the context of high-stakes testing. Objective criteria, such as school grade, exam grade, or performance-related school type, might be more acceptable to determine the starting module than, for example, teacher ratings. School grade as an indicator of number of years of education might be an especially well-accepted criterion for low-stakes as well as for high-stakes TMSTs. Exam grades or performance-related school types might be more difficult to justify in a high-stakes context and, thus, are rather recommended for low-stakes formative assessments. Consequently, TMST designs are a valuable extension of traditional MST designs to increase measurement efficiency in assessments for populations with a wide ability range, and they are particularly suitable for formative assessments.

4.5 References

- Angelone, D., Keller, F., & Moser, U. (2013). *Entwicklung schulischer Leistungen während der obligatorischen Schulzeit: Bericht zur vierten Zürcher Lernstandserhebung zuhanden der Bildungsdirektion des Kantons Zürich* [Development of school performance during compulsory school: Report on the fourth assessment for the attention of the Zurich department of education]. Zürich, CH: Institut für Bildungsevaluation (IBE). Retrieved from Institut für Bildungsevaluation (IBE) website: http://www.ibe.uzh.ch/projekte/lezh/Lernstandserhebung_9KlasseZH_Bericht.pdf
- Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, 55(1), 92–104.
- Barr, D. R., & Sherrill, E. T. (1999). Mean and variance of truncated normal distributions. *The American Statistician*, 53(4), 357–361. doi:10.1080/00031305.1999.10474490
- Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and development of differential learning and development environments]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (1st ed., pp. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bejar, I. I. (2014). Past and future of multistage testing in educational reform. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 423–438). Boca Raton, FL: CRC Press.
- College Board. (2017). *SAT Suite of Assessments Technical Manual: Characteristics of the SAT*. New York, NY: The College Board. Retrieved from <https://collegereadiness.collegeboard.org/pdf/sat-suite-assessments-technical-manual.pdf>
- Crotts, K. M., Zenisky, A. L., Sireci, S. G., & Li, X. (2013). Estimating measurement precision in reduced-length multistage-adaptive testing. *Journal of Computerized Adaptive Testing*, 1(4), 67–87.
- Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework* (Dissertation). University of North Carolina, Greensboro, NC. Retrieved from https://libres.uncg.edu/ir/uncg/f/Dallas_uncg_0154D_11394.pdf
- Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series No. 97-4). Iowa City, IA: American College Testing Program. Retrieved from American College Testing Program website: http://www.act.org/content/dam/act/unsecured/documents/ACT_RR97-04.pdf
- Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35(2), 36–49. doi:10.1111/emip.12111

- Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models. Springer series in statistics*. New York, NY: Springer.
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26(2), 44–52. doi:10.1111/j.1745-3992.2007.00093.x
- Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19(3), 203–220. doi:10.1207/s15324818ame1903_3
- Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.
- Lord, F. M. (1971a). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 31, 805–813. doi:10.1177/001316447103100401
- Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika*, 36(3), 227–242. doi:10.1007/BF02297844
- Lord, F. M. (1971c). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association*, 66(336), 707–711. doi:10.2307/2284216
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
- Luecht, R. M. (2014). Design and implementation of large-scale multistage testing systems. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 69–84). Boca Raton, FL: CRC Press.
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35(3), 229–249. doi:10.1111/j.1745-3984.1998.tb00537.x
- Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Reports Series No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
- Moser, U., Buff, A., Angelone, D., & Hollenweger, J. (2011). *Nach sechs Jahren Primarschule: Deutsch, Mathematik und motivational-emotionales Befinden am Ende der 6. Klasse* [After six years of primary school: German, Mathematics and motivational-emotional well-being at the end of grade 6]. Zürich, Switzerland: Bildungsdirektion Kanton Zürich. Retrieved from Bildungsdirektion Kanton Zürich website: <https://www.ibe.uzh.ch/de/projekte/lezh.html>

- Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multi-stage testing* (Doctoral Dissertation). University of Massachusetts Amherst, Amherst, MA. Retrieved from http://scholarworks.umass.edu/dissertations_1/3282
- Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement*, 50(4), 447–468. doi:10.1111/jedm.12028
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into multi-stage adaptive testing* (LSAC Research Report Series No. LSAC-R-97-02). Princeton, NJ: Law School Admission Council. Retrieved from Law School Admission Council website: <https://files.eric.ed.gov/fulltext/ED467816.pdf>
- Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2), 135–147. doi:10.1016/j.edurev.2010.12.001
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* [Textbook test theory, test construction] (2nd ed.). *Psychologie Lehrbuch*. Bern, Switzerland: Verlag Hans Huber.
- Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., . . . Hambleton, R. K. (2008). *Massachusetts Adult Proficiency Tests technical manual version 2* (Center for Educational Assessment Research Report No. 677). Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.
- Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable? *Applied Measurement in Education*, 19(3), 257–260. doi:10.1207/s15324818ame1903_6
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.
- Verschoor, A. J., & Eggen, T. J. H. M. (2014). Optimizing the test assembly and routing for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 135–150). Boca Raton, FL: CRC Press.
- Wainer, H. (Ed.). (2000a). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wainer, H. (2000b). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (2nd ed., pp. 1–21). Mahwah, NJ: Lawrence Erlbaum Associates.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. doi:10.1007/BF02294627
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4), 473–492. doi:10.1177/014662168200600408
- Weissman, A. (2014). IRT-based multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 153–168). Boca Raton, FL: CRC Press.

- Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, 2(1), 1–17. doi:10.7333/1401-0201001
- Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 3–20). Boca Raton, FL: CRC Press.
- Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.
- Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer.
- Zenisky, A. L., & Hambleton, R. K. (2014). Multistage test designs: Moving research results into practice. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 21–37). Boca Raton, FL: CRC Press.
- Zenisky, A. L., Sireci, S. G., Martone, A., Baldwin, P., & Lam, W. (2009). *Massachusetts Adult Proficiency Tests technical manual supplement: 2008-2009* (Center for Educational Assessment Research Report No. 715). Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Chapter 5. Development and Validation of a Vertical Scale for Formative Assessment in Mathematics⁷

Abstract

Regular formative assessment of students' abilities across multiple school grades requires a reliable and valid vertical scale. A vertical scale is a precondition not only for comparing assessment results over time and for measuring progress, but also for identifying the most informative items for each individual student within a large item bank independent of the student's grade to increase measurement efficiency. In this study, we described the development of a vertical scale for formative assessment of third- through ninth-grade students' mathematics abilities based on item response theory methods. We investigated the psychometric properties of the vertical unidimensional Rasch scale by means of item analysis, as well as by comparing two different calibration procedures, namely concurrent and grade-by-grade calibration. Furthermore, we evaluated the content-related validity of the scale by contrasting the calibration procedure's empirical outcomes—i.e., the item-difficulty estimates—with the theoretical, content-related item difficulties reflected by the underlying competence levels of the curriculum, which served as a content framework for developing the scale. Besides analyzing the general match between empirical and content-related item difficulty, we also explored, by means of correlation and multiple regression analyses, whether their match differed for items related to different curriculum cycles (i.e., primary vs. secondary school), domains, or competencies within mathematics. The results indicated satisfying item fit and a close match between the outcomes of the concurrent and grade-by-grade calibration procedures, underpinning the unidimensionality and stability of the scale from a psychometric perspective. In addition, we found strong correlations between the empirical and content-related item difficulties, which emphasized the scale's content-related validity. Further analysis showed a higher correlation between empirical and content-related item difficulty at the primary compared with the secondary school level. Across the different curriculum domains and most of the curriculum competencies, we found comparable correlations, implying that the scale is a good indicator of the math ability stated in the curriculum.

⁷ This chapter has been submitted for publication as Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (submitted). *Development and validation of a vertical scale for formative assessment in mathematics*. Manuscript submitted for publication.

5.1 Introduction

Several studies have pointed to the positive effect from formative feedback on learning and self-regulation (e.g., Campbell & Levin, 2009; Carlson, Borman, & Robinson, 2011; Cawelti & Protheroe, 2001; Lai, McNaughton, Amituanai Toloa, Turner, & Hsiao, 2009). Hattie (2009) even identified feedback as one of the most powerful influences on school achievement in his synthesis of over 800 meta-analyses. Modern computer technology can be used as a tool for providing formative feedback in classrooms on a regular basis (e.g., Brown, 2013; Hattie & Brown, 2007), and for implementing complex measurement models and item-selection algorithms that support teachers in providing objective, reliable, and valid feedback (e.g., Glas & Geerlings, 2009; Tomasik, Berger, & Moser, 2018; Wauters, Desmet, & Noortgate, 2010).

In Northwestern Switzerland, four cantons—Aargau, Basel-Landschaft, Basel-Stadt, and Solothurn—joined forces to develop a computer-based, formative feedback system for classrooms (Tomasik et al., 2018; see also <https://www.mindsteps.ch/>) to serve teachers and their nearly 100,000 third- through ninth-grade students as an instrument for data-based decision making (Schildkamp, Lai, & Earl, 2013; van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). The system's purpose is to support teachers and students in collecting objective information about students' current abilities, i.e., strengths and weaknesses, as well as learning progress within the domains and competencies stated in the curriculum for four school subjects: German, the schools' language; English and French, the two foreign languages taught; and mathematics. Based on this information, teachers and students can define appropriate learning goals, evaluate progress in realizing these goals over time, and adjust teaching, learning environments, or goals if necessary (Hattie & Timperley, 2007; van der Kleij et al., 2015).

To provide targeted feedback on students and teachers' specific needs, the system is conceptualized as an item bank with several-thousand assessment items that teachers and students can select based on curriculum-related, as well as empirical, criteria, such as curriculum-related competence levels or empirical item-difficulty estimates. As a basis for representing item difficulties on a common scale, the system uses item response theory (IRT, de Ayala, 2009; Lord, 1980), or rather the Rasch model (Rasch, 1960), as an underlying measurement model. One advantage of this measurement approach is that it allows for directly comparing assessment results related to different item sets. Thus, each student can work on a targeted selection of items within the item bank, but still compare results with those of other students, as well as with their own results from earlier assessments. In addition, the IRT approach supports targeted item selection through algorithms for computerized adaptive testing (CAT), which use preliminary ability estimates during assessment for selecting the most appropriate and informative items for each individual student (van der Linden & Glas, 2010; Wainer, 2000).

Within this context, two basic prerequisites for implementing such a competence-based feedback system are clear content specifications, which guide the assessment items'

development for the item bank (Webb, 2006), and a vertical measurement scale, which allows for monitoring and comparing students' abilities over several school grades (Young, 2006). For Northwestern Switzerland, the competence-based curriculum *Lehrplan 21*, which was made available to German-speaking cantons in Switzerland in autumn 2014 (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014), is an obvious choice as a basis for the content specifications. However, it is a more challenging endeavor to develop a vertical measurement scale based on IRT methods to represent students' competence levels over seven school grades (i.e., from third to ninth grades) as stated in the curriculum. Results from previous studies on vertical scaling indicated that the interpretation of growth or progress on a vertical IRT scale might depend on various decisions made during the calibration process (see Briggs & Weeks, 2009; Harris, 2007; Kolen & Brennan, 2014, for a general overview). One of the principal factors investigated in previous research was the impact of different linking-calibration approaches—namely concurrent vs. grade-by-grade calibration—on the resulting growth patterns. However, results from these studies are mixed and do not provide clear guidance for the practical implementation of vertical scales based on IRT methods. Some studies identified concurrent calibration as superior to grade-by-grade calibration (e.g., Hanson & Béguin, 1999; Kim & Cohen, 1998), while others reported the opposite (e.g., Béguin, Hanson, & Glas, 2000; Ito, Sykes, & Yao, 2008). In addition, several studies point out diverse interactions between the calibration approach and various other decisions during the development of the vertical scale, such as choice of IRT model, data-collection design, number of linking items, or test specifications (Briggs & Weeks, 2009; Hanson & Béguin, 2002; Lei & Zhao, 2012; Pomplun, Omar, & Custer, 2004; Tong & Kolen, 2007; see also Harris, 2007). Simulation studies might help investigate selected combinations of scaling decisions systematically. However, in practice, the best combination of decisions might depend largely on the vertical scale's specific measurement objectives, changes in ability distribution across grades (e.g., Keller & Hambleton, 2013), or the extent to which the data meet the strict assumptions of unidimensionality and parameter invariance in the underlying IRT model (Béguin et al., 2000; Pohl, Haberkorn, & Carstensen, 2015; Tong & Kolen, 2007).

One important aspect missing from the aforementioned extant studies was an external criterion for validating the vertical scale, specifically a criterion that allows for justifying the decisions taken during test development and item calibration, while considering the concrete, latent construct or ability to be measured, and for verifying the resulting growth pattern as the true one (Briggs & Weeks, 2009; Dadey & Briggs, 2012; Harris, 2007; Ito et al., 2008; Tong & Kolen, 2007). To close this gap, in the present study, we used content-related item difficulties—i.e., content experts' difficulty ratings based on the items' localization within the curriculum—as a basis for validating a new vertical IRT scale. In particular, we described the development of a vertical scale for formative assessment for third through ninth grades for the example case of mathematics. Given the mixed results from studies comparing concurrent vs. grade-by-grade calibration procedures, we applied both procedures and compared the two procedures' outcomes to detect potential model misfit (Hanson & Béguin, 2002; Kolen

& Brennan, 2014). Our study's focus lied in the validation of the final scaling decisions and, thus, the vertical scale. To this end, we contrasted the item calibration's outcomes with the underlying content specifications derived from the curriculum for mathematics. More precisely, we investigated the extent to which the empirical item-difficulty parameters resulting from the calibration reflected the items' content-related difficulties based on their assignment to specific competence levels of the curriculum by means of cross-sectional data from a pretest calibration sample.

5.1.1 Curriculum Lehrplan 21 as Content Framework

The curriculum, *Lehrplan 21*, describes the competencies that students should acquire from kindergarten until the end of compulsory school, providing teachers and schools with a basis for planning their teaching and evaluating students' progress (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014, 2016b; see also www.lehrplan.ch). Within the subject of mathematics, the curriculum is structured hierarchically into three domains (i.e., “number and variable,” “form and space,” and “measures, functions, data, and probability”), 26 competencies, and various competence levels (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016a). Within each competency, the curriculum calls for a continuous development of the competency over the school years, whereas mastery of lower competence levels is a precondition for mastering more advanced competence levels (see also Bundesinstitut für Bildungsforschung, Innovation & Entwicklung, 2011; Reusser, 2014). Furthermore, the curriculum delineates three different cycles—kindergarten to second grade, third to sixth grades, and seventh to ninth grades; defines basic requirements for each cycle, i.e., the minimal competence levels that students need to master at the end of each cycle; and states two points of orientation at the end of fourth and eighth grades. The cycles, basic requirements, and points of orientation anchor the competence levels and, thus, the development of competencies, across the 11 compulsory school years. However, the curriculum focuses much more on the development of students' competence levels across grades than on the specific competencies within a particular school grade, thereby following a domain definition of growth as defined by Kolen and Brennan (2014, pp. 429–431).

