DISSERTATION

Computerized Adaptive Testing in Inclusive Education

Nikola Ebenbeck

2023

Computerized Adaptive Testing in Inclusive Education

Inaugural-Dissertation

zur Erlangung der Doktorwürde der Fakultät für Humanwissenschaften der Universität Regensburg

vorgelegt von

Nikola Edith Ebenbeck

aus Straubing

2023

Regensburg, 2023

Erstgutachter (Betreuer): Prof. Dr. Markus Gebhardt
Lehrstuhl für Lernbehindertenpädagogik
Fakultät für Humanwissenschaften
Universität Regensburg

Zweitgutachter: Prof. Dr. Peter Zentel
Lehrstuhl für Geistigbehindertenpädagogik
Fakultät für Psychologie und Pädagogik
Ludwig-Maximilian-Universität München

To Markus, Peter, Jakob, Jana, Katharina, Miriam, Stephanie, Dominik, Lena, Katha, Woife, Beate, Melanie, Sascha, Anna, Judith, Ralph, Morten, Sven, and Myself.

Thank you for helping me grow.

Abstract

Assessing students with special educational needs (SEN) and slow learners is crucial in order to derive fitting support and instructions but can be challenging for both test administrators and examinees. Computerized adaptive testing (CAT) is an assessment technology with the chance to meet those challenges. CATs are particularly promising for students with SEN, as they have the potential to more individualized and shorter measurements and better testing in low extreme areas compared to traditional procedures. CATs are digitally conducted tests whose difficulty level adapts to the examinees. This is done with the help of underlying adaptive algorithms that re-estimate the examinee's ability after each answer and suggest the most suitable next item based on that. This technology allows tests to be shortened with minimal impact on measurement accuracy.

To investigate the benefits of CAT for student groups with heterogeneous abilities and particularly for students with SEN, three simulation studies are conducted in this work. All studies are based on a sample of 400 students with (22.5%) and without (78.5%) SEN. On the part of all students, 10.75% have intellectual disabilities, 7.75% have learning disabilities, and 3% have speech impairments. The students have completed four subtests of a digital reading screening which contain between 30 and 52 items. The results of the reading screening are used in the three studies to answer the respective research questions.

Study 1 simulates a computerized adaptive reading screening for inclusive schools. Therefore, simulations of CATs based on the subtests and generated data are done. Three different accuracy stopping rules, the test length and test accuracy are compared and further analyzed. Study 2 links the subtests to one screening CAT by incorporating the results of the previous subtest as input for the following subtest. For this purpose, a fixed and a Bayesian-based starting rule are compared based on real and generated data. Study 3 investigates the performance of a CAT for students with and without SEN in detail. For this purpose, different performing student groups are simulated separately under different conditions, and the performance of the CAT measurements as well as the students' response patterns are compared and analyzed.

The results indicate that students with SEN are measured more accurately and with a shorter test length, approximately 4 items fewer than students without SEN. This effect, demonstrating the greater effectiveness of adaptive testing for students with SEN compared to those without SEN, remains consistent across item pools of different sizes, difficulty distributions, and various starting rules. In the adaptive test, students with SEN make fewer overall incorrect

responses, although there is a slight decrease in the proportion of correctly solved items and a slight increase in the proportion of incorrectly solved items.

In general, adaptive tests are 30% to 80% shorter than the initial non-adaptive tests. The degree of test reduction depends on the size of the initial item pool. Smaller item pools, with around 30 items, result in a relatively smaller reduction in test length compared to larger item pools with 50 or 100 items. However, when the accuracy of the adaptive test is used as the stopping criterion, a shorter test length also leads to a decrease in test accuracy. Employing a higher standard error in the measurement accuracy of the adaptive test yields shorter tests with fewer items per test iteration, but it also results in slightly lower measurement accuracy. For small item pools, a standard error of 0.3 has minimal impact and rarely shortens the test. With a standard error of 0.5, depending on the size of the item pool, most test iterations can be concluded with a significantly reduced test length due to the achieved accuracy.

A larger and uniform distributed item pool (n = 100) leads to an additional average reduction of 6 items per subtest. Additionally, this item pool allows for a 30% increase in the number of test runs that can be stopped based on their accuracy. For students in extreme areas, even when individuals' abilities deviate several standard deviations from the mean, short measurements can be ensured instead of having to select and process all items from a smaller and more tightly distributed item pool until the test stops with slightly lower measurement accuracy.

The usage of a Bayesian-based starting rule, which incorporates previous test results and student information as input for the subsequent measurement, as well as the use of an easier start item, is not efficient and does not enhance the measurement in comparison to a starting rule using a fixed item with medium difficulty.

To develop a digital and adaptive reading screening, the item pools of the non-adaptive reading subtests can be utilized for further adaptation in terms of their psychometric quality, item pool size, and difficulty distribution. As a starting point, I opted for an initial item with a difficulty level of -1 to facilitate an easier test initiation for students with SEN. Subsequently, an estimation and item selection process based on Bayes estimation and maximum Fisher information is employed. The adaptive screening concludes once a measurement accuracy is attained, with a standard error of 0.5.

This work demonstrates that CATs particularly benefit students with SEN and, therefore, are useful for this target group. The opportunities and limitations of developing and using CATs for inclusive school tests are discussed. Furthermore, implications are drawn for the further development of CATs and their potential combination with artificial intelligence and digital

learning environments. Additionally, the implementation of the adaptive reading screening is considered.

Table of Contents

List of Abbreviations vi
List of Figures. viii
List of Tables… xii

1 Introduction 1
2 Literature Review 5
2.1 Struggling Students and their (Dis-) Abilities 5
2.1.1 Labeling and Identification 5
2.1.2 Types of Special Educational Needs 7
2.2 Educational Assessment in Inclusive Education 12
2.2.1 Data-based Decision Making 12
2.2.2 Psychometric Requirements for Inclusive Screening Measures 16
2.3 Computerized Adaptive Testing 18
2.3.1 Chances of Computers in Educational Assessment 18
2.3.2 Components of Computerized Adaptive Testing 21
3 Empirical Research Process 34
3.1 Instruments 34
3.2 Sample Description 37
3.3 Research Questions and Procedures 38
3.3.1 Study 1: Simulating Screening SubCATs of Inclusive Student Groups 38
3.3.2 Study 2: Simulating a Screening CAT of Inclusive Student Groups 43
3.3.3 Study 3: Comparing Students' Simulated CAT Performance 45
4 Results.. 49
4.1 Study 1: Simulation of a Computerized Adaptive Reading Screening for Inclusive School Use 49
4.1.1 Results 49
4.1.2 Discussion and Limitations 58
4.2 Study 2: Comparing Test Length and Measurement Accuracy pf Standalone and Sequentially Linked Adaptive Tests for Inclusive School Use – A Simulation Study 61
4.2.1 Results 61
4.2.2 Discussion and Limitations 65
4.3 Study 3: A Simulation-Based Comparison of the Effectiveness of Adaptive Tests for Students With and Without Special Educational Needs 67

4.3.1 Results 67
4.3.2 Discussion and Limitations 71
5 Discussion 74
5.1 Development of an Adaptive Screening 74
5.1.1 Item Pool 74
5.1.2 Starting the CAT 78
5.1.3 Stopping the CAT 79
5.2 Chances of Adaption for Students with SEN 82
5.2.1 Computerized Adaptive Testing 82
5.2.2 Machine Learning Based Adaptive Testing 83
5.2.3 Need for Digitalization 84
5.3 Limitations 85
5.4 Future Research Possibilities 87
5.4.1 Implementation of the Adaptive Reading Screening 88
5.4.2 Adaptive Testing in the Classroom 88
6 Conclusion 90
7 References I

Appendices XIX
A. Example Items of the Subtests XIX
B. Simulation Process Charts XXI
C. Psychometric Analysis XXVI
D. SubCAT Simulations XXVII
E: SubCAT Simulations of SEN Students XXXIX

List of Abbreviations

AB	Adaptive behavior
AI	Artificial intelligence
ANOVA	One-way Analysis of Variance
BI	Borderline intelligence
CAA	Computer assisted assessment
CAT	Computerized adaptive test / -ing
CBT	Computer-based test / -ing
CML	Conditional maximum likelihood estimation
DBDM	Data-based decision making
EAP	Expected a posteriori estimator
IDEA	Individuals with Disabilities Education Act
IEP	Individualized education plans
IRT	Item-Response-Theory
IQ	Intelligence quotient
JML	Joint maximum likelihood estimation
MAP	Maximum a posteriori estimator
MFI	Maximum Fisher information
ML	Maximum likelihood
MML	Marginal maximum likelihood estimation
MSQ	Mean Squared Residual based
PBT	Paper-based test / -ing
SD	Standard deviation
SE	Standard error
SEN	Special educational needs
SEN-I	Special educational needs in intellectual development
SEN-L	Special educational needs in learning
SEN-S	Special educational needs in speech, language and communication
SLD	Specific learning disabilities

USA	United States of America
θ	Person ability
σ	Item difficulty

List of Figures

Figure 1: Distribution of the number of attempted items, error score and sum score for subtest 1, 2, 3, and 4. 50
Figure 2: Person-item-maps of subtests 1, 2, 3, and 4. 52
Figure 3: Estimated and generated θ for all subtests. 53
Figure 4: σ of selected items and estimated θ while a simulated test run of a person with true θ = -1. 55
Figure 5: σ of selected items and estimated θ while a simulated test run of a person with true θ = -3. 55
Figure 6: σ of selected items and estimated θ while a simulated test run of a person with true θ = 3. 56
Figure 7: σ of generated uniform item pool. 57
Figure 8: σ of selected items and estimated θ while a simulated test run of a person with true θ = -3. 57
Figure 9: σ of selected items and estimated θ while a simulated test run with fixed first item of a person with true θ = -1.82 for subtest 1, true θ = -0.27 for subtest 2, true θ = -0.03 for subtest 3 and true θ = -2.50 for subtest 4. 63
Figure 10: σ of selected items and estimated θ while a simulated test run with θ input of the previous estimated θ of a person with true θ = -1.82 for subtest 1, true θ = -0.27 for subtest 2, true θ = -0.03 for subtest 3 and true θ = -2.50 for subtest 4. 63
Figure 11: σ of selected items and estimated θ while a simulated test run with θ input of the previous estimated θ of a person with true θ = -1 for all subtests. 64
Figure 12: Generated θ of students with (right) and without (left) SEN. 69
Figure A1: Example item of subtest 1: The title says “Where can you hear the letter?”. The graphic in the left box represents the German word for ‘book’ (“Buch”). In the right box, the letter, that has to be identified within the word, is represented (“B”). Buttons are labelled as beginning (“Anfang”), middle (“Mitte”), end (“Ende”), and nowhere (“Nirgends”). XVIII
Figure A2: Example item of subtest 2: The title says “Does the word exist?”. The word in this example is the real existing German word for ‘different’ ("anders"). Buttons are labelled as ‘Does exist’ (“Gibt es”) and ‘Does not exist’ (“Gibt es nicht”). XVIII

Figure A3: Example item of subtest 3: The title says “Which word did you see?”. The flashed word was “April”. Buttons are labelled with different words, under which the student choses the correct word. XIX
Figure A4: Example item of subtest 4: The given sentence says “A face has two… .”. The buttons are labelled with the different choices “Finger” (AE finger), “Bücher” (AE books), “Augen” (AE eyes), and “Autos” (AE cars). XIX
Figure B1: First part of the simulation process in study 1. XX
Figure B2: Second part of the simulation process in study 1. XXI
Figure B3: Simulation Process of Simulation 1 without Subtest Linking, and Simulation 2 with Subtest Linking. XXII
Figure B4: Simulation Process of Simulation 3 without Subtest Linking, and Simulation 4 with Subtest Linking. XXIII
Figure B5: Simulation Process of Simulation 5 without Subtest Linking, and Simulation 6 with Subtest Linking. XXIV
Figure C1: Graphical Model Check with median split for subtest 4 before removing bad fitting items. XXV
Figure D1: Scatterplot of true and estimated ability levels of subCAT 1 with SE = 0.5. .. XXVI
Figure D2: Scatterplot of true and estimated ability levels of subCAT 2 with SE = 0.5. .. XXVI
Figure D3: Scatterplot of true vs. estimated ability levels of subCAT 3 with SE = 0.5. ... XXVI
Figure D4: Scatterplot of true vs. estimated ability levels of subCAT 4 with SE = 0.5. .. XXVII
Figure D5: Conditional proportions of test runs of subCAT 1 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true thetas. XXVII
Figure D6: Conditional proportions of test runs of subCAT 2 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true ability levels. XXVII
Figure D7: Conditional proportions of test runs of subCAT 3 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true ability levels. XXVIII
Figure D8: Conditional proportions of test runs of subCAT 4 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true ability levels. XXVIII
Figure D9: Test length of SubCAT 1 as a function of cumulative percent of examinees. XXVIII
Figure D10: Test length of SubCAT 2 as a function of cumulative percent of examinees. XXIX
Figure D11: Test length of SubCAT 3 as a function of cumulative percent of examinees. XXIX