Figure 5.1 provides an example of a mathematics competency, namely MA.1.A.2, which covers “counting, ordering, estimating” and is part of the domain “number and variable” (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016a). Within this competency, the mathematics curriculum distinguishes between 10 different competence levels of increasing difficulty (i.e., levels a to j). For each competence level, it provides detailed descriptions of what students should know to master the level. The first three levels, a to c, belong to the curriculum's first cycle, whereas the gray, highlighted level c refers to this cycle's basic requirements. Levels d to g refer to the curriculum's second cycle, with level g including this

MA.1.A.2 Die Schülerinnen und Schüler ...	
1	a » können bis zu 20 Elemente auszählen und Zahlpositionen vergleichen.
	b » können im Zahlenraum bis 20 von beliebigen Zahlen aus vorwärts und rückwärts zählen. » können in 2er-Schritten vorwärts zählen, von 2 bis 20. » können Fingerbilder von 1 bis 10 spontan zeigen sowie Anzahlen bis 5 ohne Zählen erfassen.
	c » können im Zahlenraum bis 100 in 1er-, 2er-, 5er- und 10er-Schritten vorwärts zählen. » können im 100er-Raum Zahlen ordnen (z.B. auf dem Zahlenstrahl und auf der 100er-Tafel).
	d » können im Zahlenraum bis 100 von beliebigen Zahlen aus vorwärts und rückwärts zählen. » können im Zahlenraum bis 100 von beliebigen 10er-Zahlen aus in 2er-, 5er- und 10er-Schritten vorwärts und rückwärts zählen.
2	e » können im Zahlenraum bis 1'000 von beliebigen Zahlen aus in 1er-, 2er-, 10er- und 100er-Schritten vorwärts und rückwärts zählen. » können Zahlen bis 1'000 ordnen.
	f » können im Zahlenraum bis 1 Million von beliebigen Zahlen aus in angemessenen Schritten vorwärts und rückwärts zählen (z.B. von 320'000 in 20'000er-Schritten). » können Zahlen bis 1 Million ordnen (z.B. die ungefähre Position von 72'000 auf einem Zahlenstrahl bestimmen).
	g » können von beliebigen Dezimalzahlen aus in angemessenen Schritten vorwärts und rückwärts zählen (z.B. von 0.725 in 0.005er-Schritten). » können Brüche mit den Nennern 2, 3, 4, 5, 6, 8, 10, 20, 50, 100 ordnen. » können Dezimalzahlen ordnen (z.B. 1.043; 1.43; 1.05; 1.5; 1.403). » können Grundoperationen mit natürlichen Zahlen überschlagen (z.B. $13'567 + 28'902 \approx 40'000$; $592'000 : 195 \approx 600'000 : 200$).
	h » können Summen und Differenzen mit Dezimalzahlen überschlagen (z.B. $0.723 - 0.04 \approx 0.7$; $23'268 + 4'785 \approx 28'000$). » können in Prozentrechnungen Ergebnisse überschlagen (z.B. 263 von 830 sind etwa 30%; 45% von 13'000 sind mehr als 5'000).
3	i » Erweiterung: können Produkte und Quotienten von Dezimalzahlen überschlagen. (z.B. $0.382 : 42.8 \rightarrow 0.4 : 40 = 0.4 : 4 : 10 = 0.01$; $32.7 : 0.085 \rightarrow 30 : 0.1 = 300 : 1 = 300$).
	j » können positive und negative rationale Zahlen auf dem Zahlenstrahl ordnen.

Figure 5.1. Extract of the curriculum, *Lehrplan 21*: Example of a mathematics competency and its competence levels. Orange, blue, and green frames indicate the curriculum cycles: cycle 1 = kindergarten through second grade; cycle 2 = third through sixth grades; cycle 3 = seventh through ninth grades. Gray levels refer to the basic requirements within a cycle, and red, dotted lines serve as orientations at the ends of fourth and eighth grades. From *Lehrplan 21: Mathematik* (p. 11), by Deutschschweizer Erziehungsdirektoren-Konferenz, 2016, Luzern, retrieved from https://v-fe.lehrplan.ch/container/V_FE_DE_Fachbereich_MA.pdf Copyright 2016 by Deutschschweizer Erziehungsdirektoren-Konferenz. Reprinted with permission.

cycle's basic requirements. Levels h to j belong to the third cycle, with level j providing basic requirements. In addition, two red, dotted lines serve as orientation points for students' competence levels at the end of grades 4 and 8. Thus, the curriculum contains detailed competence descriptions for distinct competencies and levels, as well as references to minimal standards for certain school grades. As such, it serves as an ideal basis—i.e., content framework—for developing targeted items to assess students' ability formatively at various points in time during compulsory school years.

5.1.2 Vertical Scaling Based on Item Response Theory Methods

5.1.2.1 Justifying a vertical scale

To monitor students' development along the competencies and competence levels stated in the curriculum, a vertical measurement scale is required, which refers to “an extended score scale that spans a series of grades and allows the estimation of student growth along a continuum” (Young, 2006, p. 469; see also Briggs, 2013; Harris, 2007; Kolen & Brennan, 2014). Such a scale is the basis for comparing the outcomes of various consecutive measurement occasions and, thus, for measuring progress, analyzing students' growth in relation to vertically moderated standards, and CAT across grades (e.g., Cizek, 2005; Dadey & Briggs, 2012; Ferrara, Johnson, & Chen, 2005). A precondition for justifying a vertical scale is the assumption that the measured abilities or competencies are continuously stimulated and increased over time (Young, 2006). In contrast to horizontal scales, which represent one specific age or grade group's abilities, vertical scales combine test forms or item sets that vary in their mean difficulty, reflecting the broad ability range that must be covered when assessing ability over several school grades. However, even though the different assessment forms refer to different difficulty levels, the underlying latent construct needs to remain constant from a content perspective. Otherwise, a unidimensional vertical scale cannot be justified.

5.1.2.2 Test designs for vertical scaling

Kolen and Brennan (2014, pp. 431–434) distinguish between three different test designs for collecting data (i.e., student responses) to establish vertical scales based on IRT methods: (1) equivalent group designs; (2) common item designs; and (3) scaling test designs. In equivalent group designs, groups with equivalent ability distribution within a school grade are randomly assigned to answer items related to their own, adjacent lower, or adjacent higher grade. The linking within each grade is based on the assumption that all groups have comparable ability distributions. The linking between grades is based on items administered to two adjacent grades. Administering identical items to students from adjacent grades is the basic idea of common item designs, the second design type. The advantage of this design type is that it doesn't require equivalent groups, only common items, i.e., linking items, to link several item blocks. Linking items serves not only to establish a link between two grades, but also to align overlapping item blocks within one grade. Scaling test designs, the third design type, is similar to common item designs to the extent that students from different school grades solve identical items. However,

when applying a scaling test design, common items are shared not only between adjacent grades. Instead, one block of items, namely the scaling test, is administered to all involved grades. Besides the scaling test, students from each grade answer items related to their specific grade.

From these three designs, scaling test designs are the most consistent with a domain definition of growth and are the first choice in such a context from a theoretical perspective (Kolen & Brennan, 2014). In contrast, common item designs are the easiest to implement in practice under the condition that it is reasonable to administer the same items to students from adjacent grades from a content perspective (Kolen & Brennan, 2014). Equivalent group designs require more complex administration procedures within one school grade to ensure samples with equivalent ability distributions. Scaling test design requires that identical items be administered to students from even more school grades, which is difficult to justify from a content perspective if the target population spans seven school years, as in our case.

5.1.2.3 Calibration and linking based on IRT methods

After administering the items to students by means of the data-collection designs described above, the items need to be calibrated to establish a vertical measurement scale. Within the Rasch model's context, *item calibration* refers to establishing model fit and estimating the difficulty parameter, β_i , of item i based on response data by means of maximum likelihood estimation procedures (Eggen & Verhelst, 2011; Vale & Gialluca, 1988). Generally, two different procedures are used to link IRT-based item-difficulty parameters to a common vertical scale across multiple grades: concurrent and grade-by-grade calibration (Briggs & Weeks, 2009; Kolen & Brennan, 2014). Under the concurrent procedure (Wingersky & Lord, 1983), all item parameters are estimated in one single calibration run, whereby different underlying population ability distributions need to be specified for each grade group (DeMars, 2002; Eggen & Verhelst, 2011). By factoring in the groups' differences in ability, this procedure directly maps all item parameters onto one common scale by means of linking items shared by two adjacent grades.

In contrast, item parameters are estimated separately for each grade under the grade-by-grade calibration procedure. These parameters then are transformed into one common scale by means of linear transformations via

$$\beta_{ix} = A \beta_{iy} + B, \quad (5.1)$$

in which x and y refer to the two scales, and A and B refer to the two linking constants. Different methods allow for determining the linking constants, one of the most accurate and popular methods of which is the Stocking and Lord method (Stocking & Lord, 1983; e.g., Briggs & Weeks, 2009; Kolen & Brennan, 2014), which determines linking constants by minimizing differences between linking items' item characteristic curves (ICCs) between two grades. To link parameters over more than two grades, several transformation steps are required for grades, which are placed further apart from the base grade level. For example, if third grade serves as

the base level, parameters for fifth grade require a transformation with A_{34} and B_{34} , as well as with A_{45} and B_{45} .

As reported earlier, previous research comparing concurrent and grade-by-grade calibration procedures yielded mixed results and didn't provide clear guidance for practical implementation of vertical scales based on IRT methods. From a theoretical perspective, concurrent calibration might be superior to grade-by-grade calibration. According to Kolen and Brennan (2014, p. 444), it "is expected to produce more stable results because it makes use of all of the available information for parameter estimation." Furthermore, the concurrent procedure is less prone to errors because it doesn't require the estimation of linking constants, which can elicit additional estimation errors, and it is more efficient because it requires only one calibration run (Briggs & Weeks, 2009). However, Kolen and Brennan (2014) also listed several arguments that support the use of grade-by-grade calibration in a practical context (e.g., Hanson & Béguin, 2002). First, grade-by-grade calibration has the advantage of allowing for direct comparison of item-parameter estimates between two adjacent grades and, thus, for investigating potential deficiencies in parameter invariance across school grades. Investigating parameter invariance also is possible under the concurrent-calibration approach, but it requires—depending on the calibration software—additional calibration runs for subsamples of data. Second, separate calibrations for each grade group are based on smaller and simpler data matrices than concurrent calibration, which includes all available data at once. As a result, the estimation procedure converges faster, so convergence problems are less likely. Last, but not least, grade-by-grade calibration might be more robust against the violation of unidimensionality assumption because it considers only data from two school grades (Béguin et al., 2000; Béguin & Hanson, 2001; Hanson & Béguin, 2002). Thus, separate calibration might be the first choice if the empirical data cannot perfectly be described through the IRT model. Against this backdrop, Hanson and Béguin (2002), as well as Kolen and Brennan (2014), recommended applying both procedures when developing a vertical scale. Differences between the two procedures' outcomes might help detect serious problems in the calibration process, such as multidimensionality and, thus, model misfit (Hanson & Béguin, 2002).

5.1.3 The Present Study

In this study, we described the development of a vertical Rasch scale for the formative assessment of students' ability in mathematics from third through ninth grades. Furthermore, we investigated this scale's validity for mapping competency development from third through ninth grades, as stated in the curriculum, *Lehrplan 21*. Specifically, we aimed to answer the following three research questions:

- I. Do the items developed on the basis of the curriculum, *Lehrplan 21*, and targeted to third through ninth grades fit a unidimensional vertical Rasch scale?

- II. Do the item calibration's empirical outcomes—i.e., item-difficulty estimates—match the theoretical, content-related item difficulties that reflect the curriculum's underlying competence levels?
- III. Does the match between the empirical item-difficulty estimates and the theoretical, content-related item difficulties differ for items related to different curriculum cycles, domains, or competencies?

The research questions were addressed through data from a cross-sectional calibration study with third- through ninth-grade students from Northwestern Switzerland. To tackle the first research question, we performed in-depth item analyses. Moreover, we contrasted the outcomes of concurrent and grade-by-grade calibration, and compared the estimated item-difficulty parameters' stability across school grades. We hypothesized that both calibration procedures would result in similar outcomes and that item-difficulty parameters are invariant across school grades. For the second and third research questions, we assumed a strong positive relationship between the empirical item-difficulty estimates and content-related item difficulties. Furthermore, we hypothesized that the relationship's strength is similar across different curriculum cycles, domains, and competencies.

5.2 Method

5.2.1 Content-Related Item Difficulty

Our content experts identified 18 out of the 26 different mathematics competencies from the curriculum, *Lehrplan 21* (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016a) as assessable through computer-based items with clear answer formats (e.g., multiple choice, short text, drag-and-drop items). Within these competencies, they identified competence levels ranging from the basic requirements of cycle 1 to the penultimate or ultimate competence level of cycle 3 as relevant for students in third through ninth grades. The number of competence levels within this target range varies between five and 10, resulting in some competence levels that span a broader part of primary and secondary school than others. To compare the levels' relative difficulty over the different competencies, we aligned the competence levels according to the basic requirements for cycles 1, 2, and 3 and according to the orientation points for the end of grades 4 and 8 (cf. Figure 5.1). This procedure's outcome is a matrix presented in Table 5.1. On one hand, this matrix served as a basis for the item-calibration test design, which we describe in more detail in the next section, and for item development. Item developers were instructed to construct items that can be assigned clearly to one competence level each (e.g., MA.1.A.2.f) and, therefore, are related to one single competency (MA.1.A.2), domain (MA.1.A), and curriculum cycle (C2). On the other hand, we used the matrix to translate each competence level into a score on a scale from 1 to 11, which served as a measure of content-related difficulty (CRD) for our analyses (see Table 5.1, last row). Competence levels spanning

Table 5.1. Overview of 18 Mathematics Competencies from Curriculum, *Lehrplan 21*, and the Alignment of Their Competence Levels with the Scale for CRD.

			Grades		3	4	5	6	7	8	9	
			Curriculum cycles		1	2			3			
Domains Competencies			BR	OP			BR	OP			BR	
MA.1	MA.1.A.2	Counting, ordering, estimating	c	d	e	f	g	h	i	j		
	MA.1.A.3	Addition, subtraction, multiplication, division, exponentiation	b	c	d	e	f	g	h	i	j	
	MA.1.A.4	Terms and equations, principles, and rules	c	d	e	f	g	h	i	j	k	
	MA.1.B.1	Numbers and operations	c	d	e	f	g	h	i	j	k	
	MA.1.B.2	Verification of results	c	d	e	f	g	h	i	j	k	
	MA.1.C.1	Calculation pathways	c	d	e	f	g	h	i	j	k	
	MA.1.C.2	Generalization of patterns	c	d	e	f	g	h	i	j	k	
MA.2	MA.2.A.2	Decomposition and composition of figures and objects	c	d	e	f	g	h	i	j		
	MA.2.A.3	Computation of lengths, surfaces, and volumes	b	c	d	e	f	g	h	i	j	
	MA.2.B.1	Exploration of lengths, surfaces, and volumes	c	d	e	f	g	h	i	j	k	
	MA.2.C.3	Geometric figures and objects in different positions	c	d	e	f	g	h	i	j	k	
	MA.2.C.4	Coordinate systems	b	c	d	e	f	g	h	i	j	
MA.3	MA.3.A.2	Variables	c	d	e	f	g	h	i	j	k	
	MA.3.A.3	Relationships	b	c	d	e	f	g	h	i	j	
	MA.3.B.2	Statistics, combinatorics, and probability	a	b	c	d	e	f	g	h	i	
	MA.3.C.1	Data collection, ordering, presentation, analysis, interpretation	b	c	d	e	f	g	h	i	j	
	MA.3.C.2	Mathematization of situations and verification of results	b	c	d	e	f	g	h	i	j	
	MA.3.C.3	Terms, formulas, equations, tables	c	d	e	f	g	h	i	j	k	
Content-related difficulty (CRD)			1	2	3	4	5	6	7	8	9	

Note. Competence levels were aligned based on the basic requirements' (BR) definition for each cycle and the orientation points (OP) for cycles 2 and 3 (cf. Figure 5.1). MA.1 = "number and variable," MA.2 = "form and space," and MA.3 = "measures, functions, data, and probability."

more than one scale unit were represented by the underlying scale units' mean (e.g., CRD = 4.5 for level f of competency MA.1.A.2).

5.2.2 Calibration Design

5.2.2.1 Common item design

To establish a vertical scale for measuring students' abilities in mathematics from third through ninth grades, we developed a common item design, which included 520 mathematics items representing the 18 competencies described in Table 5.1. Table 5.2 provides a macro-level design overview. Generally, we administered a combination of grade-specific and linking items to each grade, with 240 grade-specific items administered to one grade only (highlighted in light blue in Table 5.2) and 280 linking items administered to two adjacent grades (highlighted in dark blue in Table 5.2) to link the various grades. The design contained one exception to the general structure. Due to the broad, overlapping ability range within the eighth and ninth grades, the design disregarded grade-specific items for eighth grade for the benefit of a larger amount of linking items dedicated to both grades.

Table 5.2. Macro-Level Common Item Design.

Target grade level	Samples per grade							N of obs. per item	
	3	4	5	6	7	8	9	per grade	total
3	60							2	2
3–4	40							1	2
4		40						2	2
4–5		40						1	2
5			40					2	2
5–6			40					1	2
6				40				2	2
6–7				40				1	2
7					40			2	2
7–8					40			1	2
8–9 a						40		2	4
8–9 b						40		1	2
9							20	2	2

Note. Grade-specific items are highlighted in light blue; linking items are highlighted in dark blue.

To reduce the workload for individual students, we further divided the items dedicated to each grade over five booklets. Most of the items were included in two booklets, whereas grade-specific items were included in two booklets within one grade group, and linking items

were included in two booklets of adjacent grade groups. To balance the design regarding the number of items per competency, 40 out of the 80 linking items between eighth and ninth grades were included in four different booklets (see Table 5.2 for an overview of the number of observations per item per grade and in total). An extract from the resulting design is displayed in Table 5.3. In total, the design comprised 35 linked booklets with 32 items in each booklet.

Table 5.3. Extract of Micro-Level Common Item Design.

Target grade level	Booklets														
	Grade 3					Grade 4					Grade 5...				
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15
...															
3–4					8			8		8					
	8							8							
		8					8								
			8						8						
				8		8									
4						8	8								
							8	8							
						8			8						
									8	8					
								8		8					
4–5						8				8				8	
							8								8
								8							
							8								
								8							
5													8		8
											8	8			
													8	8	
												8		8	
											8				8
...															
Total	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32

Note. Grade-specific items are highlighted in light blue; linking items are highlighted in dark blue.

5.2.2.2 Distribution of content within design

Four practical requirements guided content distribution within the design. First, we aimed to include approximately 29 items from each of the 18 competencies in the design (i.e., 520 divided by 18 competencies). Second, the number of items per competence level should correspond to the width of the competence levels as displayed in Table 5.1. Third, the booklets should comprise pairs of items related to the same competency to prevent constant switching between topics. Last, but not least, each booklet should include as many different competencies as possible, given the other three requirements. As a result, the 35 booklets comprised, on average, $M = 12.457$ items of the domain “number and variable” ($SD = 1.379$), $M = 8.800$ items of the domain “form and space” ($SD = 1.828$), and $M = 10.743$ items related to the domain “measures, functions, data, and probability” ($SD = 1.686$). Each booklet included, on average, items related to $M = 12.829$ different competencies ($SD = 1.361$), and each competency was represented by $M = 28.889$ items ($SD = 4.351$).

5.2.3 Item Calibration

5.2.3.1 Sample

In spring 2017 and 2018, we administered the 35 mathematics booklets to a sample of 2,733 students from schools in Northwestern Switzerland. The mathematics booklets were available for teachers besides other assessment templates in a first version of the computer-based assessment system MINDSTEPS (Tomasik et al., 2018). Teachers were invited to administer the booklets to their students during school lessons to support the calibration of the systems’ item pool. In line with the American Psychological Association’s Ethical Principles and Code

Table 5.4. Overview of Total Sample for Percentage of Excluded Students with a High Amount of Missing Responses and Calibration Sample per Grade.

Grade	N_{total}	% of $P_{NA} \geq \frac{1}{3}$	N_{calib}
3	284	13.4	246
4	215	21.9	168
5	173	13.3	150
6	392	8.7	358
7	465	7.7	429
8	667	10.2	599
9	537	9.5	486
Total	2,733	10.9	2,436

Note. P_{NA} = Percentage of missing responses for a student on his or her test.

of Conduct, as well as with the Swiss Psychological Society's Ethical Guidelines, written informed consent from students and their parents was not required for this study because it was based on the assessment of normal educational practices and curricula in educational settings (i.e., solving a computer-based mathematics assessment at school; American Psychological Association, 2017; Swiss Psychological Society, 2003). Given that the assessments were very low-stakes for students and teachers, we excluded students with a high percentage of missing responses, i.e., students who didn't answer one-third or more of the items presented to them. This yielded a calibration sample of $N = 2,436$ students. Table 5.4 displays the sample size per grade, as well as the percentage of students excluded prior to the calibration due to too many missing responses.

5.2.3.2 Calibration procedures

The computer-based assessment system automatically scored students' responses, with wrong or omitted responses scored as 0, and correct responses scored as 1. To detect possible calibration problems, we used both concurrent and grade-by-grade calibration procedures to generate marginal maximum likelihood (MML) estimates of the item parameters from the Rasch model (Rasch, 1960). Outcome differences from the two calibration procedures would indicate multidimensionality, thereby suggesting scale instability (e.g., Béguin & Hanson, 2001). For both procedures, we used the software package TAM (Kiefer, Robitzsch, & Wu, 2016) within the development environment R (R Core Team, 2016). First, the dichotomized response data were calibrated concurrently over all school grades. Separate population ability distributions were estimated for each school grade to ensure unbiased parameter estimation (DeMars, 2002; Eggen & Verhelst, 2011). The mean of the sixth-grade population (i.e., the center of the vertical scale) was constrained to 0. Second, the response data were calibrated separately for each school grade. All item parameters from this grade-by-grade calibration subsequently were transformed to the sixth-grade scale using the characteristic curve transformation method by Stocking and Lord (1983, see also González & Wiberg, 2017; Kolen & Brennan, 2014). The transformation resulted in two item-difficulty parameters for each linking item. By means of the unit-weighted average method (McKinley, 1988), we combined the two difficulty estimates for each linking item to one single parameter. To estimate student ability as a basis for determining the populations' mean abilities, the weighted maximum likelihood (WML) method proposed by Warm (1989) was used.

5.2.4 Item Analysis

To investigate whether the items fit a unidimensional vertical Rasch scale and, thus, address our first research question, we analyzed, for each item, the number of observations and response patterns, item discrimination, items' fit to the overall Rasch model, and parameter invariance across school grades. We defined four criteria for identifying problematic items, which we excluded from the final vertical scale. First, items with equal or less than three correct or incorrect responses were excluded because such a low variation within the response pattern

would result in large standard errors in the particular item's difficulty estimate. Moreover, the low variation might indicate a large mismatch between the item's target difficulty as estimated by the content experts during test development and the item's true difficulty. Second, we excluded items with a discrimination of $r_{it} \leq .10$. Third, item fit was analyzed based on the concurrent calibration by means of the root mean square deviation (RMSD), a standardized index of the difference between the expected and observed item characteristic curve (Oliveri & von Davier, 2011). RMSD was calculated separately for each school grade. For linking items assigned to two school grades, we additionally computed the weighted root mean square deviation (WRMSD; von Davier, Weeks, Chen, Allen, & van der Velden, 2016; Yamamoto, Khorramdel, & von Davier, 2016). An RMSD value of 0 indicates a perfect fit of the item to the model, whereas higher values indicate poorer fit. In this study, we rejected items with a $\text{MaxRMSD} > .20$ (i.e., with $\text{RMSD} > .20$ in at least one school grade), as well as linking items with a $\text{WRMSD} > .20$. Fourth, we investigated parameter invariance (e.g., Rupp & Zumbo, 2016) over school grades by comparing linking items' grade-specific item-difficulty estimates. For each linking-item set (e.g., linking items between third and fourth grades), we plotted a 99% confidence interval based on each item pair's mean difficulty, each pair's mean standard error, and the item set's overall mean for each of the two grades (cf. Luppescu, 1991). Items that fell outside the 99% confidence band were excluded. In addition, we computed the Pearson product-moment correlation coefficient between the difficulty estimates related to the two grades. After excluding all misfitting items based on these four criteria, we recalibrated the remaining items. This procedure was repeated until all remaining items fulfilled our evaluation criteria.

5.2.5 Data Analysis

Besides item analyses, we also addressed our first research question related to the fit of the items to a unidimensional vertical Rasch scale by comparing the outcomes of the two calibration procedures. Namely, we compared the populations' mean abilities per school grade, estimated through the two procedures. In addition, we computed Pearson product-moment correlation coefficients between the two procedures' item-difficulty estimates over all items, as well as per school grade. Divergences between the two calibration procedures could indicate multidimensionality (e.g., Béguin & Hanson, 2001) and, thus, scale instability.

To explore the scale's validity from a content perspective and, therefore, address our second and third research questions, we performed the following analyses. First, we investigated the relationship between the difficulty estimates from the final scale and the CRD by means of Pearson product-moment correlation coefficients. Because of the relatively small sample sizes, we used a simulation-based approach to factor in the estimated difficulty parameters' error when calculating the correlations. Namely, we followed five steps:

- (1) We specified a normal distribution $N_i(\hat{\beta}_i, SD[\hat{\beta}_i])$ for each item i , in which $\hat{\beta}_i$ refers to the difficulty parameter of item i from the calibration, and $SD(\hat{\beta}_i)$ refers to the standard deviation calculated based on the standard error of $\hat{\beta}_i$.
- (2) We randomly drew $k = 10,000$ samples S_{ik} of size n_i from N_i , in which n_i was equal to the number of observed responses for item i .
- (3) For each S_{ik} , we calculated the estimate $\hat{\beta}_{ik}^*$ as the mean of S_{ik} , and we stored these estimates in a matrix with $i \times k$ elements.
- (4) For each of the estimates' k samples, we computed the correlation r_k between the estimated difficulty parameters and the CRD, as well as the related p-value p_k .
- (5) Finally, we calculated the estimates r^* and p^* as the mean and related standard errors $SE(r^*)$ and $SE(p^*)$ as the standard deviation of r_k and p_k over the k samples; r^* and p^* were computed not only over the total item pool, but also separately for each curriculum cycle, domain, and competency.

With this approach, we were able to reproduce $\hat{\beta}_i$ and $SE(\hat{\beta}_i)$. For each item, $\hat{\beta}_i^*$ —i.e., the mean over $\hat{\beta}_{ik}^*$ —corresponded to $\hat{\beta}_i$, and $SE(\hat{\beta}_i^*)$ —i.e., the standard deviation of $\hat{\beta}_{ik}^*$ —corresponded to $SE(\hat{\beta}_i)$, resulting in correlations of $r = 1.00$. To compare the resulting correlation coefficients between cycles, domains, and competencies, we performed omnibus tests (Paul, 1989), as well as subsequent range tests (Levy, 1976), by means of the computer program INCOR (Silver, Zaikina, Hittner, & May, 2008).

Second, the variation in item difficulty within each CRD category was investigated by means of boxplots and scatterplots. A boxplot was used to illustrate the general variation in item difficulty, and several scatterplots were used to visualize the relationship between item difficulty and CRD on different curriculum levels (i.e., per cycle, per domain, and per competency).

Finally, we performed three multiple linear regression analyses to investigate possible interaction effects between CRD and (1) curriculum cycles, (2) domains, and (3) competencies on the prediction of the empirical item-difficulty parameters. Similarly, for estimating correlation coefficients, we also used a simulation-based approach to factor in the estimated difficulty parameters' errors for the regression analyses. Steps 1 to 3 were identical to the procedure described above. In step 4, we ran the multiple linear regressions for each of the 10,000 samples of $\hat{\beta}_{ik}^*$ and computed the regression parameters (i.e., adjusted R^2_k , F_k and the related p-value p_{Fk} , B_k and the related p-value p_{Bk}). In step 5, we then calculated the adjusted R^{2*} , F^* , p_F^* , B^* , and p_B^* as the mean and related standard errors as the regression parameters' standard deviation over the k samples.

5.3 Results

5.3.1 Item Calibration

Figure 5.2. Overview of number of observations and scores per item, as well as item-fit statistics. The red line indicates the cut-off value for excluding items from the final scale.

5.3.1.1 Item analysis and exclusion

The median of the number of observations per item corresponded to $Mdn = 113.5$ students ($IQR = 79 - 191.25$) in total, and $Mdn = 62$ students within one grade group ($IQR = 38 - 118$) for linking items. Figure 5.2 provides an overview of the number of observations and scores per item, as well as item-fit statistics from the first calibration run, including all 520 items. Based on these findings, we excluded 48 of the 520 items from the final calibration (i.e., 9.2% of the original item pool), with 18 of these excluded because of low variation in the response data (i.e., number of correct or incorrect responses ≤ 3). These items were either too easy or too difficult for the target population and, thus, could not be estimated accurately. Furthermore, 10 items were excluded because they discriminated very badly between low-ability and high-ability students (i.e., $r_{it} \leq .10$), nine indicated large item misfit (i.e., $WRMSD \geq .20$ or $MaxRMSD \geq .20$) in the first calibration run, and two indicated large item misfit after recalibration in the second and third calibration runs. The correlation between grade-specific item difficulties from adjacent grades is shown in Figure 5.3. The red curves indicate cut-off criteria for excluding items from the final scale; i.e., the upper and lower limit of the 99%

confidence interval. Generally, the item parameters for adjacent grade groups were very highly correlated. The lowest correlation was found between fourth and fifth grades ($r = .816$), and the highest correlation was found between eighth and ninth grades ($r = .962$). Based on the plots in Figure 5.3, nine linking items were excluded due to a large difference between the two grade-specific difficulty parameters.

Figure 5.3. Correlation between grade-specific item difficulties from adjacent grades. The red curves indicate the upper and lower limit of the 99% confidence interval. The blue line refers to the identity line.

In total, four calibration runs were required until all remaining items showed satisfying values on the four evaluation criteria. Table 5.5 provides an overview of selected descriptive statistics from the final item pool. As shown in Table 5.5's third column, item exclusion was rather evenly distributed over the different curriculum cycles and domains, whereas the percentage of excluded items varied between 7% for the domain MA.1 (i.e., "number and variable") and 11% for domain MA.2 (i.e., "form and space"). As for competence levels, we excluded between 0% and 25% of the original items. The largest number of items (i.e., eight) had to be excluded for competency MA.3.C.2 (i.e., "mathematization of situations and

Table 5.5. Final Item Pool's Descriptive Statistics.

Curriculum level	N. items		N. observations		Item difficult β		SE of β	
	N_{final}	$\%_{excl}$	Mdn	IQR	M	SD	M	SD
Overall	472	9%	127.0	81.0 – 192.5	–0.895	2.172	0.247	0.084
Per cycle								
C1 & C2	304	10%	86.0	68.0 – 111.0	–1.715	1.929	0.274	0.075
C3	168	9%	208.0	170.0 – 228.0	0.589	1.766	0.196	0.077
Per domain								
MA.1	188	7%	127.0	83.0 – 194.0	–1.083	2.122	0.242	0.086
MA.2	126	11%	111.0	79.0 – 181.0	–0.824	2.098	0.250	0.084
MA.3	158	10%	116.0	79.5 – 187.0	–0.729	2.283	0.249	0.083
Per competency								
MA.1.A.2	29	3%	132.0	83.0 – 206.0	–1.760	2.027	0.246	0.084
MA.1.A.3	32	6%	149.0	86.0 – 213.8	–1.341	2.088	0.228	0.069
MA.1.A.4	25	11%	96.0	69.0 – 165.0	–0.747	2.008	0.263	0.098
MA.1.B.1	24	8%	109.5	87.0 – 162.8	–1.125	2.144	0.244	0.082
MA.1.B.2	27	4%	90.0	74.0 – 175.5	–1.036	1.633	0.239	0.064
MA.1.C.1	22	12%	127.0	82.3 – 175.0	–0.602	1.825	0.225	0.085
MA.1.C.2	29	6%	145.0	88.0 – 192.0	–0.787	2.829	0.248	0.113
MA.2.A.2	26	4%	106.0	75.3 – 178.3	–1.403	1.773	0.237	0.083
MA.2.A.3	27	10%	121.0	85.5 – 219.5	–0.491	2.967	0.257	0.083
MA.2.B.1	19	17%	96.0	76.0 – 152.0	–1.032	2.080	0.256	0.077
MA.2.C.3	22	15%	104.5	70.3 – 162.8	–0.651	1.697	0.270	0.091
MA.2.C.4	32	11%	127.0	91.0 – 197.8	–0.629	1.704	0.237	0.086
MA.3.A.2	33	3%	149.0	74.0 – 200.0	–1.124	2.193	0.236	0.086
MA.3.A.3	28	7%	144.0	84.0 – 180.8	–0.155	2.271	0.241	0.091
MA.3.B.2	29	9%	144.0	86.0 – 208.0	–0.548	1.500	0.233	0.063
MA.3.C.1	26	13%	130.0	72.5 – 178.3	–1.136	1.711	0.222	0.061
MA.3.C.2	24	25%	96.0	69.0 – 151.5	–0.022	3.564	0.320	0.083
MA.3.C.3	18	0%	90.0	81.0 – 105.0	–1.540	1.776	0.261	0.080

Note. N_{final} = number of items in the final item pool; $\%_{excl}$ = percentage of excluded items from the original item pool; Mdn / IQR = median and interquartile range of number of observations per item.

verification of results”). As indicated in Table 5.5, the items related to competency MA.3.C.2 also had the highest mean difficulty ($M_\beta = -0.022$), the largest variation in difficulty ($SD_\beta = 3.564$), and the largest mean standard error of item difficulty ($M_{SE(\beta)} = 0.320$) of all 18 competencies.

In sum, a satisfying number of items (i.e., 472) fit well into the established vertical Rasch scale, and the scale covered all relevant domains and competencies within the mathematics curriculum’s two target cycles.

5.3.1.2 Comparison of two calibration procedures

To investigate scale stability, we applied both concurrent and grade-by-grade calibration procedures to establish the vertical Rasch scale. Figure 5.4 shows the estimated population means and related 95% confidence intervals based on the final calibration run for the concurrent and grade-by-grade calibration for each of the seven school grades. In both calibration procedures, the sixth-grade population mean was constrained to 0. Both procedures resulted in similar trends for the development of the mean ability from third to ninth grades and showed a steep increase in mathematics ability throughout primary school, followed by a decreased ability progression throughout secondary school. Differences in the groups’ mean ability between the two procedures were largest for third grade (i.e., $M_{conc} = -3.439$; $SE_{conc} = 0.060$, $M_{grade} = -3.557$, $SE_{grade} = 0.059$), decreased from fourth to sixth grade (i.e., both calibrations’ reference grade), and were very small for all secondary-school grades (i.e., seventh to ninth grades). None of the differences was statistically significant as the overlapping confidence intervals indicate in Figure 5.4.