Figure D12: Test length of SubCAT 4 as a function of cumulative percent of examinees. XXIX
Figure D13: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = -1 for subtest 2 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXX
Figure D14: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = -1 for subtest 3 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXX
Figure D15: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = -1 for subtest 4 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXXI
Figure D16: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = -3 for subtest 2 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXXI
Figure D17: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = -3 for subtest 3 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXXII
Figure D18: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = -3 for subtest 4 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXXII
Figure D19: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = 3 for subtest 2 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXXIII
Figure D20: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = 3 for subtest 3 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXXIII
Figure D21: Item Difficulties (sigma) and theta estimates while a test run (in items) of θ = 3 for subtest 4 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers. XXXIV
Figure D22: Scatterplot of true vs. estimated θ of the generated item pool and θ of subtest 1. XXXIV
Figure D23: Scatterplot of true vs. estimated θ of the generated item pool and θ of subtest 2. XXXV

Figure D24: Scatterplot of true vs. estimated θ of the generated item pool and θ of subtest 3. XXXV
Figure D25: Scatterplot of true vs. estimated θ of the generated item pool and θ of subtest 4. XXXV
Figure D26: σ of selected items and estimated θ while a simulated test run with fixed first item of a person with true θ = 0.03 for subtest 1, true θ = 2.59 for subtest 2, true θ = 1.98 for subtest 3 and true θ = 0.23 for subtest 4. XXXVI
Figure D27: σ of selected items and estimated θ while a simulated test run with connected first item of a person with true θ = 0.03 for subtest 1, true θ = 2.59 for subtest 2, true θ = 1.98 for subtest 3 and true θ = 0.23 for subtest 4..... XXXVII

List of Tables

Table 1: Item parameter range of the Rasch model. 24
Table 2: Item and person parameter estimators of the Item-Response-Theory. 26
Table 3: Results of subCAT simulations comparing three different SE-based stopping rules and the resulting average test length, percentage of test runs terminated with the stopping rule and correlation between true and estimated θ. 53
Table 4: Results of subCAT simulations comparing true and generated item pools. 57
Table 5: Means, standard deviations, and correlations with confidence intervals for the sum scores of the screening subtests. 61
Table 6: Results of screening CAT simulations with real and generated item pools and real and generated person abilities comparing different starting rules. 62
Table 7: Number and percentage of correct and wrong answers per subtest for different simulations of true and generated screening item pools. 71
Table C1: Item Difficulties (Sigma), Outfit and Infit MSQ of all Subtests each as one-dimensional Rasch Model. XXV
Table E1: Results of adaptive and linear screening simulations of real and generated item pools comparing different theta inputs of students with and without SEN and different start items. XXXVIII

1 Introduction

> "It's April, and a class full of students is about to tackle the rigorous, high-stakes statewide assessment. Last year they had all sharpened their number 2 pencils, sat down at their desks, opened thick test booklets, and begun filling in the blanks on paper answer forms. This year, these students are sitting in the school's computer lab (…). Two of them, a visually impaired boy and a girl who has severe dyslexia, wear headphones, working at computers that will read the test aloud. (…) The computers immediately score students' responses (…). Based on how a student answers particular questions, the computer branches to harder or easier questions, an adaptive process that yields more detailed information about what a student does or doesn't know. Because the tests are adaptive, students can also answer fewer questions than needed on a traditional paper assessment, yet the test yields more specific results. When each student is done, his or her test scores are automatically emailed to the teachers and principal."
> (Rabinowitz & Brandt, 2001, p. 3)

In 2001, Rabinowitz and Brandt described in this way their "vision of a more efficient and informative assessment process", which "sounds like a fantasy" (Rabinowitz & Brandt, 2001, p. 4). Essentially, they were describing what is now known as *Computerized Adaptive Testing* (CAT), which involves the use of algorithms to estimate student's abilities after each question and then select the next question based on their ability level. It should be noted that their vision also included considerations for students with disabilities, where the use of headphones was just one possible hardware-based accommodation for testing. This indicates that CAT was seen early on as providing opportunities for the assessment of students with various disabilities. In this example, additional tools are used to adjust the testing. But actually it is adaptive testing itself that has great potential for accurately measuring students with disabilities, as this work also will show.

But why do you even need to assess those students? Since the introduction of the UN Convention on the Rights of Persons with Disabilities in 2008, students with disabilities have had the right to be educated in an inclusive school setting. However, the enforcement of this right varies from country to country (Carrington et al., 2019; Grynova & Kalinichenko, 2018; Spulber, 2015). Despite this, the ultimate goal remains the same: to provide support measures that fit each student's unique special needs. Unfortunately, resources to support student learning are often limited (Goldan & Schwab, 2020; Hayes & Bulat, 2017; Lindsay, 2003), making it difficult to provide every student with the same level of support. Those who require more intensive support include students with disabilities or those at risk of developing disabilities. In order to provide the necessary resources for these students, it is necessary to identify their difficulties

as soon as possible and as early as possible to counteract the wait-to-fail problem. To identify these students or discern their exact level of learning, various types of educational assessments are used (see 2.1).

Educational assessment differs both in the reasons for its implementation and in its methods: On one hand, students are tested to diagnose a disability or disorder. This diagnosis is then used to assign support locations or intensities of support (Reschly, 1996). This form of diagnostics, while associated with student labeling, is necessary in the current school system to provide resources in the first place. On the other hand, assessments are also used to evaluate students' learning status and to derive support and tailored instructions from this (Bowen & Rude, 2006), or to track and compare their learning process (Harlen & James, 1997), for example in the frame of Data-based Decision Making (Lai & Schildkamp, 2013, DBDM; see 2.2). This can be done in individual, group or class settings. Overall, it can be said that assessments already play or at least should play a major role in student learning, as well as in assigning students to school systems.

Assessments in inclusive schools or classes face new challenges. Most assessments are designed for classes that show homogeneous performances or in which students do not differ significantly in their performances. However, such assessments cannot do justice to the increasing heterogeneity of students in the course of progressive inclusion because they are oriented towards the norm and expectations of performance based on the average standard. To better consider the heterogeneity of student performance within a class and to be able to provide suitable assessments in these settings, ideally, the assessments should be as widely applicable as possible while still taking into account the individual situation of each student. Moreover, they should be brief and easy to use in both individual and classroom settings, relieving the teacher of additional effort of evaluating and assigning support. Additionally, psychometric requirements should be met to deal with the heterogeneous target group of the assessments (see 2.3).

The use of digital technologies for student assessment (see 2.4.1) addresses many of these needs (Jungjohann et al., 2018; Ngwacho, 2022). In particular, adaptive testing is gaining popularity as a way of assessment, as such tests can adjust their difficulty to different target groups. CAT is the methodical and digital implementation of adaptive testing (see 2.4.2). It is a form of testing in which the difficulty level of the test adapts to the student's ability level (Meijer & Nering, 1999). This means that many students in a class can take the same test, but still receive a selection of items tailored to their level of difficulty.

The basic features of CAT have been discussed since the 1960s (Brown & Weiss, 1977; Sands et al., 1997; D. J. Weiss & Betz, 1973) and have been developed and implemented in various fields with increasing digitization. Widely used CATs are primarily used in higher education (Kalender, 2012), the military (Sands & Gade, 1983), psychology (Gibbons et al., 2012), and medicine (Papuga et al., 2018). However, there is comparatively little research in the area of primary and secondary schools. Additionally, there is a research gap in the inclusive school sector, particularly in the assessment of students with disabilities, both in special education assessment research and in methodological CAT research.

This work attempts to partially fill this research gap by presenting three studies conducted on the topic of CAT in inclusive education. The studies aim to develop a CAT for an inclusive target group and examine the performance and response behavior of students with and without disabilities. To accomplish this, both an existing inclusive reading screening and artificially generated data will be used. Different adaptive algorithms will be tested for their performance for different target groups in order to develop the CAT.

Each study poses different research questions. However, two main research questions are present throughout the entire research process:

1. What are the best methodological and technical approaches to develop an inclusive CAT based on a non-adaptive reading screening?
2. What are the performance and response patterns of students with and without Special Educational Needs (SEN) in CATs with different settings?

To address these research questions, the second chapter of this thesis covers a literature review on the theoretical background of assessing students with SEN (see 2.1-2.3). Additionally, it explains the psychometric properties and development of CAT (see 2.4) and emphasizes the prerequisites and background necessary for developing an inclusive CAT.

The third chapter explains the instruments and samples used, as well as the research questions and procedures for the three interrelated studies. In the first study (see 3.3.1), four one-dimensional adaptive tests are developed from a four-part reading screening. Various technical and methodological options for the CAT are compared. Additionally, the suitability of each of the four tests for CAT is assessed, and if necessary, the extent to which the test needs to be adapted is determined. In the second study (see 3.3.2), the results of the first study are utilized to determine whether implementing four separate CATs or combining them into a single screening

CAT is preferable. The third study (see 3.3.3) simulates and compares the performance of different groups of students with SEN on the developed CATs. This study addresses the question of whether and to what extent the tests need to be further adapted for these groups of students.

The fourth chapter (see 4.1 – 4.3) presents the findings of the three studies. These findings offer insights into the best methodological and technical approaches for developing an inclusive CAT based on a non-adaptive reading screening, as well as the performance of students with and without SEN in CATs with different settings. Each study is concluded by summarizing and discussion the study`s findings. The fifth chapter (see 5.1 – 5.4) provides a discussion of the research findings of all three studies. This chapter provides an analysis and interpretation of the research results and the research process on CAT in inclusive education, along with a reflection on the research process and limitations.

2 Literature Review

2.1 Struggling Students and their (Dis-) Abilities

2.1.1 Labeling and Identification

In every class, there are students who struggle with learning. They often are also called at-risk students (L. S. Fuchs et al., 1992; Slavin et al., 1989), low-performing students (L. S. Fuchs et al., 2015), low-ability students (Rulison & Loken, 2009) or slow learners (Kaznowski, 2004). Regardless of the terminology, what these students have in common is that they struggle to keep up with the regular level of performance in their grade level. It is estimated that up to 20% of students have problems following the subject matter in school. The reasons for this vary but are generally attributed to learning or attention issues (National Center for Learning Disabilities, 2017a).

Some students are diagnosed with a disability or what is known as *special educational needs* (SEN) before or during their school years. The term SEN has been used in English-speaking countries since the late 1960s and arose from a rejection of the medically oriented term *handicap*. Instead of categorizing children and adolescents based on their handicaps, the concept of SEN should categorize them based on personal, social, and especially educational issues. The conceptual basis of this idea was documented in the Warnock Report of 1978 (Department for Education and Science, 1978; Lindsay et al., 2020), which „pointed out that whether a disability constitutes an educational handicap for a child depends upon many factors such as the school's expertise and resources, the child's temperament and personality, the quality of support and encouragement within the family and environment." (Gulliford & Upton, 2002, p. 2). Therefore, the term SEN is recommended in the Warnock Report, in order to include not only a child's disability, but all factors „which have a bearing on his *[sic!]* progress" (Gulliford & Upton, 2002, p. 2). A broad definition is used that includes all students who need:

- „special means of access to the curriculum through special equipment, facilities or resources, modifications of the physical environment or specialist teaching techniques.
- the provision of a special or modified curriculum.
- particular attention to the special structure and emotional climate in which education takes place." (Gulliford & Upton, 2002, p. 3)

The term SEN or modifications of this term such as *Special Educational Needs and Disabilities* or *Special Needs* have persisted in the English language for students with disabilities to this day (e.g., Asbury et al., 2021, Ebenbeck et al., 2022, Lutz et al., 2022). As part of the inclusion movement, the categorization of students through such terms is increasingly viewed as a form of labeling. The extent to which this labeling is necessary is a subject of debate (Demetriou, 2020; Gebhardt, 2023, p. 50). Although labelling can lead to stigmatization and negative exacerbations in some cases (e.g., Franz et al., 2023, Shifrer, 2013, Jones, 1972, Green et al., 2005), it can also provide access to additional resources and support for students (e.g., Arishi et al., 2017, Norwich, 2009). For instance, students with diagnosed disabilities may receive special education services, including Individualized Education Plans, modifications to assessment procedures, therapies, individual support in school, or access to special schools in some countries. However, it is believed that many students with disabilities are not identified, leading to a high number of unidentified cases (National Center for Learning Disabilities, 2017b).

The designation of SEN is internationally a categorization of the school system to enable resources for individual students. The assignment of a SEN is necessary to provide these resources, which is why identifying these students is important. Therefore, it is not only important to identify whether SEN are present, but also in which area students need support. However, what falls under the different types of categorizations varies from country to country and from education system to education system. In the USA, for example, gifted students or students with a specific difficulty in a learning area are also included in the category of SEN. In Germany, the view on SEN is different, and such students are not included under this term (Gebhardt et al., 2013; Gebhardt, 2023; Grünke & Cavendish, 2016). Therefore, the definition of a SEN must be considered in the context of each country to understand the group categorized by it.

Depending on the country, there are different types of SEN. In countries like Germany, where there is not yet a comprehensive inclusive school system (Ebenbeck et al., 2022), the type of SEN also impacts the extent to which a child is likely to be educated inclusively or not. For example, in Germany, students with SEN in the area of learning (SEN-L) are most commonly educated inclusively, making up 47% of students with SEN in general schools. Students with SEN in the area of emotional and social development (SEN-E) make up 23%, students with SEN in the area of speech, language and communication (SEN-S) make up 11%, and students with SEN in the area of intellectual development (SEN-I) make up 6% (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2022, p. 6).