Figure 5.4. Estimated population means based on the final calibration run for the concurrent and grade-by-grade calibration, including the 95% confidence interval.

Figure 5.5. Comparison of item-difficulty parameters from the concurrent and grade-by-grade calibration for the total item pool and per grade. Gray circles = original $\hat{\beta}$ per grade; black circles = transformed $\hat{\beta}$ per grade, with averaged parameters for linking items.

Furthermore, Figure 5.5 compares the item-difficulty parameters from the two calibration procedures for the overall item pool, as well as separately for each grade. For the grade-by-grade calibration on the y-axis, the graphs per grade include two different item-difficulty parameters for each item, namely, the original grade-specific, item-difficulty parameter (in gray) and the parameter after the Stocking and Lord transformation, including the averaging of linking items (in black). The gray and red lines refer to the grade-specific identity

lines for the original and transformed item parameters, respectively. In line with Figure 5.4, Figure 5.5 indicates that the item-difficulty parameters from the grade-by-grade calibration corresponded very much to the parameters from the concurrent calibration. The match between the parameters from the two procedures was especially high after transforming and averaging the linking items' grade-specific item parameters over two school grades.

In sum, the two calibration procedures' results were very similar and didn't indicate any technical problems regarding item calibration. We concluded from these findings, in combination with the results from the item analysis, that the established unidimensional vertical Rasch scale showed a good fit and referred to a stable scale from a psychometric perspective. Therefore, we used only the parameters from the concurrent calibration in our analyses for validating the vertical scale from a content perspective.

5.3.2 Content-Related Validation of the Vertical Scale

5.3.2.1 Overall relationship between item difficulty and CRD

To address our second research question, regarding the validation of the vertical scale from a content perspective, we investigated the correlation between the empirical item-difficulty estimates from the concurrent calibration and the CRD as defined based on the matrix in Table 5.1 on different curriculum levels. Table 5.6 summarizes the numbers of items N per analysis, the estimates for the Pearson correlation coefficients r^* , and the related standard errors $SE(r^*)$ based on the simulation, as well as the estimated p -values p^* for the correlation coefficients and their standard errors $SE(p^*)$.

In line with our hypothesis, we found a strong, significant positive correlation of $r^*(472) = .672$ with $p^* < .001$ between the CRD and the difficulty parameters from the concurrent calibration over the whole item pool ($SE(r^*) = .004$; $SE(p^*) = .000$). Overall, items related to more advanced competence levels of the curriculum had higher difficulty estimates than items related to more basic competence levels. This finding also is supported in Figure 5.6, which illustrates the difficulty parameters' distribution within each CRD category over all competencies through boxplots. On top of Figure 5.6, the different colors represent the three curriculum cycles. The boxplots show considerable overlap of difficulties between the different CRD categories. Nevertheless, boxplots generally are located lower on the difficulty scale for low than for high CRD categories, reflecting the positive correlation between the empirical and theoretical difficulties reported in Table 5.6.

Table 5.6. Estimated Pearson Correlation Coefficients Between CRD and Difficulty Estimates from Concurrent Calibration over all Items, per Cycle, per Domain, and per Competency.

Curriculum level	<i>N</i>	<i>r</i> [*]	<i>SE</i> (<i>r</i> [*])	<i>p</i> [*]	<i>SE</i> (<i>p</i> [*])
Overall	472	.672	.004	0	0
Per cycle					
C1 & C2	304	.622	.007	0	0
C3	168	.220	.008	.004	.002
Per domain					
MA.1	188	.706	.006	0	0
MA.2	126	.603	.009	0	0
MA.3	158	.707	.006	0	0
Per competency					
MA.1.A.2	29	.742	.014	0	0
MA.1.A.3	32	.741	.013	0	0
MA.1.A.4	25	.586	.025	.002	.001
MA.1.B.1	24	.670	.020	0	0
MA.1.B.2	27	.701	.022	0	0
MA.1.C.1	22	.823	.016	0	0
MA.1.C.2	29	.742	.012	0	0
MA.2.A.2	26	.626	.019	.001	0
MA.2.A.3	27	.827	.010	0	0
MA.2.B.1	19	.433	.024	.066	.017
MA.2.C.3	22	.179	.029	.430	.077
MA.2.C.4	32	.636	.022	0	0
MA.3.A.2	33	.784	.013	0	0
MA.3.A.3	28	.784	.017	0	0
MA.3.B.2	29	.402	.029	.033	.014
MA.3.C.1	26	.746	.017	0	0
MA.3.C.2	24	.894	.008	0	0
MA.3.C.3	18	.734	.028	.001	.001

Note. *r*^{*} significant at the 5% level under consideration of *SE*(*p*^{*}) are printed in bold.

Figure 5.6. Boxplots illustrating the distribution of difficulty estimates from the concurrent calibration for each CRD category. C1, C2, and C3 refer to cycles 1, 2, and 3, respectively.

5.3.2.2 Relationship between item difficulty and CRD per curriculum cycle

Besides general correlation, we also were interested in the correlation between item difficulty and CRD within the curriculum's different cycles, domains, and competencies (see third research question). Contrary to our hypothesis, Figure 5.6 indicates that the correlation between the difficulty estimates and CRD was significantly stronger for cycles 1 and 2 (i.e., primary school) than for cycle 3 (i.e., secondary school). The estimated correlation coefficients per curriculum cycle reported in the top part of Table 5.6 confirmed this assumption: The estimated correlation coefficients were $r^*(304) = .622$ ($p^* < .001$) and $r^*(168) = .220$ ($p^* < .01$) for the first two cycles and for cycle 3, respectively. The standard errors of r^* and p^* , estimated based on the simulations, were small and, thus, didn't affect the interpretation of the results (see Table 5.6). Further analysis indicated that the correlation between the difficulty estimates and CRD within cycles 1 and 2 was significantly stronger than those within cycle 3 ($X^2_{C(F)} = 26.157$, $p > .001$; cf. Paul, 1989). In line with this finding, the regression lines of the difficulty estimates and each cycle's CRD indicate an interaction effect from CRD and curriculum cycle on the estimated item-difficulty parameters (see left graph of Figure 5.7). The results of the multiple regression analysis presented in Table 5.7 also strengthened this finding. We found a significant regression equation ($adjusted R^{2*} = .461$, $F^*(3,468) = 135.064$, $p^* < .001$), which included a significant negative interaction effect from CRD and curriculum cycle on item difficulty ($B^* = -0.357$, $p^* < .01$). In addition, we found significant positive main effects from CRD ($B^* = 0.697$, $p^* < .001$) and curriculum cycle⁸ ($B^*_{C3} = 2.204$, $p^* < .05$). Taken together, we concluded from

⁸ The combination of cycles 1 and 2 served as a base level for the dummy coding.

these results that the vertical scale represented the competencies stated in the curriculum better for cycles 1 and 2 than for cycle 3.

Figure 5.7. Relationship between difficulty estimates and CRD and related regression lines per cycle (left) and per domain (right).

5.3.2.3 Relationship between item difficulty and CRD per curriculum domain

In the middle part, Table 5.6 includes the correlation between CRD and item difficulty for each of the three curriculum domains. All three correlations were statistically significant ($p^* < .001$), and the related estimated standard errors $SE(r^*)$ and $SE(p^*)$ were negligibly small (see Table 5.6). The correlation was slightly weaker for the domain MA.2 (i.e., “form and space”) with $r^*(126) = .603$ than for the other two domains MA.1 (i.e., “number and variable”) and MA.3 (i.e., “measures, functions, data, and probability”) with $r^*(188) = .706$ and $r^*(158) = .707$, respectively. However, in line with our hypothesis, the differences between the correlation coefficients were not statistically significant ($X^2_{C(F)} = 3.040, p = .219$). Similarly, the right graph in Figure 5.7 shows that the three domains’ regression lines are rather parallel and, thus, don’t indicate any interaction effect from CRD and domain on item difficulty. Moreover, a significant regression equation was found for the multiple regression analysis of CRD and domain on item difficulty ($adjusted R^{2*} = .462, F^*(5,466) = 82.059, p^* < .001$), which indicated that neither the interaction effects of CRD and the domains⁹ ($B^*_{CRD \times MA.2} = -0.079, p^* = .283; B^*_{CRD \times MA.3} = 0.043, p^* = .521$) nor the domains’ main effects were statistically significant ($B^*_{MA.2} = 0.823, p^* = .061; B^*_{MA.3} = 0.367, p^* = .362$; see Table 5.7). CRD was the only significant predictor of item

⁹ MA.1 served as a base level for the dummy coding.

Table 5.7. Results from Multiple Linear Regression Analyses for Predicting Empirical Item Difficulty by CRD and by the Curriculum Levels Cycle, Domain, and Competencies.

Model for cycles	Main effects				Interaction effects with CRD			
	<i>Est</i> [*]	<i>SE</i> (<i>Est</i> [*])	<i>p</i> [*]	<i>SE</i> (<i>p</i> [*])	<i>Est</i> [*]	<i>SE</i> (<i>Est</i> [*])	<i>p</i> [*]	<i>SE</i> (<i>p</i> [*])
Adjusted <i>R</i> ²	0.461	0.006						
<i>F</i>	135.064	3.122	0	0				
Constant	−4.372	0.038	0	0				
CRD	0.697	0.008	0	0				
Cycle(C3)	2.204	0.113	.017	.006	−0.357	0.015	.003	.001

Model for domains	Main effects				Interaction effects with CRD			
	<i>Est</i> [*]	<i>SE</i> (<i>Est</i> [*])	<i>p</i> [*]	<i>SE</i> (<i>p</i> [*])	<i>Est</i> [*]	<i>SE</i> (<i>Est</i> [*])	<i>p</i> [*]	<i>SE</i> (<i>p</i> [*])
Adjusted <i>R</i> ²	0.462	0.006						
<i>F</i>	82.059	1.812	0	0				
Constant	−4.307	0.051	0	0				
CRD	0.581	0.007	0	0				
Domain(MA.2)	0.823	0.076	.061	.024	−0.079	0.011	.283	.070
Domain(MA.3)	0.367	0.069	.362	.091	0.043	0.010	.521	.100

Model for comp.	Main effects				Interaction effects with CRD			
	<i>Est</i> [*]	<i>SE</i> (<i>Est</i> [*])	<i>p</i> [*]	<i>SE</i> (<i>p</i> [*])	<i>Est</i> [*]	<i>SE</i> (<i>Est</i> [*])	<i>p</i> [*]	<i>SE</i> (<i>p</i> [*])
Adjusted <i>R</i> ²	0.528	0.006						
<i>F</i>	16.033	0.371	0	0				
Constant	−3.717	0.125	0	0				
CRD	0.510	0.019	0	0				
Comp(MA.1.A.2)	−1.558	0.194	.113	.046	0.229	0.032	.205	.064
Comp(MA.1.A.3)	−1.155	0.184	.221	.074	0.071	0.026	.640	.122
Comp(MA.1.A.4)	0.414	0.202	.673	.144	−0.036	0.030	.812	.118
Comp(MA.1.B.2)	0.115	0.174	.862	.099	−0.025	0.026	.858	.097
Comp(MA.1.C.1)	−1.268	0.214	.295	.082	0.199	0.030	.300	.074
Comp(MA.1.C.2)	−1.006	0.171	.261	.082	0.144	0.024	.316	.081
Comp(MA.2.A.2)	0.072	0.163	.878	.089	−0.027	0.027	.856	.096
Comp(MA.2.A.3)	−2.508	0.188	.013	.007	0.412	0.028	.010	.005
Comp(MA.2.B.1)	1.267	0.168	.174	.058	−0.206	0.025	.211	.056
Comp(MA.2.C.3)	2.258	0.213	.049	.022	−0.344	0.033	.105	.034
Comp(MA.2.C.4)	0.639	0.174	.495	.118	−0.082	0.025	.598	.114
Comp(MA.3.A.2)	−0.606	0.158	.483	.115	0.077	0.023	.597	.110
Comp(MA.3.A.3)	−0.460	0.180	.636	.131	0.238	0.028	.152	.048
Comp(MA.3.B.2)	1.789	0.177	.065	.027	−0.274	0.026	.083	.030
Comp(MA.3.C.1)	−0.113	0.174	.861	.101	0.037	0.027	.818	.109
Comp(MA.3.C.2)	−2.517	0.183	.009	.005	0.743	0.032	0	0
Comp(MA.3.C.3)	0.084	0.168	.873	.093	0.070	0.028	.724	.106

Note. *N* = 472; *Est*^{*} = estimate, representing *R*²*, *F*^{*}-statistic and *B*^{*}-values, comp = competency; *Est*^{*} significant at the 5% level under consideration of *SE*(*p*^{*}) are printed in bold; base level for dummy coding: cycle = C1 & C2, domain = MA.1, competency = MA.1.B.1.

difficulty in this model ($B^* = 0.581, p^* < .01$). Thus, we concluded from these findings that the vertical scale represented all three domains of the curriculum equally well.

5.3.2.4 Relationship between item difficulty and CRD per competency

Finally, as a third aspect within our third research question, we investigated the correlation between CRD and item difficulty for each of the 18 competencies represented by the vertical scale. The correlation coefficients, which we estimated based on our simulations, are displayed at the bottom of Table 5.6. The correlations varied between $r^*(22) = .179$ ($p^* = .430$) for competency MA.2.C.3, referring to “geometric figures and objects in different positions,” and $r^*(24) = .894$ ($p^* < .001$) for competency MA.3.C.2, referring to “mathematization of situations and verification of results.” In total, 15 of the 18 correlation coefficients were statistically significant at the 5% level when considering the standard errors of the estimated p-values $SE(p^*)$ (i.e., $p^* + 1.96 \times SE(p^*) < .05$). Besides the competency MA.2.C.3, we found insignificant correlations for the competencies MA.2.B.1, referring to “exploration of lengths, surfaces, and volumes” ($r^*(19) = .433, p^* = .066$), and MA.3.B.2, referring to “statistics, combinatorics, and probability” ($r^*(29) = .402, p^* = .033, SE(p^*) = .014$). The omnibus test for correlations’ equality indicated that significant differences existed between the different competencies’ correlation coefficients ($X^2_{C(F)} = 31.261, p < .05$). Subsequent range tests (Levy, 1976) showed that the correlation within competency MA.2.C.3 was significantly lower than all 15 significant correlations ($p < .05$). The other two insignificant correlations of competencies MA.2.B.1 and MA.3.B.2 were significantly lower than all correlations corresponding to $r > .701$ and $r > .670$, respectively ($p < .05$; i.e., 10 and 11 competencies, respectively, of the 18 competencies). On the other hand, the correlation of competency MA.3.C.2 (i.e., the highest of all correlations) was significantly higher than correlations corresponding to $r < .746$ ($p < .05$), which applied to 13 of the 18 competencies. Furthermore, the range tests showed that the correlation of competency MA.1.B.1 (i.e., “numbers and operations”), which showed an average correlation between item difficulty and CRD ($r^*(24) = .670, p^* < .001$), significantly differed only from the correlations of competencies MA.2.C.3 and MA.3.C.2 (i.e., the competencies with the lowest and highest correlations; $p < .05$).

Furthermore, we regressed CRD and the competencies on item difficulty. To facilitate the interpretation of the results from this multiple regression analysis, we specified the competency MA.1.B.1 (i.e., the competency with an average correlation coefficient) as the base level for the competencies’ dummy coding. This specification allowed us to detect competencies that deviated from the general pattern. As reported in Table 5.7, the regression equation was significant ($adjusted R^{2*} = .528, F^*(35,436) = 16.033, p^* < .001$), including significant negative main effects, as well as significant positive interaction effects with CRD on item difficulty for competencies MA.2.A.3, referring to “computation of lengths, surfaces, and volumes” ($B^*_{MA.2.A.3} = -2.508, p^* < .05, SE[p^*] = .007$; $B^*_{CRD \times MA.2.A.3} = 0.412, p^* < .05, SE[p^*] = .005$) and MA.3.C.2, referring to “mathematization of situations and verification of results” (i.e., the competency with the strongest correlation between item difficulty and CRD;

$B^*_{MA.3.C.2} = -2.517, p^* < .05, SE[p^*] = .005; B^*_{CRD \times MA.3.C.2} = 0.743, p^* < .001$). In addition, we found a significant main effect from CRD ($B^* = 0.510, p^* < .001$). All other competencies' main and interaction effects were not significant.

Figure 5.8. Relationship between difficulty estimates and CRD and related regression lines by competency. Bold blue lines = regression lines, gray lines = regression line of competencies MA.1.B.1, white background = competencies with insignificant correlations, and gray background = competencies with significant interaction effects.

Figure 5.8 visualizes the relationship between item difficulty and CRD within the different competencies of the curriculum. As a reference, each of the 18 scatterplots includes the regression line of the “average” competency MA.1.B.1 (gray line) beside the competency-specific regression line (blue line). The three competencies' scatterplots with insignificant correlations (i.e., MA.2.B.1, MA.2.C.3, and MA.3.B.2) are highlighted with purely white backgrounds. All three plots showed considerable overlap in item-difficulty estimates between the different CRD categories, and the related regression lines had lower slopes than the other 15 competencies' regression lines. The plots related to competencies MA.2.B.1 and MA.2.C.3 indicate, in addition, that these two competencies were poorly represented within cycle 3. In the curriculum (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016a), these two

competencies' level descriptions—and especially those related to cycle 3—are very abstract and complex (e.g., “students can formulate hypotheses while exploring geometrical relationships”), and partly refer to additional tools, such as dedicated computer software (e.g., “students can use dynamic geometry software to explore geometrical relationships”). Therefore, content experts only could develop a very limited number of items representing these levels. Furthermore, the curriculum only states five very broad competence levels for MA.2.C.3 and only seven competence levels, including one very broad one within cycle 2 for competency MA.3.B.2 (see Table 5.1). The low number of CRD categories might not map the variation in item difficulty sufficiently and, thus, probably also doesn't reflect the development of students' competence levels within these categories.

The scatterplots of the two competencies with significant main and interaction effects according to the regression analysis (i.e., MA.2.A.3 and MA.3.C.2) are highlighted in gray in Figure 5.8. These two competencies' regression lines differed from the general pattern in their steep slopes. On one hand, these two competencies showed the highest variation in item difficulty among all competencies, as indicated by the standard deviations of β in Table 5.5. On the other hand, compared with most of the other competencies, higher CRD and, thus, more advanced competence levels were related to clearly higher empirical item difficulties than lower CRD within these two competencies. This especially was true for competency MA.3.C.2, in which all items related to the highest competence level in the item pool (i.e., level f) had higher item-difficulty parameters than the remaining items. At the same time, the items related to level MA.3.C.2.f were among the most difficult items of the whole item pool, which, in turn, explains the high mean and large standard deviation in item difficulty within this competency compared with the other competencies (see Table 5.5).

The slopes of the remaining 13 competencies' regression lines were comparable with the slope of the reference competency MA.1.B.1's regression line. In conclusion, our findings indicate that 15 of the 18 mathematics competencies were well-reflected by our vertical scale, whereas we found a particularly strong connection between theoretical and empirical item difficulty for two of these 15 competencies.

5.4 Discussion

To assess students' abilities over the course of compulsory school and to evaluate their progress over time, a vertical measurement scale is essential, which allows for comparing scores from different measurement occasions (Briggs, 2013; Harris, 2007; Kolen & Brennan, 2014; Young, 2006). At the same time, a vertical scale is an important feature for identifying the most informative items from an item bank to assess students over a broad ability range (Dadey & Briggs, 2012; Tomasik et al., 2018), independent of whether the items are selected by a CAT algorithm or manually filtered based on difficulty by teachers or students. However, the development of a vertical scale is psychometrically challenging. Several extant studies point to the complex interactions between the practical context in which the scale is used and the

decisions that researchers need to make during the development of a vertical scale, e.g., the selection of the calibration procedure, data-collection design, or linking items (Béguin et al., 2000; Briggs & Weeks, 2009; Dadey & Briggs, 2012; Harris, 2007; Keller & Hambleton, 2013; Tong & Kolen, 2007). Given these complex interactions, no clear general recommendations exist for most of the scaling decisions (e.g., Briggs & Weeks, 2009; Harris, 2007; Tong & Kolen, 2007).

In this study, we described the development of a vertical scale for formative assessment of students' mathematics abilities from third through ninth grade based on IRT methods, and we evaluated this scale from a psychometric perspective regarding item and model fit, as well as from a content-related perspective regarding the underlying content framework, i.e., the curriculum, *Lehrplan 21* (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014, 2016a, 2016b). Hence, compared with most previous studies investigating the development of vertical scales, we not only looked at the scale's psychometric properties, but also considered the scale's validity for measuring the target construct, i.e., students' ability in mathematics as described by the curriculum. For this purpose, we compared the empirical item difficulties as estimated on the vertical scale with content-related item difficulties based on content experts' ratings. Thanks to our theory-based external validation criterion (i.e., the items' content-related difficulty regarding the curriculum), we were able to validate the specific combination of scaling decisions that we took during the development of the vertical scale within our specific application and context.

Results from item analysis indicated a satisfactory fit for more than 90% of the items from the initial item pool to the final unidimensional vertical Rasch scale. Given the Rasch model's strong assumptions, it isn't surprising that we had to exclude some items during the calibration process for establishing the scale. Nevertheless, we didn't find any systematic patterns of item exclusion across curriculum cycles or domains, and the final scale included items from all relevant competencies. Furthermore, we found strong correlations between item-difficulty parameters from adjacent school grades, as well as between the item-difficulty parameters from concurrent and grade-by-grade calibration procedures. These findings confirm the scale's unidimensionality and, thus, its stability across grades from a psychometric perspective (Hanson & Béguin, 2002; Kolen & Brennan, 2014).

From a content perspective, we found a strong positive correlation between the empirical item-difficulty parameters on our vertical Rasch scale and the content-related item difficulties based on the mapping of the items to the curriculum's competence levels by the content experts, who developed the items. As intended, items related to more advanced competence levels generally were represented by higher item difficulties on the vertical scale. However, we also found large variations in item difficulty within each content-related item-difficulty category and, therefore, a strong overlap in item-difficulty variation across the different content-related difficulty categories. At the same time, the large variation in item difficulty within the categories corresponds to the large variation in student ability within

grades. In their longitudinal study investigating the development of mathematics and German abilities during compulsory school, Angelone, Keller, and Moser (2013, p. 35) found that the standard deviation of student ability in mathematics within one grade corresponded to more than twice the average learning progress per school year (see also Stevens, Schulte, Elliott, Nese, & Tindal, 2015). Item developers might have mirrored this wide variation in abilities within a grade by creating items of different difficulty for single competence levels. However, further studies are needed to get to the bottom of this assumption.

Further analyses demonstrated a stronger correlation between empirical and content-related item difficulty for primary school (i.e., third through sixth grade, covered in the curriculum by the end of cycle 1 and the entire cycle 2) than for secondary school (i.e., cycle 3), which contradicted our hypothesis and the vertical scale's objective. The significant difference in the two correlation coefficients' strength implies that the vertical scale is more representative of the competencies described for primary school than those related to secondary school. However, we didn't find any indication of multidimensionality in our psychometric analyses, nor in our item analysis, nor in the comparison of the two calibration procedures, which were both in favor of a unidimensional stable vertical scale. One possible conclusion could be that the competence levels stated for cycle 3 differ less in difficulty than those stated for cycles 1 and 2. This assumption is in line with our finding of a steeper increase in mathematics ability throughout primary school, followed by a stagnation in ability progression throughout secondary school. Extant studies have reported similar learning trajectories (e.g., Angelone et al., 2013; Bloom, Hill, Black, & Lipsey, 2008; Moser, Oostlander, & Tomasik, 2017; Stevens et al., 2015). In particular, competence levels from cycles 1 and 2 within one competency might reflect competence development, strictly speaking, whereas competence levels of cycle 3 might differ in terms of content and complexity, rather than in pure difficulty (Angelone et al., 2013; Moser et al., 2017; Yen, 1985). Apart from that, it also might be more difficult to describe the competencies' more complex development on secondary-school levels by means of concrete competence descriptions and to differentiate clearly between distinct levels. Furthermore, it might be more challenging to create items targeted at higher competence levels and to predict their true difficulty (cf. Sydorenko, 2011). Further studies with a stronger focus on mathematics didactics and competence development are required to evaluate these hypotheses and investigate whether the competence levels described for secondary school fulfill the basic precondition for a unidimensional vertical scale of a continuous increase in the target competence over time (Young, 2006).

More detailed analyses related to the curriculum domains showed no significant differences in the correlation between empirical and content-related item difficulty across the three domains. The correlations within the domains were similar and corresponded to the general overall correlation. We concluded from these results that the vertical scale satisfactorily represented all three domains. Similar conclusions also could be drawn for most of the 18 competencies in our study, which revealed comparable correlations between empirical and content-related item difficulty. On the other hand, some findings provided interesting input for

further research. Three selected, rather abstract and complex curriculum competencies were represented poorly on the general vertical mathematics scale. At the same time, two of the 18 competencies showed remarkably strong relationships between empirical and content-related item difficulty. For one thing, these results might be related to the various competencies' content-specific characteristics. However, they also could be related to our finding that the vertical scale was less representative in measuring mathematics ability in secondary- than primary-level students. Creating difficult items related to higher-level competencies might be especially difficult for selected competencies, while other competencies might provide better foundations. Unfortunately, we are unable to investigate these interactions further based on our data set, which certainly included a large item pool, but still contained only a limited number of items representing each competency and, especially, each of the numerous competence levels. Therefore, further studies are needed to replicate our results with a larger item pool, including larger samples of items for each competence level, thereby allowing for drawing final conclusions about the causes of differences between competencies.

5.4.1 Limitations

Besides the limited size of the sample of items, other limitations in the present study should be noted. First, we included only one variation of a particular scaling decision in our analyses. Namely, we exemplarily contrasted concurrent and grade-by-grade calibration procedures' outcomes to detect possible calibration problems (Hanson & Béguin, 2002; Kolen & Brennan, 2014). The two procedures' outcomes turned out to be comparable, which verified our scale's stability. In addition, it also would be possible to vary the IRT model, data-collection design (i.e., equivalent group designs, common item designs, or scaling test designs), or number of linking items. However, the latter two variations would require a much more complicated study design, as well as larger student samples, which went beyond our study's scope. Nevertheless, in case of contradicting outcomes, an external content-related validation criterion like the one we used in our analyses could help identify those scaling decisions, which best fit a particular practical context.

Second, our study's focus lied in the validation of the empirical, vertical Rasch scale, while we didn't question the validity of the competence levels stated in the curriculum or the representativeness of the items for their assigned competence levels. Curricula or content standards, which serve as a theoretical basis for test-content specifications, often lack empirical validation of the stated domains, competencies, and competence levels and, especially, of their development over time (Fleischer, Koeppen, Kenk, Klieme, & Leutner, 2013). Thus, studies with a design like ours also could contribute to the validation of theoretical assumptions about competence development and the quality of assessment items. For example, the comparison of empirical and content-related item difficulty could be used to detect competence descriptions, which do not follow a continuous increase in difficulty over time, or test items, which do not reach the expected empirical difficulty (i.e., items deviating from the general regression line

between empirical and content-related difficulty) because they do not fit the underlying content specifications or suffer from technical problems. Even though it limits interpretation of the results to some extent that neither the theoretical framework nor the empirical vertical scale is completely bias-free, it also refers to a huge opportunity to connect the two dimensions for reciprocal validation and for gradually improving both dimensions.

A third limitation of our study is that we used only limited item formats (i.e., simple dichotomous items suitable for automated scoring). More advanced item formats allowing for assessing cognitive processes or interactions might contribute to better representing more complex competencies or competence levels. However, the question arises of whether such complex competencies are still representable by a unidimensional vertical scale or whether they would require more complex measurement models. Further studies are required to answer these questions.

Finally, it is also important to note that we developed a unidimensional vertical scale for assessing general mathematics ability. By excluding items with large drift between two age groups, we also might have excluded selected content areas that develop differently over the years in school than the average competencies (Taherbhai & Seo, 2013). Therefore, in practice, it is important to complement formative assessments based on such a general scale by additional diagnostic assessments that allow for detecting growth in very specific competencies, as well as differences in growth between competencies (e.g., Betebenner, 2009).

5.4.2 Conclusion

In sum, our study emphasizes the benefits of combining psychometric and, thus, technical verification with content-related and, thus, theoretical validation when developing a vertical scale for measuring students' abilities across multiple school grades. Studies related to other subjects might result in different correlation patterns across different curriculum levels (i.e., cycles, domains, and competencies). However, the procedure that we suggested for validating the vertical scale can be transferred to all subjects for which the curriculum states a continuous increase in ability or content difficulty. Technical procedures such as item analysis and the comparison of different calibration procedures helped us identify items with misfit and underpin the scale's stability. However, only contrasting the empirical item difficulties with content-related ones provided information about the adequacy of our decisions during the scaling process for a particular practical context and the scale's validity in representing the underlying content framework (i.e., the curriculum). Therefore, our study points out the importance of a close collaboration and discussion between psychometricians and content experts to develop valid and, thus, meaningful vertical scales.

5.5 References

- American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Washington, DC: American Psychological Association. Retrieved from American Psychological Association website: <https://www.apa.org/ethics/code>
- Angelone, D., Keller, F., & Moser, U. (2013). *Entwicklung schulischer Leistungen während der obligatorischen Schulzeit: Bericht zur vierten Zürcher Lernstandserhebung zuhanden der Bildungsdirektion des Kantons Zürich* [Development of school performance during compulsory school: Report on the fourth assessment for the attention of the Zurich department of education]. Zürich, CH: Institut für Bildungsevaluation (IBE). Retrieved from Institut für Bildungsevaluation (IBE) website: http://www.ibe.uzh.ch/projekte/lezh/Lernstandserhebung_9KlasseZH_Bericht.pdf
- Béguin, A. A., & Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.
- Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the 2000 annual meeting of the National Council of Measurement in Education, New Orleans, IL.
- Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51. doi:10.1111/j.1745-3992.2009.00161.x
- Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4), 289–328. doi:10.1080/19345740802400072
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204–226. doi:10.1111/jedm.12011
- Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3–14. doi:10.1111/j.1745-3992.2009.00158.x
- Brown, G. T. L. (2013). AsTTle – A national testing system for formative assessment: How the national testing policy ended up helping schools and teachers. In S. Kushner, M. Lei, & M. Lai (Eds.), *Advances in Program Evaluation: Volume 14. A national developmental and negotiated approach to school self-evaluation* (Vol. 14, pp. 39–56). Bradford, UK: Emerald Group Publishing Limited. doi:10.1108/S1474-7863(2013)0000014003
- Bundesinstitut für Bildungsforschung, Innovation & Entwicklung (BIFIE) (Ed.). (2011). *Kompetenzorientierter Unterricht in Theorie und Praxis* [Competence orientation in theory and practice]. Graz, AT: Leykam. Retrieved from https://www.bifie.at/wp-content/uploads/2017/06/bist_vs_sek1_kompetenzorientierter_unterricht_2011-03-23.pdf

- Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability*, 21, 47–65. doi:10.1007/s11092-008-9063-x
- Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, 33(3), 378–398. doi:10.3102/0162373711412765
- Cawelti, G., & Protheroe, N. (2001). *High student achievement: How six school districts changed into high-performance systems*. Arlington, VA: Educational Research Service.
- Cizek, G. J. (2005). Adapting testing technology to serve accountability aims: the case of vertically moderated standard setting. *Applied Measurement in Education*, 18(1), 1–9. doi:10.1207/s15324818ame1801_1
- Dadey, N., & Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment, Research & Evaluation*, 17(14), 1–13. Retrieved from <http://pareonline.net/getvn.asp?v=17&n=14>
- de Ayala, R. J. (2009). *The theory and practice of item response theory. Methodology in the social sciences*. New York: Guilford Press.
- DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, 15(1), 15–31. doi:10.1207/S15324818AME1501_02
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2014). *Lehrplan 21: Rahmeninformationen*. Luzern. Retrieved from http://www.lehrplan.ch/sites/default/files/lp21_rahmeninformation_%202014-11-06.pdf
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2016a). *Lehrplan 21: Mathematik*. Luzern. Retrieved from https://v-fe.lehrplan.ch/container/V_FE_DE_Fachbereich_MA.pdf
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2016b). *Lehrplan 21: Überblick*. Luzern. Retrieved from https://v-fe.lehrplan.ch/container/V_FE_Ueberblick.pdf
- Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica*, 32(1), 107–132.
- Ferrara, S., Johnson, E., & Chen, W.-H. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education*, 18(1), 35–59. doi:10.1207/s15324818ame1801_3
- Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft*, 16(S1), 5–22. doi:10.1007/s11618-013-0379-z
- Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, 35(2), 83–88. doi:10.1016/j.stueduc.2009.10.006