2.1.2 Types of Special Educational Needs

As already explained, the definition of individual types of SEN depends on the country and perspective. From an international and medical perspective, students with SEN are referred to as individuals with *learning disabilities*. In the 1970s, the definition and identification of a learning disability heavily relied on the intelligence quotient (IQ), which follows a normally distributed value range and is the result of an intelligence test (D. Fuchs et al., 2003). An IQ between 70 and 84, which represents at least one standard deviation below the mean, was considered to indicate a learning disability.

Based on the criticism of IQ and its psychometric properties (e.g. Beaujean et al., 2018; Gardner, 2005; Restori et al., 2009; Sternberg et al., 2001; Weinberg, 1989), recent research increasingly includes the student's environment and instructional response in the definition and identification while opening up the definition of learning disability (Fletcher et al., 2004; D. Fuchs et al., 2003; Kovaleski et al., 2013; Lyon et al., 2001; Madalyn, 2021; Maki & Adams, 2019). Also, there are indicators, that learning disabilities are not connected to an traditional IQ range per se and the IQ therefore should not be used for identifying (Siegel, 1989; Francis et al., 1996; Gunderson & Siegel, 2001; Naglieri & Reardon, 1993; Rispens et al., 1991; for a contrary opinion please see Torgesen, 1989; Lyon, 1989). Although IQ should no longer be used as the determining criterion for diagnosing learning disabilities, it still plays a role for categorizing. For example, distinctions are made between people with learning disabilities and with high and low IQ for research purposes (Lovett & Lewandowski, 2006; Rooney et al., 1985).

Learning disabilities are defined more medically (Cortiella & Horowitz, 2014). The current international valid medical classification system of the American Psychiatric Association (2022) DSM-5 uses the term *specific learning disorder* and defines it as follows. It does not use the IQ as diagnostic criterion, but as delimination criterion from SEN-I or intellectual disabilities:

> "Difficulties learning and using academic skills (…) that have persisted for at least 6 months, despite the provision of interventions that target those difficulties. (…) The affected academic skills are substantially and quantifiably below those expected for the individuals chronological age, and cause significant interference with academic or occupational performance, or with activities of daily living, as confirmed by individually administered standardized achievement measures and comprehensive clinical assessment. (…) The learning difficulties begin during school-age years but may not become fully manifest until the demands for those affected academic skills exceed the individual's limited capacities (…). The learning difficulties are not better accounted for by intellectual disabilities, uncorrected visual or auditory acuity, other mental or neurological disorders, psychosocial adversity, lack of proficiency in the language of

> academic instruction or inadequate educational instruction." (American Psychiatric Association, 2022, 66-67).

In the context of the school system, SEN-L is a term used for students with difficulties in learning. In the USA, students with SEN-L are diagnosed most frequently and make up about 5% of the total student population (Cortiella & Horowitz, 2014, p. 12), which also includes students with specific difficulties in individual areas of learning. Compared to other types of SEN, it is harder to identify as there are no clear diagnostic criteria such as an IQ threshold or noticeable physical or sensory deficits (Zydney et al., 2020). In contrast, in Germany, only those students who have significant difficulties in academic learning that cannot be adequately addressed by regular schools are categorized under SEN-L. The reasons for the difficulties in academic learning are complex and include risk factors of the child and the environment (KMK, 2019). Therefore, in addition to the child's individual experiences and competencies, both social comparison to peers and the adaptability of the general school are crucial for the emergence, identification, and perception of learning problems (Gebhardt, 2023).

Students with SEN-L often experience challenges with their behavior or emotional development. In a long-term study conducted by Vaughn and Schumm (1995), this group of students was found to have significantly lower social skills coupled with significantly higher behavioral problems compared to students without SEN. Students with SEN-L have been shown to exhibit higher levels of hyperactivity and distractibility, as well as more emotional problems, lower frustration tolerance, and weaker work habits in the classroom than students without SEN (Bender & Smith, 1990; Cullinan et al., 1981; Silver, 1981; Toro et al., 1990).

In an educational context, the student's response to support, instruction, and intervention is observed to determine the presence of learning disabilities, which can affect a variety of sub-areas such as math, spelling, or reading. Consequently, some students with learning disabilities may also have weaknesses in prerequisites that impact their reading abilities. Dyslexia is the most common type of learning disability, characterized by difficulties with phonemic awareness, phonological processing, and word decoding, fluency, rate of reading, rhyming, spelling, vocabulary, comprehension, and written expression. Other forms of learning disabilities may also struggle with some of these areas, such as students with auditory or visual processing deficits (Cortiella & Horowitz, 2014, pp. 3–4). Students with SEN-L frequently read significantly worse than their peers without SEN. This was confirmed by a meta-analysis by Gilmour et al. (2019), which found that students with SEN-L in the US read 1.44 standard deviations worse than students without SEN, which equates to falling about four years behind in learning.

Students with Special Educational Needs in Intellectual Development

People with an *intelligence disability* make up around 1% of the global population (American Psychiatric Association, 2022, p. 38). The definition and identification of intellectual disability are primarily medically influenced internationally. DSM-5 defines intellectual disability as a neurodevelopmental disorder and includes three criteria: first, deficits in intellectual functioning confirmed by clinical and intelligence assessment, second, deficits in adaptive behavior (AB), and third, the fact that these deficits have been present since childhood.

Country-specific definitions are based on this definition. In the USA, from an educational perspective, intellectual disability is defined as follows:

> Intellectual disability "is characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social and practical adaptive skills. This disability originates during the developmental period, which is defined operationally as before the individual attains age 22." (Schalock et al., 2021, p. 439)

While the IQ is still used as a diagnostic criterion for intellectual disability, it is also criticized (Arvidsson & Granlund, 2018; Greenspan & Woods, 2014; Zucker & Polloway, 1987). An IQ score of two standard deviations or more below the mean is considered to diagnose an intellectual disability based on the IQ. However, in addition to the IQ, AB is becoming increasingly important for identifying children with intellectual disability. AB "is defined as the collection of conceptual, social, and practical skills that have been learned and are performed by people in their everyday lives" (Tassé et al., 2012, pp. 291–292).

People with intellectual disabilities are a diverse group with a wide range of abilities. As a result, it is common to classify subgroups using different indicators such as IQ scores or AB scores. IQ and AB scores use the same value ranges. Typically, students are classified as mild (IQ between 50/55 and 70/75), moderate (IQ between 40/45 and 50/55), severe (IQ between 25/30 and 40/45), or profound intellectual disability (IQ < 20/25) (Patel et al., 2018).

In Germany, the medical perspective is still used for the identification and categorization of intellectual disability, although it is increasingly being supplemented by the definition of AB (Dworschak & Kölbl, 2022).

IQ is commonly used to distinguish between people with intellectual disabilities and those with learning disabilities, although the latter group is no longer defined by IQ. However, the IQ range of 70-84 is still sometimes used as a differentiator between individuals with learning difficulties

(Cornoldi et al., 2014). This range was previously considered to be characteristic of learning disabilities but is now referred to as borderline intelligence in psychology and medicine. The term is used to distinguish individuals with an IQ between 70 and 84, which overlaps with the group of people with learning disabilities, from those with intellectual disabilities. However, IQ is a weak indicator for categorizing, and people with borderline intelligence and mild forms of intellectual disabilities are sometimes treated as one group in research, even though they may differ in their performance and characteristics (Arvidsson & Granlund, 2018; Bouck & Satsangi, 2015; Cohen et al., 2001; Nouwens et al., 2017).

The transitions between students with learning disabilities, borderline intelligence, and mild intellectual disabilities are fluid and often not clearly delineated. According to Snell et al. (2009), individuals with mild intellectual disabilities share many characteristics with those with borderline intelligence, leading to an overlap between students with learning disabilities and intellectual disabilities. In a comparative study, Gresham et al. (1996) examined the differences between struggling students with low academic achievement, students with learning disabilities, and students with mild intellectual disabilities. The results showed that the differences were mainly related to cognitive abilities and academic achievement, with fewer differences in social skills or problem behavior. For instance, in academic achievement, students with learning disabilities showed higher skills than those with mild intellectual disabilities. In terms of cognitive abilities, students with learning disabilities scored the highest, followed by students with low academic achievement, and students with intellectual disabilities who scored the lowest. Although cognitive measurements could differentiate between students with learning disabilities and mild intellectual disabilities with 99% accuracy, reading skills only achieved a 71% differentiation rate. Gresham et al. (1996) also argue, that despite similarities between the groups, “the differences that do exist are probably not educationally relevant in terms of differential placement options or interventions”.

Another conclusion that can be drawn from the results of Gresham et al. (1996) is related to the reading skills of students with learning disabilities, borderline intelligence, and intellectual disabilities. Although IQ is commonly used to differentiate these student groups, there may be a connection between reading skills and IQ or student group diagnosis. Other studies support this idea: Cohen et al. (2001, pp. 71–72) found a correlation between IQ and reading performance ($r = .44$) in adults with intellectual disabilities and borderline intelligence. Individuals with IQ < 65 were unable to solve the reading items adequately. Levy (2011) also observed a correlation

between word reading and IQ in adults with intellectual disabilities and borderline intelligence, with a correlation coefficient of up to r = .893. In a study by Di Blasi et al. (2019, pp. 1028–1029) significant ($p < 0.01$) differences in reading skills were reported between students with mild intellectual disabilities, BI, and those without a disability. Students with intellectual disabilities had worse reading skills than those with borderline intelligence. The greatest differences between students with mild intellectual disabilities and borderline intelligence were observed in reading accuracy, followed by reading speed. There was little difference in text comprehension, with both groups of students showing average performance slightly above the minimum possible performance.

Students with Speech and Language Impairment

Similar to SEN-I, the medical perspective is also predominantly used internationally for the definition and identification of students with SEN-S, which serves as the basis for identifying SEN in this area in schools. In the USA, these students are referred to as students with Speech and Language Impairment, while in Germany, the term "Förderbedarf Sprache" (translatable as SEN Speech, Language, and Communication) is used as a description in schools.

Students with SEN-S face difficulties in academic learning due to developmental difficulties in the area of language and communicative action. Students with SEN-S are not defined by reduced intelligence and have a non-verbal IQ in the normal range (Botting, 2005; Earle et al., 2017). Instead, they exhibit various issues in processing language and communication. The processing of language is a necessary requirement to integrate learning content into one's own cognitive system. As this ability is not sufficiently developed in many students with SEN-S, these students increasingly develop problems in various academic areas. One significant area is reading, particularly reading comprehension, which makes it difficult for them to independently acquire learning content. In addition to these difficulties in acquiring academic content due to impaired language and reading comprehension, children with language acquisition disorders frequently lack the necessary semantic-lexical and grammatical skills to adequately express academic content (Mayer & Motsch, 2016). Students with SEN-S exhibit a heterogeneous skills profile in reading. Depending on the specific issue, students with SEN-S read weaker than students without SEN (Boudreau & Hedberg, 1999; Simkin & Conti-Ramsden, 2006; Werfel & Krimm, 2017). Their difficulties are evident in both phonological awareness (e.g. Claessen et al., 2013, Farquharson et al., 2014), reading fluency (e.g. Isoaho et al., 2016,

Puranik et al., 2008), and reading comprehension (e.g. Coloma et al., 2015, Jungjohann, 2022, Kelso et al., 2007, Ricketts, 2011).

In summary, it can be said that students with SEN are not a clearly defined group. Their identification depends on various circumstances and assessments. SEN itself should therefore be viewed as a label that offers both advantages and disadvantages. Although there are different groups of students with SEN, the assignment to a specific SEN category depends on different factors depending on the type of SEN. However, the boundaries between SEN categories are fluid. In particular, the assignment in the border areas between SEN categories is contentious and cannot be clearly defined. In the area of cultural techniques, such as reading, there are also many students who have difficulties. These difficulties are not linked to a specific SEN category or SEN particulary, but are noticeable in several categories of SEN. This also suggests that the type of SEN is secondary for school support and that the performance of students in their skills and assessments should be more decisive.

2.2 Educational Assessment in Inclusive Education

According to Gebhardt (2023), categorizing students in schools for the provision of resources, such as SEN, is no longer considered contemporary. As previously demonstrated, categorizing students into different types of SEN is insufficient for predicting their performance in a subject or skill area. Therefore, students can have problems with reading regardless of whether they are classified as having SEN or the type of SEN they are assigned. Additionally, interventions should focus on the individual who requires tailored support, rather than a systemic category. Therefore, providing effective support and instructions requires the right support decision to be made.

2.2.1 Data-based Decision Making

Struggling students, regardless of whether they have a diagnosed disability or SEN, need to be supported in their learning as soon as possible. Early interventions can prevent long-term issues from arising (e.g. D. Fuchs & Fuchs, 2011; Menzies et al., 2008; Partanen & Siegel, 2014; Schwartz, 2005). For example, if a student is already experiencing reading difficulties, it can become a persistent issue that is difficult to improve in the long term. This is supported by

research: Jacobson (1999) observed the word reading of students with and without disabilities in grade 2 and again seven years later in grade 9. Most of the students observed were unable to make up for the reading difficulties during this period and continued to struggle. These persistent reading difficulties are also decisive for further academic performance problems since reading is an important basis for information acquisition. In this context, Hakkarainen et al. (2013) confirmed that reading difficulties can have a strong impact on a student's academic performance in general, even up to grade 9. Therefore, it is crucial to identify and address early signs of reading problems with appropriate instruction and support to prevent long-lasting school problems. This is especially true for students with SEN, who acquire reading skills more slowly than students without SEN and exhibit weaker reading skills (Badian, 1999, p. 56; Vaughn & Wanzek, 2014).

To do so, teachers evaluate their students' performance in various ways and then assign them learning methods, support, and instructions. The correct assessment of student performance is critical to assigning appropriate instruction. However, the accuracy of the student's assessment depends heavily on the teacher. While teachers are generally capable of accurately judging the performance of their students, their judgment accuracy is lower for low-performing students than for high-performing students (Begeny et al., 2008; Coladarci, 1986, p. 144). This is critical because these low-performing students rely on early and appropriate intervention to avoid developing persistent difficulties.

To achieve a more accurate and justified support and instruction planning in classrooms, modern principles of special and inclusive education, like *response to intervention* (D. Fuchs & Fuchs, 2006), the *Rügen Inclusion Model* (Voß & Blumenthal, 2019), the *Multi-Tiered System of Supports* (Stoiber & Gettinger, 2016), *Positive Behavioral Interventions and Supports* (Sugai & Simonsen, 2012), or *Data Wise* (Boudett Parker et al., 2006) therefore rely on *Data-based Decision Making* (DBDM). DBDM describes an approach to data-based support planning, where decisions about support are made based on data about each student's learning, rather than the teacher's personal perception and judgement (Lai & Schildkamp, 2013). Using DBDM can lead to increased student learning progress. In mathematics, reading comprehension and spelling, positive effects with an improvement of 0.37 standard deviations (*SD*) in student achievement have been observed. Low-performing students have the strongest effects on learning with DBDM (Filderman et al., 2018). In a meta-analysis on DBDM reading interventions

for struggling readers, Filderman et al. (2018) also reported medium-strong weighted mean effects (.24 < g < .27) of DBDM instructions.

DBDM is a four-step iterative process (Lutz et al., 2022; Schildkamp, 2019). The first step is to set individual learning goals for each student. This is determined by a team that may include teachers, educators, therapists, and parents. The team discusses the learning and development goals, as well as associated support measures, based on the student's developmental level. They take into account existing data and determine what additional data needs to be collected for problem-solving. The second step is to collect data related to the learning goal. To improve the quality of teaching and student support, a variety of data sources can be used, including both informal and formal sources. In the third step, the collected data is analyzed and interpreted. Teachers must have sufficient diagnostic competencies to assess the material properly and use it to inform support decisions. In the final step, planned support is implemented and evaluated in relation to the collected data and adopted learning goals. Support methods are continuously evaluated during the student's learning process to determine if appropriate support is being provided. If not, the support can be adjusted during the learning process. Ideally, only evidence-based instructions whose effectiveness has been confirmed in research should be used to ensure the effectiveness of the support for students.

There are several types of data that can be collected for DBDM, such as formal data, informal data, research results, and big data (Schildkamp, 2019, pp. 261–262). It is recommended to focus on formal, systematic, and standardized assessment data, as these data ensure better comparability, reliability, and validity. Four types of formal assessment for decision making can be distinguished (Gebhardt, 2022; Hasbrouck & Tindal, 2006):

- *Screening measures*, which are brief assessments focusing on main components of a larger skill area (e.g. reading) that predicts the student's future development. They mostly are conducted within a school class at the beginning of a school year or intervention to identify students struggling with the measured skill area.
- *Diagnostic measures*, which are standardized tests that can be administered at any time during the school year in order to test a student as comprehensively as possible in one competency area at one point in time. Such tests are used to create a fully comprehensive profile of a student to determine, for example, a disability.
- *Progress-monitoring measures*, which consist of at least three shorter tests collected at continuous intervals to show the learning progress of a child over time. These tests are

used to assess the student's learning development, identify students who are not making adequate progress, and evaluate different instructions for struggling students.

- *Outcome measures*, which are comprehensive school achievement tests to determine whether the student has met the class target or assessment criterion. Such tests are administered either at the end of an intervention or at the end of the school year.

The combination of screening and progress-monitoring measures is a common concept in inclusive education used to assess skills related to school subjects such as reading, writing, or mathematics (e.g. L. S. Fuchs et al., 2007; Saddler & Asaro-Saddler, 2013, Stecker et al., 2008; Wilcox et al., 2021, p. 5). Screening measures are typically used at the beginning of the school year to identify students who are at risk of developing difficulties or who already struggle with a particular skill. All students in a class are usually assessed using these measures. Students who score below a certain criterion or percentile may be considered at risk of developing difficulties or disabilities and may receive preventive or intensified instruction and support. The students' learning progress is then evaluated using progress-monitoring tests administered at regular intervals (Stecker et al., 2008). If students have achieved the targeted learning goals, additional support can be discontinued. The long-term goal of DBDM is not to permanently stigmatize students with a label but rather to provide early and preventive support to avoid identifying and labeling students in the school context (Lutz et al., 2022).

DBDM has become a widely recognized and promising approach to educational assessment and support planning. It has been successfully implemented in various countries around the world, such as the United States (Blumenthal et al., 2021; Jimerson et al., 2015; Preston et al., 2016) and the Netherlands (Gelderblom et al., 2016; van Geel et al., 2016). In recent years, DBDM has also been utilized in developing countries (Schildkamp et al., 2019). However, despite its potential benefits, implementing DBDM in schools can be challenging for both teachers and schools (Hoogland et al., 2016).

One of the key challenges of DBDM is the need to gather a wide range of data on students' learning abilities through the use of multiple forms of assessment. This can be time-consuming and resource-intensive, and it requires teachers to be trained in the administration and analysis of these assessments. Many teachers may not feel prepared to handle these tasks and may struggle to make sense of the resulting data. Additionally, the use of multiple forms of assessment can be overwhelming for teachers, who may already be struggling with a heavy workload and limited resources. It can be difficult for them to integrate the use of these assessments into their

daily teaching practices, especially if they do not have the necessary support and training. This can lead to frustration of teachers, which can ultimately hinder the success of DBDM in schools (Wilcox et al., 2021, pp. 2–4).

Despite these challenges, it is important to recognize the potential benefits of DBDM for both students and teachers. When implemented effectively, DBDM can provide valuable insights into students' learning abilities and needs. It can also help teachers to identify struggling students early on and provide targeted support, leading to better outcomes for students. With training and support, teachers can successfully implement DBDM into their teaching practices and use it to support student learning.

2.2.2 Psychometric Requirements for Inclusive Screening Measures

Assessments in inclusive education, such as those conducted within the framework of DBDM, should focus on identifying struggling students in the classroom so that they can receive additional support and instruction. Standardized instruments have high value in DBDM, as they can reliably and accurately assess students' performances and enable comparisons across different students. Screening measures, which can assess the whole class in a single sitting, are a type of assessment that can efficiently identify students who may need additional support. However, in order for standardized testing to work effectively in an inclusive setting, several psychometric aspects must be considered. While this work focuses on the requirements for screening measures, many of the aspects mentioned can also be applied to other types of assessments.

Screening tests must meet general scientific test quality criteria, regardless of whether they are designed for use in inclusive or general classroom. According to Moosbrugger and Kelava (2020), these criteria include reliability, validity, objectivity, economy, usefulness, reasonableness, fairness and unforgeability (for a detailed description of each criterion, see Moosbrugger & Kelava, 2020). The reliability and validity of a test depend on the underlying psychometric theory, while the other mentioned quality criteria are more influences by the general planning of test design and administration (Moosbrugger & Kelava, 2020, p. 27).

Objectivity in testing typically results from the standardization and scaling of the test (Moosbrugger & Kelava, 2020, pp. 18–19). The more precise the instructions for administration and evaluation, the more objective the instrument can be assumed to be. However, experimenter effects or mistakes cannot be completely eliminated, even in standardized assessments

(Vormittag, 2011). Experimenter effects refer to the influence that the experimenter or test examiner has on the examinee's behavior. This can be due to factors such as the examiners personality, experience, expectancy, sex or gender (Kintz et al., 1965). In the context of assessment, experimenter effects can also influence the evaluation and interpretation of test scores as in the case of DBDM (Kintz et al., 1965, pp. 229–230). Additionally, examiners may unintentionally make mistakes while administering an assessment or evaluating test scores. For example, experienced examiners may rely on their memory instead of reading out the standardized instructions of a test, leading to minor mistakes that can affect the test score and objectivity of the test (Bundschuh & Winkler, 2019, p. 84). Errors may also occur when the test examiner lacks sufficient experience with the respective test instrument.

A test is considered economic if it requires few financial or time resources. The time needed to administer the test and prepare, follow up, and evaluate it is important for its use in the classroom (Moosbrugger & Kelava, 2020, p. 24). Screening tests that require little time do not take away valuable student support time and can be conducted efficiently in a group or class setting (Bundschuh & Winkler, 2019, p. 120). Simple evaluation procedures can save time for the teacher, which can then be used for student support or instruction (Buchwald et al. 2022). Short test administration times are particularly beneficial for students with SEN, who may struggle with concentration and attention during testing (Tarver & Hallahan, 1974). Therefore, shorter test administration times also contribute to the fairness and reasonableness of a test.

A test is considered reliably if it measures accurately without measurement error (Moosbrugger & Kelava, 2020, p. 27). Most tests have high accuracy in the medium performance range, as the difficulty of the items is normally distributed. However, in more extreme areas two or three SD away from the mean, their accuracy is much lower. This is because there are fewer items for these performance ranges, and there are not enough easy items for the lower end of the normal distribution (Schurig & Gebhardt, 2022, p. 242). As a result, low-performing students cannot be accurately measured with these tests, or they are measured much more inaccurately than other students. The lack of sufficient items in the lower difficulty range means that their performance can only be poorly depicted. These so-called *bottom effects* should be avoided as much as possible in inclusive assessments. To avoid bottom effects and use differential measures, there should be enough easy items in the item pool (Eigner, 2022, p. 423).

In summary, educational assessment plays a crucial role in data-driven, inclusive approaches such as DBDM, which aim to use data and analytics to improve student outcomes.

Approximately 20% of students struggle with learning and require tailored support and instruction. Some of these students may have been diagnosed with a disability or SEN, but the boundaries between these categories are often blurred in practice. However, low-performing students, including those with disabilities, are often evaluated less accurately than high-performing students by teachers. To address this issue, it is important to use assessments that are both accurate and efficient. Teachers need assessments that are easy to administer, analyze, and evaluate, and that take minimal time. Administration methods that minimize experimenter effects and mistakes should be preferred. Students, especially those with low performance, need assessments that accurately measure them in a short amount of time. Many assessment instruments are more accurate in the middle ranges and less accurate in extreme areas. Therefore, inclusive assessment instruments should include enough easy items at the lower end to accurately measure students with low performance.

2.3 Computerized Adaptive Testing

2.3.1 Chances of Computers in Educational Assessment

Many of these requirements can be met by changing the way assessments are conducted. Traditionally, tests were administered using a paper-and-pencil format (paper-based testing, PBT). However, with the widespread availability of computers in schools, computer-based testing (CBT) is becoming increasingly popular for educational assessments (Bennett, 2002).

There are different kinds of CBTs depending on how computers are used throughout the assessment process (Conole & Warburton, 2005). For example, many standardized PBTs are evaluated with the aid of computer-based software. This way of using computers for assessment is sometimes also called computer-assisted assessment (CAA, Conole & Warburton, 2005). For CAA, the test examiner must transfer the students' analog answer sheets into the software. Therefore, the use of CAA does not affect the student themself. After the transfer to the computer, the answers are then evaluated automatically. The execution of the test is still analog, only the evaluation has been transformed. This digital evaluation saves time, is easier to perform for teachers than an analog evaluation and is less prone to errors.

Most often, however, the term CBT is used to refer to tests that examinees take themselves on a computer or other digital device (Thelwall, 2000). In this context, CBTs are distinguished between four generations (Bunderson et al., 1988; Redecker & Johannessen, 2013): The first

generation is *computerized testing*, the second generation is *computerized adaptive testing*, the third generation is *continuous testing*, and the fourth generation is *intelligent testing*. Tests of Generation 1 or 2 focus on the measurement itself, while tests of Generation 3 or 4 focus on personalized learning and have integrated measurement inside their learning environments (Redecker & Johannessen, 2013, S. 82).

In *computerized testing* (generation 1), examinees take a conventional test, but with a changing administration mode. This can be due to the digitalization of existing paper-and-pencil tests or the creation of new items formats for computerized testing (Parshall et al., 2000). The execution takes place on the computer or tablet by displaying the items on the screen and indicating the answer via mouse, keyboard or touch display.

This computer-only administration leads to higher psychometric quality, as it promises better objectivity, economy, and usability for teachers and fairness for students (Liebers et al., 2019, pp. 211–212). Experimenter effects or mistakes during the administration can be reduced as students administer the items on the computer, and no examiner is necessary to provide the items (Schaper, 2009, p. 26; Walter & Schuhfried, 2004, p. 265). Also, like CAA, test examiners no longer have to transfer the answers into software. Instead, the answers are automatically scored and evaluated digitally after administration. This means that no errors can occur when transferring the answers to the software, and the administration and evaluation of the tests are accelerated. Initial concerns about the validity and reliability of CBTs could not be confirmed (Piaw Chua, 2012): When comparing PBT and CBT, there were no systematic differences in the scores (Ebrahimi et al., 2019; Mason et al., 2001). This has been shown in higher education assessments as well as tests of skills and competencies like reading (Mojarrad, 2013; S. Wang et al., 2008).