- González, J., & Wiberg, M. (2017). *Applying test equating methods*. Cham, Switzerland: Springer International Publishing.
- Hanson, B. A., & Béguin, A. A. (1999). *Separate versus concurrent estimation of IRT parameters in the common item equating design* (ACT Research Report Series No. 99-8). Iowa City, IA: ACT.
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3–24. doi:10.1177/0146621602026001001
- Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–251). New York: Springer.
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.
- Hattie, J. A. C., & Brown, G. T. L. (2007). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, 36(2), 189–201. doi:10.2190/ET.36.2.g
- Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487
- Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*, 21(3), 187–206. doi:10.1080/08957340802161741
- Keller, L. A., & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement*, 50(4), 390–407. doi:10.1111/jedm.12025
- Kiefer, T., Robitzsch, A., & Wu, M. L. (2016). TAM: Test analysis modules.
- Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22(2), 131–143. doi:10.1177/01466216980222003
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.
- Lai, M. K., McNaughton, S., Amituanai Toloa, M., Turner, R., & Hsiao, S. (2009). Sustained acceleration of achievement in reading comprehension: The New Zealand experience. *Reading Research Quarterly*, 44(1), 30–56. doi:10.1598/RRQ.44.1.2
- Lei, P.-W., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement*, 36(1), 21–39. doi:10.1177/0146621611425171
- Levy, K. J. (1976). A multiple range procedure for independent correlations. *Educational and Psychological Measurement*, 36(1), 27–31. doi:10.1177/001316447603600103

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
- Luppescu, S. (1991). Graphical diagnosis. *Rasch Measurement Transactions*, 5(1), 136. Retrieved from <https://www.rasch.org/rmt/rmt51j.htm>
- McKinley, R. L. (1988). A comparison of six methods for combining multiple IRT item parameter estimates. *Journal of Educational Measurement*, 25(3), 233–246. doi:10.1111/j.1745-3984.1988.tb00305.x
- Moser, U., Oostlander, J., & Tomasik, M. J. (2017). Soziale Ungleichheiten im Leistungszuwachs und bei Bildungsübergängen [Social disparities in performance gains and transitions probabilities]. In M. P. Neuschwander & C. Nägele (Eds.), *Bildungsverläufe von der Einschulung bis in den ersten Arbeitsmarkt. Theoretische Ansätze, empirische Befunde und Beispiele* (pp. 59–77). Wiesbaden, Germany: Springer VS.
- Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, 53(3), 315–333. doi:10.2307/3314850
- Paul, R. S. (1989). Test for the equality of several correlation coefficients. *The Canadian Journal of Statistics*, 17(2), 217–227.
- Pohl, S., Haberkorn, K., & Carstensen, C. H. (2015). Measuring competencies across the lifespan: Challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Springer proceedings in mathematics & statistics: volume 145. Dependent data in social sciences research. Forms, issues, and methods of analysis* (pp. 281–308). Cham, Switzerland: Springer.
- Pomplun, M., Omar, M. H., & Custer, M. (2004). A comparison of Winsteps and Bilog-Mg for vertical scaling with the Rasch model. *Educational and Psychological Measurement*, 64(4), 600–616. doi:10.1177/0013164403261761
- R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Reusser, K. (2014). Kompetenzorientierung als Leitbegriff der Didaktik [Competence orientation as a key concept of teaching]. *Beiträge zur Lehrerinnen- und Lehrerbildung*, 32(3), 325–339.
- Rupp, A. A., & Zumbo, B. D. (2016). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66(1), 63–84. doi:10.1177/0013164404273942
- Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education: Challenges and opportunities*. Dordrecht, the Netherlands: Springer.

- Silver, N. C., Zaikina, H., Hittner, J. B., & May, K. (2008). incor: a computer program for testing differences among independent correlations. *Molecular ecology resources*, 8(4), 763–764. doi:10.1111/j.1755-0998.2008.02107.x
- Stevens, J. J., Schulte, A. C., Elliott, S. N., Nese, J. F. T., & Tindal, G. (2015). Growth and gaps in mathematics achievement of students with and without disabilities on a statewide achievement test. *Journal of School Psychology*, 53(1), 45–62. doi:10.1016/j.jsp.2014.11.001
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201–210. doi:10.1177/014662168300700208
- Swiss Psychological Society. (2003). *Ethische Richtlinien für Psychologinnen und Psychologen der Schweizerischen Gesellschaft für Psychologie* [Ethical guidelines for psychologists of the Swiss Psychological Society]. Bern, Switzerland: Swiss Psychological Society. Retrieved from Swiss Psychological Society website: https://www.swisspsychologicalsociety.ch/sites/default/files/ethische_richtlinien18.dt_.pdf
- Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, 8(1), 34–52. doi:10.1080/15434303.2010.536924
- Taherbhai, H., & Seo, D. (2013). The philosophical aspects of IRT equating: modeling erift to evaluate cohort growth in large-scale assessments. *Educational Measurement: Issues and Practice*, 32(1), 2–14. doi:10.1111/emip.12000
- Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, 9, 2245. doi:10.3389/fpsyg.2018.02245
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227–253. doi:10.1080/08957340701301207
- Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement*, 12(1), 53–67. doi:10.1177/014662168801200106
- van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, 22(3), 324–343. doi:10.1080/0969594X.2014.999024
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.
- von Davier, M., Weeks, J. P., Chen, H., Allen, J., & van der Velden, R. (2016). Creating simple and complex derived variables and validation of background questionnaire data. In *Technical report of the survey of adult skills (PIAAC)* (Ch. 20). Paris: OECD.
- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. doi:10.1007/BF02294627
- Wauters, K., Desmet, P., & Noortgate, W. van den. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, 26(6), 549–562. doi:10.1111/j.1365-2729.2010.00368.x
- Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wingersky, M. S., & Lord, F. M. (1983). *An investigation of methods for reducing sampling error in certain IRT procedures* (ETS Research Reports Series No. RR-83-28-ONR). Princeton, NJ: Educational Testing Service. Retrieved from Educational Testing Service website: <http://dx.doi.org/10.1177/014662168400800312>
- Yamamoto, K., Khorramdel, L., & von Davier, M. (2016). Scaling PIAAC cognitive data. In *Technical report of the survey of adult skills (PIAAC)* (Ch. 17). Paris: OECD.
- Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50(4), 399–410. doi:10.1007/BF02296259
- Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–485). Mahwah, NJ: Lawrence Erlbaum Associates.

Chapter 6. Epilogue

This thesis reflects on vertical scaling based on item response theory (IRT; e.g., de Ayala, 2009), within the practical context of implementing and validating a vertical Rasch scale for mathematics. IRT offers powerful techniques for implementing a vertical scale, since different item sets can be representative of the same unidimensional latent ability (e.g., Rost, 2004; Wainer & Mislevy, 2000). A vertical scale is particularly relevant for the data-based decision making approach to formative assessment (Schildkamp, Lai, & Earl, 2013; van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015) because it serves as a foundation for monitoring students' progress over time (Young, 2006). Thus, the scale allows both students and teachers to evaluate students' progress toward their learning goals; to adapt learning goals, learning strategies, or teaching methods, if necessary; and to define new learning goals that are appropriate for each student's learning path (Black & Wiliam, 1998; Hattie, 2009; Hattie & Timperley, 2007). This thesis contributes to the existing literature about IRT, vertical scaling, and formative assessment by proposing and evaluating practical solutions for practical challenges. In particular, this thesis elaborates on the challenges related to implementing and validating a vertical mathematics scale based on IRT methods to assess third through ninth grade students from Northwestern Switzerland using two different assessment instruments: a set of four standardized tests and an online item bank for formative assessment (Bildungsraum Nordwestschweiz, 2012; Tomasik, Berger, & Moser, 2018). This chapter reviews the research questions of the four studies presented in this thesis (see Chapter 1 for an overview), summarizes the related theoretical and empirical findings, and discusses their implications for future research.

6.1 Insights into Implementing a Vertical Scale for Mathematics in Northwestern Switzerland

Linking different assessments to a common vertical IRT scale can only be justified when the assessments share enough similarities, and when the measured ability or competency is continuously stimulated and increased over time (Young, 2006). To investigate whether the four standardized tests and the online item bank for formative assessment share enough similarities for a common vertical scale (research question Q2.1; see Chapter 1), Chapter 2 compared the two instruments using four assessment feature categories (Kolen, 2007; Kolen & Brennan, 2014). This comparison revealed substantial similarities between the two assessment instruments in three of the four categories; specifically, both instruments (1) are dedicated to the same population of third through ninth grade students in Northwestern Switzerland, (2) intend to provide objective reports reflecting the current ability levels of individual students, and (3) are based on the same content specification, namely the curriculum, *Lehrplan 21* (Deutschschweizer Erziehungsdirektoren-Konferenz, 2014). Differences were identified in the

instruments' measurement conditions; specifically, the compulsory standardized tests are administered following strict rules, while students and teachers can voluntarily and flexibly engage with the online item bank for formative assessment. Overall, it was concluded that a common vertical scale is justified. However, the final decision regarding calibration and linking procedures requires careful validation, not only from a psychometric perspective, through item analysis or the investigation of parameter invariance, but also from a content perspective, by means of external content-related validation criteria.

Taking into account the distinct characteristics of the two instruments, as well as several practical constraints, Chapter 2 addressed how a common vertical scale might be realized in practice (Q2.2) by proposing a concept for the scale's practical implementation. The most relevant constraint addressed in the concept was the ratio between the 10,000 items, which needed to be calibrated for the two instruments, and the limited sample of students who were willing to participate in calibration studies. A four-step calibration procedure was suggested: (1) developing a horizontal scale for the standardized third-grade test; (2) extending the third-grade scale to a vertical scale by means of dedicated calibration assessments; (3) linking the standardized tests for sixth, eighth, and ninth grade to the vertical scale; and (4) calibrating and linking additional items to the scale through online calibration. Only one of these four steps (step 2) referred to a classical calibration study in the sense that dedicated calibration assessments were proposed. The remaining three calibration steps combined item calibration with the administration of actual assessments, which provide student reports. The concept suggested assembling the first version of each standardized test without empirical pre-testing, based on the expertise of content experts (steps 1 and 3). This approach benefits from the large student sample and the homogenous administration conditions of the compulsory standardized tests. However, this approach also requires in-depth item analysis prior to releasing the reports to students and teachers in order to ensure reliability and fairness of the test results. The fourth calibration step was based on an algorithm for online calibration, with the goal of estimating item difficulty while students engage with the online item bank for formative assessment (Verschoor & Berger, 2015). To justify this approach, further studies should investigate the mechanisms for ensuring estimates of students' abilities are reliable during the early stages of the online item bank, when only a limited number of calibrated items are available. Furthermore, online calibration should be thoroughly monitored through offline calibration to verify the accuracy of the item difficulty parameters and to exclude potentially dysfunctional items.

6.2 Insights into Efficient Item Calibration Under Practical Constraints

To further address the limited resources available for item calibration, common-item nonequivalent group designs and common item designs were suggested for horizontal and vertical scaling, respectively (Kolen & Brennan, 2014), in steps 1, 2, and 3 of the calibration procedure described in Chapter 2. Given that, under the Rasch model, item calibration is most

efficient if an item's difficulty corresponds to the student's ability (Berger, 1991; Rost, 2004; Stocking, 1988; van der Linden, 1988), the question arose of whether the efficiency of targeted test forms within the common item designs could be further increased by means of performance-based routing (Q3.1). To this end, Chapter 3 extended the previous research into the efficiency of calibration designs (Berger, 1991; Stocking, 1988) by introducing targeted multistage calibration designs as a new design type, combining features of both traditional targeted calibration designs (Eggen & Verhelst, 2006; Mislevy & Wu, 1996) and multistage calibration designs (Eggen & Verhelst, 2006; Yan, von Davier, & Lewis, 2014; Zenisky, Hambleton, & Luecht, 2010; Zwitser & Maris, 2015). Targeted multistage calibration designs consider ability-related background variables and performance to optimize the match between item difficulty and student ability. The results of the simulation study presented in Chapter 3 indicated that targeted multistage calibration designs are more efficient than targeted designs for calibrating a given item pool, under the condition that the items are optimally located within each design. The reported efficiency gain was also relevant in practice. The student sample would have to be increased considerably to calibrate items through traditional targeted design in order to compensate for the efficiency gain of the targeted multistage calibration design. Furthermore, the multistage procedure improved the match between item difficulty and student ability, especially for easy and difficult items. Accurate estimates of item difficulty for these groups of items are highly relevant for the item bank for formative assessment and the related adaptive assessments, in which each student is assessed based on a different subset of items from the overall item pool (van der Linden & Glas, 2010).

The theoretically superior targeted multistage calibration designs, however, come at a certain price, if a priori knowledge about item difficulty is limited. In practice, empirical item difficulties are usually missing during the assembly stage of calibration assessments. Furthermore, content experts might under- or overestimate the difficulty of some items (e.g., Bejar, 1983; Hambleton & Jirka, 2006; Sydorenko, 2011; Wauters, Desmet, & van den Noortgate, 2012). In contrast to the previous research into efficient item calibration designs (e.g., Berger, 1991; Stocking, 1988), Chapter 3 addressed this practical constraint by investigating how limited a priori knowledge about item difficulty can affect the efficiency of both targeted calibration designs and targeted multistage calibration designs (Q3.2). The results of the simulation study showed that limited knowledge was related to a considerable decrease in efficiency for both the traditional targeted and targeted multistage calibration designs. Moreover, the results indicated that limited knowledge about item difficulties could result in inaccurate routing rules for targeted multistage calibration designs, which, in turn, may severely decrease the efficiency of a targeted multistage calibration design by unevenly distributing students over the modules. Targeted calibration designs, however, provide full control over the assignment of students to test booklets, independent of the expert's knowledge of the items' difficulties. Further studies are required to analyze the factors that contribute to the stability of targeted multistage calibration designs. In particular, future studies should investigate whether step-by-step adaptation of the routing rules or of the composition of modules during calibration

(cf. Ali & Chang, 2014; Kingsbury, 2009) might reduce the risk of efficiency loss due to imbalanced distribution of the number of observations per item. Furthermore, it would be interesting to extend the research into targeted multistage calibration designs to include more complex IRT models. In sum, targeted multistage calibration designs are a good option for enhancing calibration efficiency under the condition that accurate knowledge about item difficulty is available; however, targeted designs are still the safer choice for calibration studies whenever there are doubts about the accuracy of predicted item difficulty.

6.3 Insights into Relative Efficiency of Targeted and Multistage Testing

Chapter 4 continued the research into targeted multistage designs by addressing the fact that neither targeted testing based on ability-related background variables nor adaptive testing by means of multistage test (MST) designs can ensure that all students receive items that completely match their true abilities. Specifically, Chapter 4 extended the previous research into the efficiency of different test designs (e.g., Hendrickson, 2007; Kim & Plake, 1993; Yan, Lewis, & von Davier, 2014; Zenisky & Hambleton, 2014) by analyzing the efficiency of targeted multistage test (TMST) designs to estimate students' abilities under the Rasch model, and by providing insights into the relative efficiency of targeted and MST designs. The efficiency of targeted test designs had not been systematically studied in the past. The main research question in Chapter 4 was whether TMST designs achieve more efficient estimates of students' abilities than traditional targeted test designs or MST designs with one starting module (Q4.1). The results of the related simulation study showed that the performance-based module assignment used by the MST and TMST designs could substantially increase measurement efficiency, compared with targeted module assignment based on ability-related background variables. In addition, TMST designs were generally more efficient than MST designs.

Three specific factors that might influence the efficiency of TMST designs were explored in this chapter: (1) the correlation between the ability-related background variable and students' true abilities (Q4.2), (2) students' ability levels and categorization into ability groups (Q4.3), and (3) the length of the starting module in relation to the total test length (Q4.4). In the simulation study, the MST and TMST designs achieved comparable measurement efficiency when the target population spanned a narrow ability range, and the ability-related background variable was a poor indicator of students' true abilities. However, if the target population spanned a wide ability range and an accurate ability-related background variable was available, the TMST design was more efficient because it considered both low and high abilities from the first stage onward. The TMST design was particularly efficient for estimating the abilities of both low- and high-ability students, and might also protect these students from underload and overload caused by test items that are too easy or too difficult (Asseburg & Frey, 2013; Wise, 2014). Finally, the study showed that the length of the starting module is much more relevant for MST than for TMST designs.