For students, the administration of CBTs was challenging in the early days of CBT. They were not used to handling computers and using mice to answer questions (Latu & Chapman, 2002). Therefore, the question arose as to what extent computer skills are necessary to successfully complete CBTs and whether students with less experience with computers are disadvantaged in such assessments. However, it has since been found that computer familiarity has no effect on performance in CBTs (Jeong, 2014; McClelland & Cuevas, 2020, p. 87). This means that students without much experience with computers are not at a disadvantage. Students with SEN are also not disadvantaged by CBTs (Calhoon et al., 2000; Taherbhai et al., 2012). Furthermore, students prefer CBTs over PBTs (Mojarrad, 2013) and tend to be more motivated (Piaw Chua,

2012). Students complete CBT faster than PBT (Piaw Chua, 2012) while achieving the same scores, leading to a more efficient measurement. However, flexible modes that allow students to review their answers do not yield additional benefits in speed or accuracy (Bodmann & Robinson, 2004).

Computerized adaptive testing (CAT; generation 2) is an advancement of computerized testing. Like computerized tests, CATs are also performed completely digitally and therefore yield the same benefits as computerized tests. Both the test execution and the evaluation take place automatically. As an extension to computerized testing, however, an individual item selection is made for each examinee by an adaptive algorithm. The algorithm is used to estimate how strong the examinee's abilities are in the tested area and selects items at an appropriate level of difficulty.

In their literature review, Stone and Davey (2011) summarized the benefits of using CAT compared to computerized tests or PBTs, focusing on students with disabilities. CATs require less time for administration since examinees only need to take a smaller number of items. In the optimal case, they only need to answer fitting items that provide information about their ability. This leads to a test reduction of between 20% and up to 50%, depending on the psychometric model and the settings for the CAT's test termination (Flens et al., 2016; Stone & Davey, 2011, 6). With the selection of appropriate items, the issue of item pools that only focus on the middle range of difficulty also decreases. CAT has been shown to measure accurately at the extremes of the ability distribution, which leads to a more accurate measurement of gifted or low-performing examinees (Ebenbeck & Gebhardt, 2022; Flens et al., 2016; Stone & Davey, 2011, 6). The combination of testing in less time with growing accuracy makes CAT especially promising for assessments in inclusive school settings focusing on struggling students and students with disabilities. While early research claimed to have observed psychological effects of CAT on examinees (Betz, 1977; Linacre, 2000), recent research contrasted these findings: In a meta-analysis, Akhtar et al. (2022) compared the effects of CAT on test anxiety and motivation and concluded that there is no significant effect. Therefore, neither positive nor negative impacts of CAT on the psychological aspects of testing could be evaluated, but CAT seems to have the same psychological effects as CBTs with a fixed item order.

Tests of generation 3 and 4 take different approaches to measurement. In continuous measurement (generation 3), students' progress is evaluated over a period of time instead of a single measurement point. Data are collected regularly through various forms of assessments or

evaluations, and the data is later used to track students' progress and identify areas where they may be struggling or require additional support. Continuous measurement employs assessments to derive tailored support for all students (Bunderson et al., 1988).

In *intelligent testing* (Generation 4), artificial intelligence (AI) and machine learning or other advanced technologies are additionally used within the tests for various purposes (González-Calatayud et al., 2021). This can include adapting the test itself, classifying, evaluating and analyzing test results, categorizing student profiles, or providing feedback for students and teachers (McCusker et al., 2013, Shute & Zapata-Rivera, 2010, Zheng et al., 2020). Depending on the exact use of AI and machine learning, intelligent measurement can offer different benefits, such as real-time feedback and support for students and teachers, customization, improved accuracy and validity, enhanced efficiency, or greater insights into student learning and performance (Afzaal et al., 2021, Asthana & Hazela, 2020, Böhme et al., 2022, Hilbert et al., 2021, Korkmaz & Correia, 2019, Martin, 2008, Nafea, 2017). AI itself also offers various opportunities for people with intellectual disabilities besides assessment (Zentel et al., 2019).

Even if they provide more functions and benefits for educational assessment use, tests of generation 3 and 4 are more challenging to develop and implement at the moment compared to tests of generation 1 and 2. As they often involve the use of advanced technologies like AI or machine learning algorithms, they can be complex to develop and maintain. Such systems require large amount of data to be collected and analyzed, which is time-consuming and resource intensive (van Ooijen et al., 2022). This is also why the development and implementation of generation 3 and 4 tests are more expensive compared to generation 1 and 2 tests (Luckin, 2017), what can be a barrier for some educational institutions and organizations. An implementation of these types of tests also for smaller tests is therefore desirable for the future, but still faces technological challenges. In comparison, adaptive testing technologies are well-research and easier to develop. With new tools and the use of free programming environments (Chalmers, 2016; Han, 2012; Magis & Barrada, 2017; Nydick, 2022; Rizzo Meneghetti, 2016; Sorrel et al., 2021), they also get more and more accessible and cheaper in their development.

2.3.2 Components of Computerized Adaptive Testing

There are different types of adaptive testing. Two common types are tailored testing and branched testing (Amelang & Zielinski, 1997; Kubinger, 2003).

Branched testing presents the examinee with groups of items and selects a new group of items based on their performance on the previous group. There are fixed branched sequences that dictate which item groups must be selected (Kubinger, 2003). Branched testing can be administered in analog (Kubinger, 2007) or digital formats. Analog branched testing requires one-on-one administration, as the test administrator selects the appropriate item groups.

In *tailored testing*, the test adapts to the examinee by selecting a specific set of items based on an algorithm calculated by a computer. The algorithm follows the principle that after a correct answer, a more difficult item is selected, while after an incorrect answer, an easier item is chosen. Due to technical requirements, tailored testing can only be performed digitally, and is commonly known as CAT. Tailored tests or CATs can be administered in both one-on-one and group settings (Lord, 1968). They require fewer items per test run than paper-and-pencil tests or branched tests, and are therefore shorter while maintaining the same level of accuracy (Kubinger, 2017; Latu & Chapman, 2002).

CATs consist of four components that are characteristic and necessary for tailored testing (Bock & Gibbons, 2021, pp. 244–245):

- *A pre-calibrated item pool*. All items in the item pool must be calibrated according to a test theoretical model, such as an item response theory (IRT) model (for more detail, see section 2.3.2.1). This calibration allows for the estimation or calculation of information about every item. To obtain this information, all items need to be taken by a sample of the test's target group, which allows for the determination of item difficulty and other relevant metrics.
- *An item selection procedure,* which is a crucial part of the CAT algorithm. In order to select the next item after an examinee's response, the pre-calibrated item pool's information and a repeated estimation of the examinee's ability after each response are required (for further details, see section 2.3.2.2).
- *A test determination procedure.* Different stopping rules can be used in CAT, and the choice of stopping rule can affect the outcome of the test. The stopping rule used can vary from a fixed test length to achieving a certain level of accuracy in the estimation of the test-taker's ability, after which the test is concluded (see section 2.3.2.4 for more details).

- *A test scoring procedure.* In a CAT, examinees may get different numbers of items in a test run or items of different difficulties. Dependent on the type of test, a suitable scoring method must be found (see for more detail 2.3.2.5).

Additionally, there also needs to be a decision on *the test starting procedure* of a CAT (Thompson & Weiss, 2011, see 2.3.2.3). Different starting rules can be used for CAT and may also impact the resulting test. CATs can start with a fixed item or use previous information for the item selection (Magis & Raîche, 2012, p. 8).

2.3.2.1 Pre-calibrated Item Pool

2.3.2.1.1 Item Response Theory as Basis of CAT Item Pools

A CAT's item pool can be calibrated using different test theories. The use of item response theory (IRT) has been proven to be especially useful and effective (Bock & Gibbons, 2021, p. 245). It has been successfully used for the development of CAT and in CAT research since the 1970s and is the most widely used test theory for CAT. Therefore, this work focuses on IRT-based item pools for CAT.

In IRT, it is assumed that a person's latent ability cannot be directly observed. Instead, it is assumed that a person possesses a particular latent ability and processes an item. Only the response to the item can be observed and measured as a manifest variable, not the latent variable itself. The person's response to the item is dependent on their latent ability and the parameters of the item, such as difficulty or discrimination. A person with a stronger latent ability has a higher probability of answering an item of a certain difficulty than a person with a weaker latent ability (Moosbrugger et al., 2020, pp. 260–261).

The IRT is a family of different mathematical measurement models, which are unified by three principles (van der Linden, 2016, p. 8):

> "The first principle is a focus on the responses by human subjects to test items rather than an a priori chosen score on the entire test, as in more traditional test theory. The second is the recognition of the random nature of these responses and the acknowledgment of the need of a probabilistic model to explain their distribution. The third principle is the presence of separate parameters for the effects of the subjects' abilities, skills, or attitudes and the properties of the items. This parameter separation is the ultimate defining characteristic of IRT." (van der Linden, 2016, p. 8)

In general, there are four item parameters in IRT that can be estimated. Every model employs the item difficulty parameter (σ), which determines the difficulty of an item. The other item parameters are the item discrimination parameter (a_i), the guessing parameter (c_i), and the carelessness parameter (d_i). The a_i is a measure of an item's differential capability. A high a_i indicates an item that can effectively differentiate among examinees (An & Yung, 2014, p. 3). The c_i is an item's lower asymptote that represents its guessability. The d_i is an item's upper asymptote that indicates the examinee's carelessness.

All item parameters are described on a logit scale, which is a measurement scale used in IRT. The logit scale is used to measure a person's ability (θ) and an item's difficulty (σ) on a common metric. Based on the logarithmic function, the logit scale transforms the probability of a person getting an item right into a continuous scale ranging from negative infinity to positive infinity. A difference of one unit on the logit scale corresponds to a difference of one standard deviation in the underlying trait being measured. In IRT, the logit scale is used to estimate the item parameters. σ is defined as the point on the logit scale where there is a 50% probability of an individual answering the item correctly. The use of the logit scale in IRT allows for the comparison of item difficulties across different tests and populations and for the estimation of individual abilities on a common metric (Linacre & Wright, 1989).

There are a large number of IRT models. One of them is the 1PL model, also known as the dichotomous Rasch model. This model was mainly defined by Georg Rasch (1960) and can be used for dichotomous answer categories. Dichotomous answer categories are response options in which the examinee can only choose between two possible answers, such as *correct* or *incorrect*. In the Rasch model, a person's likelihood of solving an item depends only on θ and σ (An & Yung, 2014, pp. 1–2). Therefore, σ is the only item parameter estimated within the Rasch model, and all other item parameters are set to be the same for all items. For every item, the expectation of a_i is 1, the expectation of c_i is 0, and the expectation of d_i is 1 (Table 1).

Table 1

Item parameter range of the Rasch model.

Item Parameter	Value Range
a_i	1
σ	$]-\infty; \infty[$
c_i	0
d_i	0

The Rasch model follows several main assumptions (Kelderman, 1984). The assumption of *unidimensionality* (Verhelst, 2001) means, that the test items are measuring a single underlying construct or trait. The assumption of *local independence* means, that the responses to each test item are independent of the responses to other test items (Baghaei, 2008). In other words, the response to one item should not influence the response to another item. The assumption of *sample independence* means, that the responses to the test items are independent of the sample who take the test (Scheiblechner, 2009). In other words, the results obtained from one sample should be similar to the results obtained from another sample:

> "The concept of sample independence allows for the assessment of much broader subject populations and item universes than the classical sample-dependent test criteria. This gain in the precision of concepts of latent dimensions and of generalizeability [*sic!*] of psychological assessment procedures is the true achievement of Rasch models." (Scheiblechner, 2009, p. 188)

2.3.2.1.2 Model Calibration

CATs rely on item parameters, which are used to define the item and to select the best fitting item for each person. In order to obtain those item parameters for a CAT based on an IRT model, the first step involves estimating the item parameters from a sample of examinees (Bock & Gibbons, 2021, p. 141). Therefore, the target group of the test must be selected to obtain reliable information about the items. Examinees from the target group then take the test, which comprises the item pool. Their responses to the test items are used to calibrate the item pool and obtain item parameters for each item. These item parameters are essential input for the CAT.

There are different methods for calibrating an IRT item pool. The three most well-known methods are *joint maximum likelihood estimation* (JML), *conditional maximum likelihood estimation* (CML), and *marginal maximum likelihood estimation* (MML) (Molenaar, 1995). However, there are also alternatives, such as the *pairwise method* of estimation (Heine & Tarnai, 2015; Zwinderman, 1995). Each of these methods is based on the maximum likelihood (ML) method, which is a statistical method used to estimate the parameters of a probability distribution by finding the set of values that maximizes the likelihood function. The likelihood function measures how well a particular set of parameters fits the observed data. Therefore, ML is a way to determine the most likely values of the parameters of a given probability distribution based

on a set of observed data (Myung, 2003). All IRT estimation methods can be used to estimate the item parameters of the item pool. However, the models differ in their handling of person parameters (Molenaar, 1995, pp. 40–41).

JML is one of the earliest methods for calibrating IRT models. This method estimates σ and θ jointly, meaning that they are estimated simultaneously (Bock & Gibbons, 2021, p. 147). However, this approach has several flaws. JML cannot be applied when an examinees provides all correct or incorrect answers or when one item is fully correct or incorrect answered. Additionally, "JML item estimates are known to be biased and inconsistent" (Magis et al., 2017, p. 21), and θ is estimated less precise than σ (Heine, 2016, p. 5).