It is important to note that the findings reported in Chapter 4 were limited to the Rasch model, a fixed test length, and a particular set of design conditions. Therefore, further studies should explore whether the use of more complex IRT models, different test lengths, more adaptive MST and TMST designs, or different variations in item difficulty within the modules (i.e., flat vs. peaked module information) could affect the efficiency of the TMST design, regarding estimation of students' abilities, compared to the efficiency of the MST and the traditional targeted designs (cf. Dallas, 2014; Verschoor & Eggen, 2014).

Chapter 4 demonstrated that MST designs might be a better choice to assess a population with a narrow ability range because these designs require fewer items in the first test stage than do TMST designs to achieve a comparable accuracy (i.e., one general instead of several targeted modules), making MST designs easier and cheaper to implement. TMST designs are a good option when the target population spans a wide ability range and an accurate ability-related background variable is available. However, the ability-related background variable must be perceived as a fair criterion by the test takers. Within the context of a vertical scale, school grade might be an especially well-accepted criterion for both low-stakes and high-stakes TMSTs. Within a school grade, exam grades or performance-related school types might be more difficult to justify in a high-stakes context and, thus, are rather recommended for low-stakes formative assessments. Consequently, TMST designs are particularly suitable for formative assessments.

6.4 Insights into the Vertical Scale's Validation from a Content Perspective

Chapter 5 directed attention toward content-related validation of the vertical mathematics scale, which was pointed out, in Chapter 2, as an essential step in the implementation of a vertical scale. The combined consideration of the scale's psychometric properties and content validity differentiates the study presented in Chapter 5 from previous studies of vertical scaling (e.g., Briggs & Weeks, 2009; Hanson & Béguin, 2002; Keller & Hambleton, 2013; Tong & Kolen, 2007). From a psychometric perspective, and based on empirical cross-sectional data gathered from 520 mathematics items and 2,733 third through ninth grade students, Chapters 5 investigated whether the items that were developed on the basis of the curriculum, *Lehrplan 21* (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016), fit a unidimensional vertical Rasch scale (Q5.1). The results of the item analysis indicated a satisfactory fit between the majority of the items and the final unidimensional vertical Rasch scale. Furthermore, item difficulty parameters from adjacent school grades, as well as item difficulty parameters from concurrent and grade-by-grade calibration procedures, were strongly correlated. These findings confirm the scale's unidimensionality and, thus, its stability across grades from a psychometric perspective (Hanson & Béguin, 2002; Kolen & Brennan, 2014).

From a content perspective, Chapter 5 examined whether the item difficulty estimates gathered from calibration matched the theoretical, content-related difficulties reflected in the

curriculum's underlying competence levels (Q5.1). To this end, the empirical item difficulty parameters on the final scale were compared with content-related item difficulties, which were based on the content experts' mapping of the items to the curriculum's competence levels. The strong positive correlation found between these two difficulties supported the content validity of the vertical mathematics scale. Chapter 5 further investigated whether the match between empirical and content-related item difficulties differed for items related to different curriculum cycles (i.e., primary vs. secondary school), domains, or competencies (Q5.2). The analysis demonstrated a stronger correlation between empirical and content-related item difficulty for primary school (i.e., third through sixth grade) than for secondary school (i.e., seventh through ninth grade), contradicting the psychometric evidence for the vertical scale's unidimensionality, and its objective of measuring the same mathematics ability throughout compulsory school. The vertical scale appears to be more representative of the competencies described for primary school than for secondary school. Further studies with a stronger focus on mathematics didactics and competence development are required to evaluate the potential cause of this result. In contrast, the comparable correlations within the mathematics domains and competencies indicated that the vertical scale satisfactorily represents the mathematics domains and competencies as stated in the curriculum, *Lehrplan 21* (Deutschschweizer Erziehungsdirektoren-Konferenz, 2016). However, further studies should replicate the results using a larger item pool and including larger samples of items for each competence level, thereby allowing for drawing conclusions about the possible differences between particular competencies. Future studies might also contribute to the validation of theoretical assumptions regarding competence development, as stated in the curriculum (Fleischer, Koeppen, Kenk, Klieme, & Leutner, 2013), especially for secondary school, as well as quality inspection of the items.

6.5 Conclusion and Outlook

Within the practical context of implementing a vertical mathematics scale for formative assessment in Northwestern Switzerland, this thesis introduced targeted multistage designs as a new design type with the aim of achieving as accurate estimates of item difficulty and student ability as possible, based on limited item and student samples. The research presented in this thesis shows that targeted multistage designs are promising for item calibration and testing under the condition that the practical context fulfills certain criteria; specifically, efficient targeted multistage calibration designs require accurate knowledge about the items' difficulties in order to assemble functional designs. Similarly, TMST designs rely on accurate and fair ability-related background variables. Given that the standardized tests and the online item bank for formative assessment are based on newly developed items, without empirical difficulties, the risks related to targeted multistage calibration designs are too high to justify their implementation in this particular practical context. Nevertheless, the practical constraint of a limited calibration sample, as described in Chapter 2, emphasizes the relevance of efficient calibration designs and, thus, underlines the need for further research in this area. Further

studies should evaluate whether more complex and adaptive targeted multistage calibration designs can overcome the limitations of the basic designs that were introduced in this thesis, and whether they may be a valid alternative to the fully adaptive online calibration algorithms (e.g., Ban, Hanson, Wang, Yi, & Harris, 2001; Fink, Born, Spoden, & Frey, 2018; Makransky & Glas, 2010; Stocking, 1988; Verschoor & Berger, 2015).

TMST designs are free of comparable risks and might be a good option for increasing measurement efficiency of the standardized eighth and ninth grade MSTs in Northwestern Switzerland. In these school grades, students are categorized into three performance-related school types, which might serve as a basis for assigning students to targeted starting modules at the beginning of the MSTs. Low- and high-ability students, in particular, could profit from such a design, as they would receive more appropriate items at the beginning of their tests. As a result, the tests would not only be more efficient, but would also prevent these students from experiencing underload or overload caused by test items that are too easy or too difficult. However, given the reported overlap of ability distributions among the three school types (e.g., Angelone, Keller, & Moser, 2013; Baumert, Stanat, & Watermann, 2006), whether such an approach is perceived as fair by the tested students, their teachers, and their parents should be carefully evaluated.

The validation study presented in Chapter 5 provided empirical evidence to justify implementing a common vertical scale to assess third through ninth grade students' mathematics abilities, according to the curriculum, *Lehrplan 21*. However, this study focused on the calibration assessments as one of four calibration steps, which are required in order to link the four standardized tests and the online item bank for formative assessment to a common vertical scale. Further studies are needed to validate the link between the two assessment instruments and ensure that both instruments measure the same mathematics ability even though they differ in their primary assessment purpose and measurement conditions. The differences between the two instruments might affect students' motivation during test-taking and, thus, might affect the outcome of item calibration (Mittelhaeuser, Béguin, & Sijtsma, 2011, 2013, 2015). Possible effects of the test mode should also be evaluated, as two of the four standardized tests are administered on paper. The transformation of the linking items between the paper-based and the computer-based modes might distort the scale if one mode facilitates or hinders students to correctly answer an item (e.g., Flaugher, 2000; Kröhne & Martens, 2011; Robitzsch et al., 2017; Wainer & Mislevy, 2000).

In conclusion, this thesis established a basis for implementing and validating a common vertical measurement scale for four standardized tests and an online item bank for formative assessment to evaluate seven school grades. This thesis underpinned the justification of an assessment system, which will enhance the validity of assessment results by facilitating interpretation of the results for students, teachers, and other stakeholders, and by offering a unique opportunity to monitor students' learning trajectories throughout compulsory school.

6.7 References

- Angelone, D., Keller, F., & Moser, U. (2013). *Entwicklung schulischer Leistungen während der obligatorischen Schulzeit: Bericht zur vierten Zürcher Lernstandserhebung zuhanden der Bildungsdirektion des Kantons Zürich* [Development of school performance during compulsory school: Report on the fourth assessment for the attention of the Zurich department of education]. Zürich, CH: Institut für Bildungsevaluation (IBE). Retrieved from Institut für Bildungsevaluation (IBE) website: http://www.ibe.uzh.ch/projekte/lezh/Lernstandserhebung_9KlasseZH_Bericht.pdf
- Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, 55(1), 92–104.
- Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38(3), 191–212. doi:10.1111/j.1745-3984.2001.tb01123.x
- Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and development of differential learning and development environments]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (1st ed., pp. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, 15(3), 293–306. doi:10.1177/014662169101500310
- Bildungsraum Nordwestschweiz (BR NWCH). (2012). *Checks und Aufgabensammlung im Bildungsraum Nordwestschweiz: Porträt. Bildungsraum Nordwestschweiz (BR NWCH)*. Retrieved from Bildungsraum Nordwestschweiz (BR NWCH) website: <http://www.bildungsraum-nw.ch/medien/dokumente-pdf>
- Black, P., & Wiliam, D. (1998). Assessment and classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. doi:10.1080/0969595980050102
- Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3–14. doi:10.1111/j.1745-3992.2009.00158.x
- Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework* (Dissertation). University of North Carolina, Greensboro, NC. Retrieved from https://libres.uncg.edu/ir/uncg/f/Dallas_uncg_0154D_11394.pdf
- de Ayala, R. J. (2009). *The theory and practice of item response theory. Methodology in the social sciences*. New York: Guilford Press.

- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2014). *Lehrplan 21: Rahmeninformationen*. Luzern. Retrieved from http://www.lehrplan.ch/sites/default/files/lp21_rahmeninformation_%202014-11-06.pdf
- Deutscheschweizer Erziehungsdirektoren-Konferenz (D-EDK). (2016). *Lehrplan 21: Mathematik*. Luzern. Retrieved from https://v-fe.lehrplan.ch/container/V_FE_DE_Fachbereich_MA.pdf
- Eggen, T. J. H. M., & Verhelst, N. D. (2006). Loss of information in estimating item parameters in incomplete designs. *Psychometrika*, 71(2), 303–322. doi:10.1007/s11336-004-1205-6
- Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60(3), 327–346.
- Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (2nd ed., pp. 37–60). Mahwah, NJ: Lawrence Erlbaum Associates.
- Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft*, 16(S1), 5–22. doi:10.1007/s11618-013-0379-z
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3–24. doi:10.1177/0146621602026001001
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.
- Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26(2), 44–52. doi:10.1111/j.1745-3992.2007.00093.x
- Keller, L. A., & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement*, 50(4), 390–407. doi:10.1111/jedm.12025
- Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.
- Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

- Kröhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, 14(2), 169–186. doi:10.1007/s11618-011-0185-4
- Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, 11(1), 1–20.
- Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Reports Series No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
- Mittelhaeuser, M.-A., Béguin, A. A., & Sijtsma, K. (2011). *Comparing the effectiveness of different linking designs: The internal anchor versus the external anchor and pre-test data* (Report No. 11-01). Retrieved from http://www.cito.nl/~media/cito_nl/Files/Onderzoek%20en%20wetenschap/cito_mrd_report_2011_01.ashx
- Mittelhaeuser, M.-A., Béguin, A. A., & Sijtsma, K. (2013). Modeling differences in test-taking motivation: Exploring the usefulness of the mixture Rasch model and person-fit statistics. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *Springer proceedings in mathematics & statistics: Vol. 66. New developments in quantitative psychology. Presentations from the 77th annual Psychometric Society meeting* (pp. 357–370). New York: Springer.
- Mittelhaeuser, M.-A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, 52(3), 339–358. doi:10.1111/jedm.12080
- Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien [Challenges in Estimations of Trends in Large-Scale Assessments: A Calibration of the German PISA Data]. *Diagnostica*, 63(2), 148–165. doi:10.1026/0012-1924/a000177
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* [Textbook test theory, test construction] (2nd ed.). Psychologie Lehrbuch. Bern, Switzerland: Verlag Hans Huber.
- Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education: Challenges and opportunities*. Dordrecht, the Netherlands: Springer.
- Stocking, M. L. (1988). Scale drift in on-line calibration. *ETS Research Report Series*, 1988(1), i-122. doi:10.1002/j.2330-8516.1988.tb00284.x
- Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, 9, 2245. doi:10.3389/fpsyg.2018.02245
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227–253. doi:10.1080/08957340701301207
- van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative

- assessment. *Assessment in Education: Principles, Policy & Practice*, 22(3), 324–343. doi:10.1080/0969594X.2014.999024
- van der Linden, W. J. (1988). *Optimizing incomplete sample designs for item response model parameters* (Research Report No. 88-5). Enschede, the Netherlands: University of Twente.
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.
- Verschoor, A. J., & Berger, S. (2015, September). *Computerized adaptive testing with online JML calibration*. Paper presented at the IACAT Conference, Cambridge, England.
- Verschoor, A. J., & Eggen, T. J. H. M. (2014). Optimizing the test assembly and routing for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 135–150). Boca Raton, FL: CRC Press.
- Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing. A primer* (2nd ed., pp. 61–100). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, 2(1), 1–17. doi:10.7333/1401-0201001
- Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 3–20). Boca Raton, FL: CRC Press.
- Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.
- Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–485). Mahwah, NJ: Lawrence Erlbaum Associates.
- Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer.
- Zenisky, A. L., & Hambleton, R. K. (2014). Multistage test designs: Moving research results into practice. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 21–37). Boca Raton, FL: CRC Press.
- Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, 80(1), 65–84. doi:10.1007/s11336-013-9369-6

Summary

A vertical measurement scale is the basis for repeatedly assessing and monitoring students' abilities throughout their school years. Advanced computer technology serves as a foundation for implementing vertical scales and related complex measurement models, such as item response theory (IRT) models, within computer-based assessment systems. Such innovative assessment systems have immense potential for formative assessment because they can assist students and teachers to evaluate students' progress toward their learning goals; to adapt learning goals, learning strategies, or teaching; and to define new learning goals that are appropriate for each student's learning path. This thesis was motivated by practical challenges related to implementing and validating a vertical Rasch scale to measure students' mathematics abilities throughout compulsory school in Northwestern Switzerland. The goal of this vertical scale is to provide third through ninth grade students with objective, reliable, and valid assessment reports based on two different assessment instruments: (1) a set of four standardized tests and (2) an online item bank for formative assessment. This thesis contributes to the existing literature about IRT, vertical scaling, and formative assessment by evaluating practical solutions for practical challenges, which can complicate the implementation and validation of a vertical Rasch scale. This thesis elaborated on practical constraints, including time and financial resources; willingness of schools, teachers, and students to participate in calibration studies; and the number of new items that need to be calibrated as a basis for repeatedly assessing students over multiple school years.

Chapter 1, the general introduction, provided an overview of educational assessment in Northwestern Switzerland as the practical context for this thesis, and introduced vertical scaling and efficient testing based on IRT methods as the common theoretical themes of the studies presented in this thesis. Chapter 1 also outlined the research objectives and research questions.

In **Chapter 2**, a theoretical concept was proposed to implement IRT-related calibration procedures and data-collection designs in order to establish the vertical mathematics scale for Northwestern Switzerland. Specifically, Chapter 2 described the four standardized tests and the online item bank for formative assessment in more detail, evaluating their similarities and differences regarding target population, assessment types and purposes, content specifications, and measurement conditions. Chapter 2 also provided an overview of different IRT-related calibration procedures and data-collection designs for both horizontal and vertical scaling. This chapter elaborated on the idea of targeted and adaptive testing based on the Rasch model to increase measurement efficiency under the practical constraints of a limited student or item samples. By integrating the two instruments' similarities and differences with the theoretical background on data-collection designs and item calibration within a Rasch framework, a four-step item calibration process was proposed to establish a vertical scale and link the two instruments. The concluding discussion pointed out a need for empirical research into efficient

calibration and test designs under practical contexts and constraints, and stressed the need to validate the final scale from a psychometric, as well as a content, perspective.

Chapter 3 directed focus toward the fact that calibration of an item bank for computerized adaptive testing, such as the online item bank for formative assessment, requires substantial resources, and addressed the need for empirical research into efficient calibration designs. This chapter presented a study that investigated whether calibration efficiency under the Rasch model could be enhanced through targeted multistage calibration designs, which consider ability-related background variables and performance for assigning students with suitable items. This chapter also investigated whether uncertainty about item difficulty could impair assembly of an efficient design. The results indicated that targeted multistage calibration designs were more efficient than ordinary targeted designs under optimal conditions. Limited knowledge about item difficulty reduced the efficiency of one of the two investigated targeted multistage calibration designs, whereas the targeted design was more robust.