CML estimates σ and θ separately. Therefore, θ is eliminated by conditioning (Molenaar, 1995, p. 40). Due to this approach, some information for σ estimation can be lost during the estimation process. This effect becomes smaller with increasing item pool size and is almost negligible with item pools of 20 items or more (Eggen, 2000). CML can only be used for Rasch models and, like JML, cannot be applied when all answers of an examinee or of one item are fully correct or incorrect (Magis et al., 2017, p. 21).

MML estimates σ and θ also separately, but does so by integrating out the θ levels (Molenaar, 1995, p. 41). MML is suitable for various unidimensional and multidimensional IRT models and can handle fully correct or incorrect responses. Additionally, MML estimates can be used to compare the fit of different models using Likelihood Ratio tests. However, the use of MML can be problematic with small datasets or selective samples, as the assumption of normal distribution may be violated in these cases (Heine, 2016, p. 5).

An alternative calibration method is the *pairwise* method (Choppin, 1982; Heine & Tarnai, 2015; Rasch, 1960, p. 172). The pairwise method also estimates σ and θ using conditional probability. It estimates the difficulty of a set of items by comparing two or three items at a time, instead of estimating all items at once (Choppin, 1982, p. 1). "To describe the difficulty of a set of items, a matrix is constructed in which each element is the number of people who responded correctly to one item and incorrectly to another item" (Choppin, 1982, p. 1) of the item set. The pairwise method is only applicable to the Rasch model. Like JML and CML, the pairwise method cannot handle fully correct or incorrect responses to an item. However, due to the use of item sets, the pairwise method can be used to estimate θ and σ even with incomplete data (Choppin, 1982; Heine & Tarnai, 2015). Incomplete data may occur when examinees do not take all items of a test.

Accordingly, the choice of the most appropriate model for IRT calibration depends on the type of test and sample being used. Therefore, the selection of a suitable model can only be made based on the data structure of each test and cannot be applied universally to all item pools for CAT. An overview of the usability of the estimators can be seen in Table 2.

Table 2

Item and person parameter estimators of the Item-Response-Theory.

	JML	**CML**	**MML**	**Pairwise**
IRT models	All	Rasch	All	Rasch
Person parameter estimation	Jointly	Separately	Separately	Separately
Usable with fully (in)correct responses	No	No	Yes	No
Usable with incomplete data	No	No	No	Yes
Usable for model comparison	No	No	Yes	No

2.3.2.1.3 Item Pool Size and Distribution

For every IRT test, the size and difficulty distribution of an item pool refers to the range and distribution of difficulty levels of the items included in the pool. In addition to the fact that a CAT item pool should ideally be calibrated to an IRT model, the size and distribution of the item pool are also relevant to its quality and performance.

Item pools for CAT can either be created from scratch based on a blueprint or sourced from existing tests. When creating a CAT and its item pool from scratch on a blueprint, there are various ways to design such an "optimal" item pool. Two well-known methods for this are the *linear programming method* (Veldkamp & van der Linden, 2000, 2010) and the *bin-and-union method* (He & Reckase, 2014; Reckase, 2003). However, item pools generated by both methods show little difference in comparison (Hsu, 2019).

CATs can also be created by using item pools from existing tests (Ebenbeck & Gebhardt, 2022; Mills & Stocking, 1996; Mizumoto et al., 2019; Thompson & Weiss, 2011). A well-designed item pool for CAT should have a sufficient number of items to provide an accurate assessment of a test taker's ability, and the items should be representative of the content and skills being measured (Reckase, 2003). Recommendations for a specific number of items vary. Stocking (1994) recommends an item pool size of 12 times the length of a fixed-length CAT, if the test length should be reduced by 50% compared to a linear test. Way (2005) extends this

recommendation by suggesting that this item pool size is also reasonable for CATs with variable length. However, he notes that such large item pools are especially necessary when the CAT is intended to be administered frequently and practice effects should be avoided.

Practically developed and implemented CATs sometimes use smaller item pools ranging from about 30 to 130 items (e.g. Forkmann et al., 2009, Ludewig et al., 2022, Petersen et al., 2010). Weiss and Kingsbury (1984) note that an item pool with 100 items can provide sufficient results as long as it covers the range of θ in the sample. Wyse and Albano (2015) use an item pool with 167 items for a math assessment for grades three to eight for students with and without disabilities, as this number of items can cover multiple grade levels and performance levels. Ludewig et al. (2021) developed a CAT for reading comprehension in the third and fourth grades in Germany with an item pool size of 132 items and compared three different fixed test lengths (n = 85, n = 25, n = 8). According to the results, the accuracy and reliability of the CAT decreased with shorter test lengths.

In general, a larger item pool may be more effective for CAT, as it can provide a wider range of items to choose from and reduce the risk of biases or measurement errors. The difficulty distribution of the item pool should also be carefully considered because an unbalanced or poorly calibrated item pool may result in inaccurate or unreliable test scores (Ebenbeck & Gebhardt, submitted; Segall, 2005). One option is to use items with a uniformly distributed difficulty, as this ensures a similar number of items are available for each ability level, and in particular, examinees in extreme θ ranges have sufficient items available (Chen et al., 2000).

2.3.2.2 Item Selection

Item selection describes the process of selecting the next item after an examinee has completed an item. Therefore, two steps are necessary: Firstly, the examinee's ability needs to be estimated after each item. Secondly, based on the pre-defined item parameters and the examinee's current estimated ability, the most suitable item needs to be selected.

To estimate the examinee's ability, different ability estimators can be used. Two popularly ability estimators used in this case are *ML* and the family of *Bayesian estimators* (Magis et al., 2017, p. 41). Each estimator uses the examinee's responses to the items they have taken. After each answer, the new response is considered to re-estimate the examinee's ability. As more items are completed, more information about the examinee's ability can be derived. In this way, the estimation of the examinee's ability becomes more accurate with each item taken. For a

detailed mathematically description of each of these estimators in the context of CAT, please refer to van der Linden & Pashley, 2000.

It is necessary to note, that ML estimators are unable to handle fully correct or incorrect response patterns (van der Linden & Pashley, 2000, p. 9). This can be particulary problematic at the start of a CAT. When the first item in answered, the answer is automatically the only response of the test so far. As a result, the first answer is always a fully correct pattern (if the first answer was correct) or an incorrect pattern (if the first answer was incorrect). If the second answer is the same as the first answer (i.e. both answers are either correct or incorrect), the response pattern of the examinee is again fully correct or incorrect. However, as more items are taken, the probability of a fully correct or incorrect response pattern diminishes. This is why ML estimators are more and more reliable as the CAT progresses but may exhibit performance issues in the beginning.

Instead, Bayesian estimators can be utilized at the start of a CAT until the examinee no longer displays a fully correct or incorrect response pattern (Wang & Vispoel, 1998, p. 111). Two frequently used Bayesian estimators in CAT are the *maximum a posteriori estimator* (MAP) and the *expected a posteriori estimator* (EAP). One characteristic of Bayesian estimators is that they "incorporate prior information into the data in deriving ability estimates" (Wang & Vispoel, 1998, p. 110):

> "Initially, it is typically assumed that the population ability distribution is normally distributed (…). This initial assumed ability distribution is called the prior distribution. After the examinee answers the first item, the likelihood associated with the response is combined with the information about the prior ability distribution to create an adjusted ability distribution called the posterior distribution. This posterior distribution then becomes the prior distribution to be combined with the likelihood associated with the examinee's response to the second item." (Wang & Vispoel, 1998, p. 112)

It is not clear which method of ability estimation is best for CAT in general. Instead, based on the advantages and disadvantages of each estimation method, it must be decided for each CAT which method is the most appropriate. An overview of the advantages and disadvantages in the context of CAT is provided by Wang and Vispoel (1998, p. 131) based on various simulation studies: ML estimates particulary high or low abilities less biased than EAP and MAP. However, it also estimates those high or low abilities with a higher *SE* when the item pool lacks items at those corresponding difficulty levels. When compared to MAP and EAP, it results in

the longest test length for *SE*-based test terminations (Wang & Vispoel, 1998, p. 131). Also, ML is not usable on its own at the start of a CAT as it cannot work with fully correct or incorrect result patterns.

EAP and MAP both estimate abilities with lower *SE*s than ML. In comparison, EAP leads to the lowest *SE*s, but MAP requires the shortest test length for *SE*-based test terminations. Both estimators also exhibit bias in estimation for low and high abilities levels (Wang & Vispoel, 1998, p. 131). The use of EAP and MAP is not optimal, as the assumption of a normal distribution influences ability estimation, especially in the beginning of CAT. However, with growing test length (e.g. of more than 20 items), the influence and bias resulting from those prior distributions decrease (van der Linden & Pashley, 2000, p. 9).

Since all methods have their flaws, it is also possible to combine ML and a Bayesian estimator for CAT. In those cases, CAT ability estimation starts with a Bayesian estimator and switches to ML, when the response pattern is no longer fully correct or incorrect (Magis et al., 2017, p. 41).

After the examinee's ability is estimated after every answer, a suitable item has to be selected afterwards. To date, 14 different item selection methods have been introduced for CAT (For a detailed mathematical description and overview, please see van der Linden & Pashley, 2010, pp. 11–27, Magis et al., 2017, pp. 42–47 and Bock & Gibbons, 2021, pp. 245–257). However, a comparative simulation study by Chen et al. (2000) found that the most commonly used item selection methods showed nearly no precision advantage.

The mostly used rule for item selection is the maximum Fisher information (MFI). MFI is a standard mathematical rule of ML. Lord (1980) introduced the use of the MFI to select items in CAT in order to get the most efficient measurement. Under MFI, a set of eligible items is chosen based on the estimated examinee's ability after an answer. Among the set of eligible items, the most informative item for the current estimated ability is selected (Bock & Gibbons, 2021, pp. 247–248; Magis et al., 2017, p. 43). The examinee takes this selected item next.

2.3.2.3 Test Starting

A CAT starts with the selection of the first item (Magis & Raîche, 2012). The first item of a CAT can be selected from the item pool in various ways. Usually, a fixed first item is selected, which is chosen for every person at the beginning of the CAT. This first item is often selected based on its σ. Usually, an item with $\sigma = 0$ is selected as the scale in IRT is centered on the

examinees θ. An item with σ = 0 therefore can be seen as the most informative item when no further information about the examinee is available.

Another possibility is to incorporate existing information about the examinee. If information about the examinee is already available at the beginning of the test, this information can be used to select a first item. "For instance, knowing from previous tests that the examinee has rather high or low ability level, one can adjust the initial ability level to values larger or smaller than the average prior ability level, respectively" (Magis & Raîche, 2012, p. 8). If a test is conducted more frequently in the context of formative assessment, the result of the first test provides an ideal starting point for selecting the first item for the next test (Thompson & Weiss, 2011). Adapting the first item to the pre-estimated ability of an examinee does not affect the results and accuracy of the CAT. Instead, it offers the possibility of a shorter test with fewer items since the examinee's ability can be reached more quickly (Weiss & Kingsbury, 1984).

2.3.2.4 Test Termination

When developing a CAT, it is necessary to define the conditions under which the test should end. Test termination rules must be established for every test, as there is no general rule about which criterion is the best, but rather, the performance depends on the type and purpose of the test. One possibility are *fixed-length CATs*, which establish a fixed test length, i.e., a set number of items that every examinee must answer. Afterward, the examinee's ability is estimated. Another possibility are *variable-length CATs*, which do not have a fixed test length. Instead, another criterion for stopping the CAT algorithm is used (Thompson & Weiss, 2011). Once the CAT is stopped, the examinee's ability is estimated. Fixed-length CATs do not provide any psychometric advantages over variable-length CATs. Additionally, variable-length CATs do not result in biased testing, as earlier studies had suggested (Babcock & Weiss, 2009).

There are multiple stopping rules for variable-length CATs. One often considered rule is the *precision criterion* (Magis et al., 2017, pp. 47–48): The precision criterion stops the test when the ability can be estimated with a predetermined level of accuracy, based on a *SE* that must be equal to or smaller than the target *SE*. Michiel et al. (2008) compared maximum $SE(\theta) < 0.3$, < 0.4, < 0.5, < 0.6, < 0.7, and < 0.8. Using $SE(\theta) < 0.3$ led to an average test length reduction of about 22% (8 items less). Using $SE(\theta) < 0.4$ led to an average test length reduction of about 67% (25 items less). With $SE(\theta) < 0.5$, the average test length reduction was 81% (29 items less). The measurement accuracy decreased with increasing *SE* ($SE(\theta) < 0.3$: $r = .996$, $SE(\theta) < 0.4$: $r = .949$, $SE(\theta) < 0.5$: $r = .895$). Similar results were also found by Ebenbeck and Gebhardt

(2022), who also compared *SE* values of 0.3, 0.4, and 0.5, and demonstrated a reduction in test length and measurement accuracy for students with and without SEN as *SE* increased.

Another stopping rule that can be used for variable-length CATs is the *information criterion*. The information criterion stops the test when the remaining items in the item pool do not have enough information to further estimate the examinee's ability. To prevent the administration of items that would only result in a longer test without providing additional information, a minimum information value is defined for each item before it can be selected. It is also possible to combine multiple stopping rules, such as a length and precision rule, to customize the test to its intended purpose. Stafford et al. (2019) conducted a comparison of the precision criterion with two other variable-length stopping rules, namely the information criterion and the expected change in θ. The study found that the precision criterion offered the best balance between measurement accuracy and test length, resulting in the shortest tests. Moreover, the addition of a maximum number of items in the form of a fixed-length CAT further improved the performance of the CAT. Similarly, Dodd et al. (1993) demonstrated that the precision criterion leads to shorter and more accurate tests compared to the information criterion. This finding was consistent across item pools containing 30 or 60 items and across various CAT settings.

2.3.2.5 Response Pattern and Test Scoring

In a CAT, students do not answer every item in the item pool. Instead, they are presented with a selection of items that match their estimated ability level. While this approach shortens the test length compared to non-adaptive tests, it can pose challenges when it comes to scoring the test (Wang & Kolen, 2001):

In non-adaptive tests, the maximum score is determined by the number of items. For example, in a basic Rasch model test, every correct answer would give one credit, and the score would be the number of correct answers. More complex non-adaptive tests could assign multiple credits per item. However, in a CAT, the number of correct answers cannot be used for scoring. Firstly, due to the stopping rule, students could have different test lengths and therefore different maximum numbers of items to answer, leading to different maximum scores between students. Secondly, students can receive a different selection of items with varying difficulty distributions. Even if students answered the same number of items correctly, their difficulty level could differ, making the results not comparable. Hence, other scoring methods are used in CAT, with proficiency estimates (i.e., θ values) being the most commonly used method. These

estimates are calculated after every answer and again after the final answer (Stone & Davey, 2011, 4).

For the same reasons, sum scores cannot be used to compare results from non-adaptive and adaptive tests. Instead, the percentage of correctly solved items can be used, although this method also does not take into account the difficulty of each task (Ling et al., 2017; Vispoel et al., 1994). In their study, Vispoel et al. (1994) showed that the percentage of correctly solved items did not significantly differ for each examinee between a non-adaptive test and the same test as an adaptive test, despite the tasks being of varying difficulty. The average proportion of correctly solved items was about 80%. Ling et al. (2017) found similar results. The proportion of correctly solved items did not differ significantly between adaptive and non-adaptive tests, with both having a proportion of about 50%. Students with higher abilities also proportionally solved more items correctly in the adaptive test compared to students with lower abilities. Helwig et al. (2002) also demonstrated that students without SEN performed significantly better on both an adaptive and a non-adaptive test than students with SEN. The results between the adaptive and non-adaptive tests did not differ significantly.

In summary, CAT is a procedure with a lot of potential for DBDM in schools. This is partly due to the fact that CAT simplifies and shortens the implementation and evaluation process for both teachers and students. With CAT, it is likely that student groups with heterogeneous performance distributions can be measured with the same test. However, there is a research gap in the area of CAT for students with SEN. There are few empirical studies that have included this group of students in their samples. Most of the findings are of a methodological nature and do not relate to specific application contexts or groups of people. Furthermore, there is no literature on the response patterns of students with and without SEN and their comparison between different types of tests. This is problematic, as students in the lower performance spectrum and with SEN could benefit greatly from CAT.

To partially address this research gap, three studies are conducted in the empirical part of this study. These studies focus on the development and simulation of multiple CATs based on data from students with and without SEN. The performance of the CATs and the simulated students themselves, as well as their response patterns and response behavior, are evaluated. The methodology of the studies is presented in the next chapter.

3 Empirical Research Process

CAT has the potential to improve the assessment of struggling students or those with SEN. This study aims to investigate the development and performance of CAT for students with different performance levels through simulation studies. The chapter is structured as follows: First, the instruments used in the three studies will be described. Second, the sample, sample selection, and how they will be used throughout the studies will be explained. Third, the designs and procedures of the three simulation studies will be described in detail, starting with study 1 and ending with study 3. The different simulation runs will be addressed, and the simulation processes will be explained and illustrated. The research questions for each study will also be explained.

All research questions will use an existing reading screening as an example and generated data. Reading is considered an exemplary academic performance area since it is a fundamental skill required for success in many academic subjects and in life outside of the classroom. Therefore, it is important to assess and improve reading skills for students with low abilities or SEN, who may face additional challenges in developing this skill. Additionally, reading is a skill that is included in the curriculum for both students with and without SEN, ensuring comparability. Finally, the choice to focus on reading as an exemplary academic performance area was made with the understanding that the methodology and results of the study can be applied to other academic subjects. While reading is the primary focus of this work, the findings of the studies are relevant to other skills and subjects.

3.1 Instruments

To develop an adaptive test, a German modular reading screening test comprising four subtests was used. These subtests measure different subareas of reading that are relevant to reading intervention, including phonological awareness (subtest 1), security of lexical recall (also known as "vocabulary," subtest 2), speed of lexical recall (also known as "flash reading," subtest 3), and sentence comprehension (subtest 4). Initially, the screening was developed as an analog paper-pencil test (Ebenbeck et al., 2022). In a second step, it was modified and transformed into a digital web-based version that can be used on a tablet or computer (Ebenbeck et al., 2023).

Subtest 1 evaluates the level of phonological awareness as a proximal ability that students have developed. Phonological awareness is a crucial aspect of phonological information processing (Wagner & Torgesen, 1987), which is essential for reading development. Each item in this subtest displays a pictogram representing a word from the basic vocabulary, along with a graphemic representation of a sound. In this task, students are required to identify whether the corresponding sound appears at the beginning, middle, end, or not at all in the word depicted in the pictogram (Figure A1). The subtest consists of 35 items, and students are given a time limit of five minutes to complete them.

Subtest 2 assesses the reliability of students' ability to recall knowledge about words from their mental lexicon. The reliability of lexical recall is demonstrated by students' ability to access their word knowledge. Insufficient vocabulary can have an impact on students' reading acquisition and later reading comprehension (Röthlisberger et al., 2021). Reading a text fluently and understanding its meaning is only possible if the reader knows a sufficient number of words. The yes/no method (Richter et al., 2012; Trautwein & Schroeder, 2019) is used to measure this skill in subtest 2. In this method, students read a word and then decide whether it is a real word or a pseudoword (Figure A2). The subtest consists of 52 items, with 50% being real words and 50% pseudowords. Students are given a time limit of five minutes to complete the subtest.

In addition to the security of lexical recall, the speed of lexical recall from the mental lexicon is also an important factor for reading speed (Ennemoser et al., 2013). To achieve high reading speed, it is necessary that words are recognized directly as whole words rather than being recoded through their letter structure (Coltheart et al., 2001). Therefore, subtest 3 presents words briefly on the screen, with a display duration between 0.5 and 2 seconds, to measure the speed of lexical recall. After the display time, students select which of four possible words they have read. The subtest comprises 30 items, and there is no time limit to complete it (Figure A3).

Subtest 4 assesses sentence comprehension through a gap-filling task. Each item consists of a sentence with a gap that can be filled by one of four answer choices. The gap may be caused by various omissions in the sentence. The item pool for this subtest includes 75 items, and students have a time limit of five minutes to complete them. However, for the development of the adaptive test, only the first 35 items of subtest 4 were evaluated because at least 25% of the students were able to solve this number of items (Figure A4).

The reading screening was specifically developed for use in inclusive classes and heterogeneous learning groups to account for the performance heterogeneity of students and focus on those

with low reading abilities. The target group for this screening is students in second to fourth grade at inclusive primary schools and special schools. As all subtests are competency-based, they can be administered as soon as a student has acquired the necessary prerequisites. For subtest 1, students should have completed letter acquisition or phoneme-grapheme correspondence. For the other three subtests of the screening, at least synthetic reading of words or sentences is required. Synthetic reading involves recoding words "letter by letter" and is not yet done with whole words (Coltheart et al., 2001).

The screening can be administered digitally through a tablet or compute using the website www.levumi.de, with the use of a tablet being recommended. All subtests are designed to be simple to navigate, with no special motor skills required such as drag and drop. All subtests are presented in a single-choice format.

Subtests 1, 2, and 4 are *speeded tests* (Gulliksen, 1950; Pomplun et al., 2002), with a maximum completion time of five minutes. Speeded tests have varying levels of difficulty, allowing for the measurement of a wide range of abilities, and also to enable student performance to be measured based on the number of items attempted within the time limit. The use of a time limit can be helpful for both teachers and technical considerations. It allows teachers to better plan the duration of the assessment and ensures that the test does not take up too much time in the classroom. It also allows for more efficient use of technical resources, such as tablets or computers, as the test can be completed in a set amount of time. It is important to note that the time limit is not intended to create pressure for measurement, but rather to allow for easier planning of the assessment and test run and to prevent low-performing students from becoming overwhelmed by a lengthy test.

Subtest 3, on the other hand, has no maximum processing time, and the test ends once all 30 words have been presented. As a result, this subtest is considered a *power test* (Kline, 2015), with varying item difficulties used to measure a range of abilities. The number of items worked, therefore, is not an indicator of performance. The total processing time for the screening is between 15 and 20 minutes.

It has been demonstrated that each subtest of the analog paper-pencil test version fits the unidimensional Rasch model (Jungjohann et al., submitted). This allows for flexible use in everyday teaching, as all four tests can be administered consecutively, or each subtest can be administered separately and independently of the others. Each subtest is evaluated individually. Therefore, each subtest has a single score, but there is no total score for the entire screening.

3.2 Sample Description

In an inclusive class, there can be a high degree of heterogeneity in abilities, such as reading. To reflect this heterogeneity in simulation studies, different grade levels and school types were included in the sample. Between November 2022 and January 2023, 400 German students were tested using the digital reading screening. The sample included 357 students from grades 2 (n = 123), 3 (n = 135), and 4 (n = 99) at inclusive primary schools. The students' ages ranged from 6 to 11 years (M = 8.43, SD = 1.07, 16 NA), and they were in their first to fifth school attendance year (M = 3.06, SD = 0.91). Of these students, 31 had an identified SEN-L, 12 had an identified SEN-S and 12 had an identified dyslexia (for an explanation of the view and definition on these SEN in Germany, please see 2.1).

To include students with SEN-I in the sample, additional 43 students from private special schools with a focus on intellectual development were included. To ensure that students had at least completed letter acquisition, teachers selected students from their classes in advance, but it was not possible to verify the accuracy of their assessments. At these schools, students are typically only divided into grade levels in primary school, and in middle school, students of similar ages are placed together in one class, which may not necessarily correspond to their grade level. Of the 43 students, 13 are in primary school (6 in grade 3 and 7 in grade 4), and 30 are in secondary school. The students range in age from 8 to 15 years old (M = 11.57, SD = 1.91) and have attended school for 2 to 9 years (M = 6.03, SD = 2.24). The class teachers provided information on the severity of the disability, with 19 students having a mild intellectual disability, 12 students having a moderate intellectual disability, and 4 students having an IQ value in the range of BI. Since these students still require support for intellectual development, they were evaluated along with the students with intellectual disabilities. Nine of the students did not complete subtest 4.

The total sample consists of 400 students, with 78.5% without SEN, 7.75% with SEN-L, 10.75% with SEN-I and 3% with SEN-S. Some students have multiple disabilities, such as a behavioral disorder in addition to SEN-L, but these additional disabilities are not evaluated separately in this sample. Of the total sample, 42.75% are female, 34.75% are male, and for 22.5% no information about gender was available.

3.3 Research Questions and Procedures

Three studies are conducted successively in order to further develop the digital reading screening into an adaptive one and to compare its performance towards a generated item pool for different student groups with and without SEN. For all statistical data analyses, the free programming language R (R Core Team, 2022) is used with the programming interface RStudio (RStudio Team, 2020). For data cleaning, wrangling, reshaping, and visualization, the open-source package collection *tidyverse* (Wickham et al., 2019) is used. The syntax of the analyses is available on the Open Science Framework (Ebenbeck, 2023).

3.3.1 Study 1: Simulating Screening SubCATs of Inclusive Student Groups

3.3.1.1 Research Questions

Study 1 investigates how to set up an adaptive test based on the item pool of the reading screening. To achieve this, four research questions are explored:

Are the item pools of the reading screening suitable for CAT?

The effectiveness of a CAT largely depends on the quality of the item pool that is used. If the item pool is unsuitable for CAT, then the adaptive testing approach may produce inaccurate results or fail to identify individuals who require further intervention or support. To mitigate this issue, it is necessary to evaluate the suitability of the item pool. This involves ensuring that the items meet the necessary psychometric criteria of the Rasch model, such as reliability and unidimensionality. Also the items should be checked for fairness and coverage of a range of difficulty levels, from easy to difficult, to provide sufficient measurement precision and variability. The difficulty values of each item are also needed to set up the CAT algorithm later. Therefore, assessing the suitability of the item pool is critical for ensuring that the adaptive testing process is effective and can accurately measure individual differences in reading ability.

Does the inclusive sample show sufficient performance heterogeneity?

Simulation studies will be conducted to simulate the performance of a CAT in an inclusive setting. Assuming that inclusive learning groups are characterized by their performance heterogeneity, it is important to ensure that the input data for the simulation are also as heterogeneous as possible in their performance. Insuring the heterogeneity of the sample is also crucial to

ensure that the simulation results are generalizable to all students, regardless of their disability status.