Chapter 4 further investigated the idea of combining targeted and multistage testing and addressed the fact that neither targeted testing by means of ability-related background variables nor adaptive testing through multistage tests (MSTs) can ensure that all students receive items that completely match their abilities. Targeted designs do not consider that student abilities might significantly differ within each group. MST designs usually include a starting module of moderate difficulty, which does not account for differences in student abilities. This chapter investigated whether measurement efficiency can be improved through targeted multistage test (TMST) designs that consider ability-related background information for a targeted assignment at the beginning of the test, as well as performance during test-taking, for selecting matching test modules. Through simulations, the efficiency of the TMST design for estimating student ability was compared to that of the traditional targeted test design and the MST design. Chapter 4 further analyzed the extent to which each design's efficiency depends on the correlation between the ability-related background variable and students' true abilities, each student's ability level and categorization into an ability group, and the length of the starting module compared to the total test length. The results indicated that TMST designs were generally more efficient than targeted and MST designs, especially if the ability-related background variable was highly correlated with students' true abilities. TMST designs were also particularly efficient for estimating the abilities of low- and high-ability students within a given population. Finally, longer starting modules resulted in a less efficient estimation of low and high abilities than did shorter starting modules. However, this finding was more prominent for MST than for TMST designs. This chapter concluded by recommending TMST designs to assess students with a wide range of abilities when a reliable ability-related background variable is available.

Chapter 5 resumed and expanded upon the conclusion provided in Chapter 2, that vertical scales require validation from a psychometric, as well as a content, perspective. Specifically, this chapter described the actual implementation of the calibration assessments,

proposed in Chapter 2, as one of four calibration steps to establish a vertical Rasch scale for assessing the mathematics abilities of students in third through ninth grade. The psychometric properties of the vertical scale were examined through item analysis, as well as by comparing two different calibration procedures: concurrent and grade-by-grade calibration. The content-related validity of the scale was evaluated by contrasting the empirical item difficulty estimates with the theoretical, content-related item difficulties reflected in the underlying competence levels of the curriculum. Through correlation and multiple regression analyses, this chapter explored whether the match between empirical and content-related item difficulty differed for items related to different curriculum cycles (i.e., primary vs. secondary school), domains, or competencies within mathematics. The results indicated a satisfying item fit and a close match between the outcomes of the concurrent and grade-by-grade calibration procedures, supporting the scale's unidimensionality and stability from a psychometric perspective. In addition, strong correlations between the empirical and content-related item difficulties were found, emphasizing the scale's content validity. Further analysis showed a higher correlation between empirical and content-related item difficulty at the primary level when compared with the secondary school level. Correlations were comparable across the different curriculum domains and competencies, implying that the scale is a good indicator of students' math abilities, as outlined in the curriculum.

The Epilogue, **Chapter 6**, reviewed the primary research questions answered by the four studies presented in this thesis, summarized the related theoretical and empirical findings, and discussed their implications for future research.

Samenvatting

Aan herhaaldelijk meten van vaardigheden en het volgen van leerlingen tijdens hun jaren op school ligt een verticale vaardigheidsschaal ten grondslag. Geavanceerde computertechnologie vormt een basis voor het implementeren van verticale schalen en daarvoor benodigde complexe meetmodellen zoals itemresponstheorie (IRT) modellen in computergestuurde toetssystemen. Zulke innovatieve systemen kunnen een grote bijdrage leveren aan formatieve toetsing aangezien ze leerlingen en docenten kunnen ondersteunen in het evalueren van de voortgang, in het bereiken en in het aanpassen van leerdoelen van leerlingen, leerstrategieën of lesgeven, en in het definiëren van nieuwe leerdoelen die geschikt zijn voor de individuele leerwegen van leerlingen. De praktische uitdagingen bij het implementeren en valideren van een verticale Raschschaal voor het meten van wiskundevaardigheden voor leerlingen uit noordwest Zwitserland gedurende hun leerplichtige periode zijn de basis voor deze dissertatie. Het doel is het objectief, betrouwbaar en valide rapporteren van vaardigheden aan de hand van twee meetinstrumenten: (1) een serie van vier gestandaardiseerde toetsen en (2) een online itembank voor formatieve toetsing. Dit proefschrift levert een bijdrage aan literatuur over IRT, verticale schalen en formatieve toetsing door het vinden van praktische oplossingen voor praktische vragen bij het ontwikkelen van deze verticale schalen. Specifiek wordt er uitgeweid over de context van beperkte beschikbare middelen, van lage bereidheid van docenten en leerlingen om mee te werken aan grootschalige proefonderzoeken, en van de noodzaak om grote hoeveelheden nieuwe items te ontwikkelen en te kalibreren.

Hoofdstuk 1 geeft een overzicht over de context van de toetspraktijk in noordwest Zwitserland en introduceert verticale schalen als een efficiënte methode, gebaseerd op IRT-modellen, als het centrale thema in dit proefschrift. Daarnaast beschrijft het de onderzoeksdoelen en -vragen voor de rest van de studie.

In **hoofdstuk 2** wordt een theoretische basis geïntroduceerd voor dataverzameling en kalibratieprocedures voor het ontwikkelen van de verticale schaal voor wiskunde. Hoofdstuk 2 beschrijft de vier gestandaardiseerde toetsen en de online itembank in detail door in te gaan op overeenkomsten en verschillen tussen de twee instrumenten: doelpopulatie, doel en vorm van de toetsen, toetsspecificaties en -omstandigheden. Verder komen procedures en designs voor het linken van horizontale en verticale IRT schalen aan bod. Dit hoofdstuk werkt ook het idee uit van targeted en adaptieve toetsing om de meetefficiëntie te verhogen onder praktische randvoorwaarden van een beperkte inzetbaarheid van leerlingen en items. Door de overeenkomsten en verschillen tussen de beide instrumenten te combineren met het theoretische kader van designs en kalibratie, wordt een procedure voorgesteld die voorziet om in vier kalibratiestappen een schaal te ontwikkelen die beide instrumenten met elkaar verbindt. De noodzaak voor onderzoek naar efficiënte toetsdesigns wordt duidelijk, evenals de noodzaak tot validering van de schaal vanuit psychometrisch en inhoudelijk perspectief.

In **hoofdstuk 3** ligt de focus op het feit dat kalibreren van een itembank voor adaptieve toetsing een groot beroep doet op resources, en daarom werd er aandacht besteed aan empirisch onderzoek naar efficiënte kalibratiedesigns. Onderzocht werd of de efficiëntie van kalibratie onder het Raschmodel verhoogd kon worden door middel van targeted multistage kalibratiedesigns. Deze designs gebruiken zowel achtergrondvariabelen als de vaardigheid om leerlingen aan geschikte items toe te wijzen. Verder werd onderzocht wat de impact is van onzekerheid over de moeilijkheid van de items op de efficiëntie. Het resultaat was dat targeted multistage kalibratiedesigns efficiënter zijn dan gewone targeted kalibratiedesigns onder optimale omstandigheden. Inschattingsfouten van itemmoeilijkheden reduceerden de efficiëntie van targeted multistage designs terwijl de gewone targeted designs robuuster bleken te zijn.

De combinatie van targeted testing en multistage testing wordt verder uitgewerkt in **hoofdstuk 4**. Hier wordt ingegaan op de situatie dat zowel targeting aan de hand van achtergrondvariabelen als multistage tests (MSTs) niet kunnen garanderen dat items aan leerlingen worden aangeboden die geheel bij hun vaardigheid aansluit. Targeted designs negeren het feit dat er grote spreiding in de subgroepen, waaraan toegewezen wordt in de designs, blijft. Daarentegen gebruiken MSTs doorgaans één startmodule van gemiddelde moeilijkheid, waarbij geen rekening wordt gehouden met verschillen in vaardigheid. In dit kader werd onderzocht of targeted multistage tests (TMST), die achtergrondvariabelen gebruiken voor de toewijzing van leerlingen aan een geschikte startmodule en waarbij aan de hand van de vaardigheid de overige modules worden gekozen, de efficiëntie kunnen verhogen. Door middel van simulaties is de efficiëntie van TMSTs vergeleken met traditionele targeted tests en MSTs. In hoofdstuk 4 is verder onderzocht in hoeverre de efficiëntie van TMSTs afhangt van de correlatie tussen de achtergrondvariabele en de vaardigheid, van de toewijzing van leerlingen in groepen en van de lengte van de startmodules relatief ten opzichte van de totale toetslengte. In het algemeen bleek dat TMSTs efficiënter waren dan MSTs en targeted tests, en des te beter waren indien de achtergrondvariabele hoger correleerde met de vaardigheid. In het bijzonder bleek dat TMSTs efficiënter bleken om laag- en hoogvaardige leerlingen te schatten. Tenslotte bleek dat lange startmodules minder efficiënt waren voor deze extremere leerlingen dan korte startmodules. Dit effect was echter groter voor MSTs dan voor TMSTs. Het hoofdstuk sluit af met de aanbeveling TMSTs te gebruiken indien de vaardigheden in de groep een grote spreiding vertonen en er een betrouwbare, met de vaardigheid correlerende, achtergrondvariabele beschikbaar is.

Hoofdstuk 5 gaat verder met de conclusie van hoofdstuk 2 dat verticale schalen, zowel vanuit psychometrisch als vanuit inhoudelijk perspectief, gevalideerd dienen te worden. Het hoofdstuk beschrijft de kalibratietoetsen die in hoofdstuk 2 werden voorgesteld als een van de stappen om een verticale schaal te ontwikkelen voor wiskundeleerlingen van leerjaar 3 tot en met 9 in noordwest Zwitserland. De psychometrische eigenschappen van de verticale schaal werden onderzocht door itemanalyses, en door het vergelijken van twee kalibratieprocedures: gelijktijdige en groep-voor-groep kalibratie. Verder werd de schaal inhoudelijk gevalideerd

door de empirisch vastgestelde moeilijkheden te vergelijken met de ingeschatte moeilijkheden aan de hand van de competentieniveaus van het curriculum. Door middel van multiple regressie werd onderzocht of de niveaus overeenkwamen met de empirische moeilijkheden. De resultaten lieten zien dat de itemfit redelijk goed was en dat er een duidelijke overeenkomst was tussen de verschillende kalibratieprocedures, zodat bewijs werd geleverd voor een unidimensionele, stabiele schaal. Verder werd er een hoge correlatie gevonden tussen de empirische moeilijkheden en de moeilijkheden gebaseerd op de inhoud van de curriculumonderdelen, waarmee inhoudelijke validiteit werd aangetoond. Nadere analyses lieten een hogere correlatie zien in het primair onderwijs dan in het secundair onderwijs. In de meeste curriculumonderdelen en competentieniveaus waren de correlaties met elkaar vergelijkbaar, waarmee de schaal een goede indicator vormt voor de wiskundevaardigheden zoals in het curriculum is beschreven.

Tot slot komt **hoofdstuk 6**, de epiloog, terug op de onderzoeksvragen die in de voorgaande hoofdstukken zijn gesteld, verbindt het de theoretische en empirische resultaten en behandelt hun implicatie voor vervolgonderzoek.

Acknowledgments

More than four years ago, I started my PhD project, inspired by an assignment of the cantons of Northwestern Switzerland to develop an innovative online item bank for formative assessment. As a senior researcher at the Institute for Educational Evaluation (IBE) at the University of Zurich, I was involved in the conception and practical implementation of this assessment system. My PhD project gave me the opportunity to delve deeper into related theoretical aspects such as item response theory, computer adaptive testing and vertical scaling. I'm very thankful for what I've learned from both my practical work and my PhD project, and I would like to thank all the people who accompanied me during these exciting and educational years.

In particular, I thank my supervisors Theo Eggen, Urs Moser and Angela Verschoor. I thank Theo for giving me the opportunity to address practical psychometric challenges in my thesis and always providing me valuable and targeted feedback, which helped me refine and concretize my research questions and analyses. Often my ideas became clearer to me after discussing them with him. I'm also thankful for his support in aligning my PhD project with the sometimes challenging timeline of the subordinate project. To my boss and supervisor Urs, I'm grateful for the opportunities to contribute to the exciting endeavor of developing an assessment system from scratch, and to learn from international experts in psychometrics and educational assessment. He encouraged me to start and finish my PhD project and helped me balance my daily work at the IBE with my PhD project, and lately also with my responsibilities as a mother. To Angela, I'm thankful for introducing me to the world of simulations and algorithms and for her support during numerous conferences. I've learned a lot from her, especially during my stay in Arnhem. I also very much appreciated her hospitality whenever I was in the Netherlands.

Furthermore, I thank the whole IBE team for the excellent cooperation over the last years. It is very inspiring to work in such a great team and I very much appreciate the joyful and loyal atmosphere at the office. Thank you for taking over my daily work during my more intense PhD periods, for motivating me to continue with my PhD project and for your supporting advice. I also would like to thank the international and psychometrics departments at Cito for sharing their remarkable knowledge and for their hospitality during my visits to Arnhem. Especially, I am grateful to Henk Moelands, who encouraged me to write my PhD thesis at the University of Twente and introduced me to Theo. In addition, I thank Maaïke van Groen and Frans Kleintjes, who were—together with Angela and Henk—closely involved in the development of the assessment system. Thank you for the inspiring collaboration and sharing your valuable advice.

Finally, I thank my friends and family for accompanying me during my PhD project, for their continued interest in my progress, their open ears during more difficult periods and their

understanding for my sometimes busy schedule. I thank my parents, Monique and Emil, for providing me the support, confidence and freedom to develop into a critical, open-minded and applied researcher, and my brother Manuel for his encouragement and flexible support with numerous short-term proof-reading sessions. I give a special thanks to my husband Remo, who always believed in my abilities to finalize this project, and to my daughter Laila. Thank you very much for fulfilling my life with memorable, lighthearted moments. I'm very grateful for all your encouragement, patience, love and support.

Research Valorization

Publications in Scientific Journals

Published

- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Efficiency of targeted multistage calibration designs under practical constraints: A simulation study. *Journal of Educational Measurement*, 56 (1), 121–146. doi: 10.1111/jedm.12203
- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education*, 4, 1. doi: 10.3389/educ.2019.00001
- Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, 9, 2245. doi:10.3389/fpsyg.2018.02245
- Berger, S., & Freund, A. M. (2012). Fear of failure, disorganization, and subjective well-being in the context of preparing for an exam. *Swiss Journal of Psychology*, 71 (2), 83–91. doi:10.1024/1421-0185/a000074

Submitted

- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (submitted). *Development and validation of a vertical scale for formative assessment in mathematics*. Manuscript submitted for publication.

Conference Contributions

- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2018, November). *Improvement of measurement efficiency in multistage tests by targeted assignment*. Paper presented at the AEA-Europe Conference, Arnhem–Nijmegen, The Netherlands.
- Berger, S., Verschoor, A., Eggen, T., & Moser, U. (2018, February). *Konzept und Kalibrierung eines adaptiven formativen Feedbacksystems für den Unterricht*. Vortrag auf der 6. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Basel, Schweiz.
- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2017, August). *Efficiency of targeted multistage calibration designs under practical constraints*. Paper presented at the IACAT Conference, Niigata, Japan.
- Tomasik, M. J., & Berger, S. (2017, July). *After the PISA debacle in Switzerland: Lessons learned and measures taken*. Invited talk at the Annual Meeting of the Psychometric Society, Zurich, Switzerland.

- Berger, S., Moser, U., Verschoor, A. J., & Eggen, T. J. H. M. (2016, November). *Computer-based formative assessment in classrooms*. Paper presented at the AEA-Europe Conference. AEA-Europe, Limassol, Cyprus.
- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2016, October). *Efficiency of targeted multistage calibration designs under practical constraints*. Paper presented at the RCEC Workshop on IRT and Educational Measurement, Enschede, The Netherlands.
- Berger, S., Moser, U., Verschoor, A. J., & Eggen, T. J. H. M. (2015, November). *Development of an online item bank for adaptive formative assessment*. Paper presented at the AEA-Europe Conference. AEA-Europe, Glasgow, Scotland.
- Berger, S., Oostlander, J., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2015, September). *Development of multistage tests based on teacher ratings*. Paper presented at the conference of the International Association for Computerized Adaptive Testing, Cambridge, England.
- Verschoor, A. J., & Berger, S. (2015, September). *Computerized adaptive testing with online JML calibration*. Paper presented at the IACAT Conference, Cambridge, England.
- Berger, S., Verschoor, A. J., Moser, U., & Oostlander, J. (2014, November). *Educational assessment in Northwestern Switzerland: Psychometrical challenges in linking standardized tests to an online item bank for formative assessment*. Paper presented at the RCEC Workshop on IRT and Educational Measurement, Enschede, The Netherlands.
- Berger, S., Verschoor, A. J., Moser, U., & Oostlander, J. (2014, November). *Educational assessment in Northwestern Switzerland: Psychometrical challenges in linking standardized tests to an online item bank for formative assessment*. Paper presented at the AEA-Europe Conference, Tallinn, Estonia.