> *Which stopping rule is suitable for test length reduction? Can the accuracy of testing be maintained despite shorter test runs?*

Adaptive testing tailors the test to the individual's ability level, potentially shortening test length while maintaining accuracy. However, a suitable stopping rule must be identified to end the test early without compromising accuracy. A weak stopping rule may prolong testing, causing student fatigue and decreased motivation and negatively affecting accuracy. Conversely, a stronger stopping rule may lead to inaccurate measurements. Shortening test length is also crucial in schools with limited time and resources for diagnostics.

> *What influence does the size and distribution of the item pool have on test length reduction and measurement accuracy for different groups of students?*

The influence of item pool size and distribution on test length reduction and measurement accuracy is an important consideration in CAT. A larger item pool has been shown to lead to more accurate measurement, provided that the difficulty distribution is wide enough (Ebenbeck & Gebhardt, submitted). However, if the item difficulty is clustered within a narrow range, it may limit the adaptive range of the test and result in less accurate measurement. Conversely, a wider distribution of item difficulty may allow for a more precise estimation of an examinee's ability and result in more efficient testing. Understanding the influence of item pool size and distribution on test length reduction and measurement accuracy is crucial in designing and implementing effective CATs, especially for different groups of students with SEN.

3.3.1.2 Research Process

To assess the performance heterogeneity of students in the sample, the number of attempted items ("attempts"), the number of correctly solved items ("sum score"), and the number of incorrectly solved items ("error score") for each student are considered. The maximum, minimum, and average values for each subtest are calculated.

For the psychometric analysis of the subtests, the R package *pairwise* (Heine, 2022) was used. This package provides functions for analyzing the one-dimensional Rasch model and uses the pairwise method to calculate σ and θ. The pairwise method is chosen because some students could not answer all items within the given time, which resulted in missing values in the data

set. The pairwise method is capable of handling missing values and calculating σ and θ accurately, nonetheless.

To assess item fairness, we compared the fairness of items for each subtest using a graphical model test that employed the *Andersen Likelihood Ratio test* (Andersen, 1973) with median and random split. The Andersen Likelihood Ratio test is a goodness-of-fit test that evaluates whether the observed responses to items fit the Rasch model. The log-likelihood ratio of the observed data is compared to the log-likelihood ratio of the expected data under the Rasch model. If the difference between these two log-likelihood ratios is statistically significant, it indicates that the observed data does not fit the Rasch model well. The median split has been shown to have better statistical power than the random split and is therefore recommended in the literature (Krammer, 2018). However, the random split can provide additional information if it is not used as the only split criterion. Misfitting items, which do not conform to the Rasch model's assumptions, can have too few or too many response categories, display differential item functioning across different groups, or exhibit unexpected response patterns. By identifying misfitting items, one can revise or remove them to improve the overall fit of the Rasch model to the data. Conspicuous items are removed as appropriate.

After removing any misfitting items, a one-dimensional Rasch model for each subtest is computed and σ and θ for each subtest are extracted. In the Rasch model, σ are typically expressed on a logit scale and take values between -∞ and +∞. Negative values indicate easier items and positive values indicate harder items. Additionally, the *mean squared residual based* (MSQ) *infit* and *outfit statistics* are used to assess the fit of items or persons to the Rasch model. The infit statistic is a measure of how well the response of a person or an item fits the model at a particular point on the measurement continuum. It is calculated by comparing the expected response to the observed response, and the squared difference is weighted by the estimated variance of the item or person. The infit statistic is sensitive to unexpected responses that occur close to the expected value. The outfit statistic is also a measure of how well the response of a person, or an item fits the model, but it is less sensitive to unexpected responses that occur close to the expected value. Instead, it is more sensitive to unexpected responses that occur at the extreme ends of the measurement continuum (Linacre, 2002).

Optimal values for MSQ infit and outfit statistics fall between 0.5 and 1.5. Values up to 2.0 are generally acceptable and do not compromise the measurement system. However, values

exceeding 2.0 can distort or degrade the measurement system, while values below 0.5 are less productive for measurement but do not cause degradation (Linacre, 2002).

Additionally, person-item maps, also known as Wright Maps, are generated for each subtest to show the relationship between θ and σ in the Rasch model (Lunz, 2010). In these maps, θ are plotted on the horizontal axis, while σ are plotted on the vertical axis. Each person is represented by a dot, and each item is represented by a horizontal line or a series of adjacent points. The position of each item on the map represents its difficulty level, with easier items located closer to the left-hand side and more difficult items located closer to the right-hand side. The position of each person on the map represents their θ level, with more able persons located higher on the map and less able persons located lower on the map. The person-item maps are also used to identify gaps in the coverage of the skill being measured by the item pool.

The subtests measure θ to some extent, but they have a ceiling effect since the contents of the subtests focus on the lower ability range. Particularly in the first three subtests, a high percentage of students are expected to solve all or almost all items correctly. Therefore, their actual θ may be assessed as lower than they really are. To avoid this ceiling effect, in the next step, the students' θ are generated (n=1000) based on the true *M* and *SD* without the ceiling effect.

With the calculated σ and the generated θ per subtest, adaptive testing is simulated for each subtest ("subCAT") using the R package *catR* (Magis & Barrada, 2017). This package is useful for generating response patterns and CATs based on one-dimensional IRT-based item pools. To set up a CAT algorithm for each subtest, four algorithm steps must be defined with *catR*: the Initial Step, the Test Step, the Stopping Step, and the Final Step. In the Initial Step, the first item is selected. In the following Test Step, items are selected based on the student's answers and θ estimates are updated after each answer. If the given stopping rule is satisfied, the item administration ends in the Stopping Step. Finally, in the Final Step, the student's final θ estimate is calculated (Figure B1).

The subCATs are set up with the following settings: The first item of each subCAT is one of average difficulty of all items, which is the common way to proceed (Weiss, 1985; Magis and Raîche, 2012), since an item with average difficulty is most meaningful at the beginning. The first item is pre-set to ensure the same starting item for each test run. For ability estimation in the Test Step and Final Step, a recommended procedure of Magis et al. (2017) is used to avoid the problem of infinite estimators for fully correctly or incorrectly answered test items. That way, the subsequent item selection within a test run is performed using MFI, which is also the

most commonly used method of item selection (Barrada et al., 2009), in combination with the Bayesian modal estimator (Birnbaum, 1969) and a normal distributed prior distribution. The final ability is estimated using ML. This combination of CAT settings has already been successfully proven for samples including students with SEN (Ebenbeck & Gebhardt, 2022; Ebenbeck & Gebhardt, submitted).

In order to simulate subCATs that measure with a certain level of accuracy, three different stopping criteria are used for each subtest simulation. Specifically, the simulations use three different *SE* values as stopping criteria ($SE(\theta) = 0.3$, $SE(\theta) = 0.4$, $SE(\theta) = 0.5$). The *SE* value represents the level of precision with which θ can be estimated by the CAT. Terminating the CAT with $SE(\theta) = 0.3$ results in the most accurate measurement of θ compared to termination with $SE(\theta) = 0.4$ or $SE(\theta) = 0.5$. However, a smaller *SE* also means a longer test length. Since the initial item pools of the non-adaptive screening are relatively small, the CAT measures more efficiently than the non-adaptive screening if it can achieve an average test length shorter than the number of items in the screening, using any of the three *SE* stopping criteria.

The resulting test length, test accuracy (in terms of the correlation between true θ and estimated θ), and the percentage of successful test discontinuations due to the discontinuation criteria are evaluated and compared. The resulting test lengths are compared to the average and maximum achieved test lengths of the non-adaptive screening to determine whether CAT is a suitable extension for the subtest.

To evaluate the performance of the CAT in extreme ranges, three additional individual simulations are conducted for one person each with θ values of -1, -3, and 3, and their results are compared. These individual simulations reveal that the screening item pools have a higher number of missing easy items. To eliminate the possibility that the study's results are solely due to the narrowness of the screening item pools' σ, another item pool ($n = 100$) is generated with a range of $-3 < \sigma < 3$ for further simulations. This larger and more widely distributed item pool allows for a more comprehensive and accurate measurement of students' θ.

Using this generated item pool and the stopping rule $SE(\theta) = 0.5$, along with the other CAT settings from previous simulations, four additional simulations of 1000 test runs each are performed, one for each subtest, and compared to the screening item pools in terms of test length, accuracy, and the percentage of successfully terminated test runs (Figure B2).

3.3.2 Study 2: Simulating a Screening CAT of Inclusive Student Groups

3.3.2.1 Research Questions

In the previous study, each subCAT started with a fixed item, which provided the CAT algorithm with the initial information about the examinee. As more information was obtained over the course of the test, the estimation of the examinee's θ became more accurate. However, in a screening, students typically complete multiple subtests in succession. When starting a new subtest, there may already be information available from previous tests, which could potentially reduce the number of items needed to obtain information about the student's θ. Therefore, the hypothesis is that using this previous information can lead to a more accurate and shorter measurement in the subsequent subCATs. This approach involves combining the subCATs into one screening CAT. Study 2 investigates whether connecting different subCATs into one large screening CAT leads to a more efficient measurement. To accomplish this, three research questions are explored:

> *Does linking the subCATs through a modified starting rule lead to a reduction in test length? Does this linking have an impact on the accuracy of the measurement?*

This question aims to determine whether linking the subCATs and utilizing information from previous subtests can enable the CAT algorithm to adapt more quickly and accurately to the student's θ level, potentially resulting in a more efficient and accurate measurement. This involves changing the start rule of the following test, where no fixed item is used and the estimated θ of the person from the previous test is used as input. This could lead to not only a shorter and more accurate measurement, but also a more convenient implementation, as only one screening test needs to be started in a session instead of four subtests.

> *Does the size and distribution of the item pool have an impact on whether the linking of the subCATs is effective?*

The size and distribution of the item pool can have an impact on the effectiveness of linking subCATs in a screening CAT. If the item pool is too small, there may not be enough items to accurately measure the true θ of all students. Additionally, if the item pool is not well-distributed in terms of σ, it may lead to inaccurate measurement. If a student's θ was measured inaccurately in the subCAT before, it could lead to even more inaccurate measurement in the linked subCAT. Therefore, the influence of item pool size and σ distribution on the effectiveness of

linking subCATs needs to be investigated in order to optimize the design and implementation of screening CATs.

> *Does the correlation of the subtests have an impact on whether the linking of the sub-CATs is effective?*

The linking of subCATs is based on the assumption that the θ level of the examinee in the previous test provides relevant information for the subsequent subCAT. However, this assumption is likely only valid if there is a certain degree of correlation between the examinee's θ levels across different subtests. Therefore, analyzing the correlation of θ levels between subtests can offer valuable insights into the effectiveness of the linkage.

3.3.2.2 Research Process

To answer these questions, the total scores of the reading subtests in the screening test are correlated with each other using the *Pearson product-moment correlation coefficient* (Pearson, 1920). The Pearson product-moment correlation coefficient is a statistical measure that evaluates the linear relationship between two continuous variables. It measures the degree of association or relationship between the two variables. The Pearson correlation coefficient ranges from -1 to +1, where -1 represents a perfect negative correlation, +1 represents a perfect positive correlation, and 0 represents no correlation.

Next, six different simulations of screening CATs are performed using the R package *catR*. Therefore, the CAT algorithm settings that were checked in the previous study are used: the θ after each answer is estimated via Bayesian Modal Estimator and the next item is selected using MFI. The final θ is estimated using ML. Each test run is terminated after $SE(\theta) = 0.5$ is achieved.

Because a connected screening CAT is being estimated, it is not possible to use the generated θ as in study 1. With the generated θ, it is not possible to track how a person performs across the various subtests and how θ are related. To maintain these relationships in the simulation, real θ are used instead, and they are not adjusted for the ceiling effect.

The simulations differ in their data input and start rule. Simulation 1 uses the screening test's item pools and real θ as input. The subCATs are not linked and use a fixed start item, as in study 1. Simulation 2 also uses the screening test's item pools and real θ as input but links the subCATs (Figure B3). Therefore, subCAT 1 has a fixed start item, while the following subCATs use the previously estimated θ as input to select an appropriate first item.

To ensure that the correlations between the subtests are not too low for connecting, simulations 3 and 4 use perfectly correlated θ (Figure B4). This means that people are simulated to have identical θ in every subtest, for example, θ = -1 in every subtest. In simulation 3, the subCATs are not connected, while in simulation 4, they are connected.

To investigate whether the size of the item pool and the distribution of σ are crucial for a successful connection of subCATs, simulation 5 and 6 use the item pool generated in study 1 as the basis for each subCAT in the simulated screening (Figure B5). In simulation 5, the subCATs are not connected, while in simulation 6, they are connected.

All simulations are evaluated in terms of their average test length, the correlation between the true and estimated θ, and the proportion of successfully stopped test runs. Additionally, the difference between a connected and unconnected CAT for real and fixed θ is demonstrated and analyzed in individual test runs to identify characteristics.

3.3.3 Study 3: Comparing Students' Simulated CAT Performance

3.3.3.1 Research Questions

Study 3 investigates how students with SEN perform on different kinds of CATs and how their performance differs from that of students without SEN. To achieve this, the following research questions are explored:

> *Do students with and without SEN perform differently on the reading screening that serves as the basis for the simulation of screening CATs?*

To determine whether students with and without SEN perform differently on the screening is crucial when simulating those student groups. If the θ distribution of students with and without SEN differs, it is more sensible to simulate these groups separately. If the abilities of student groups, such as those with and without SEN, do not differ, a combined evaluation and simulation are more appropriate.

> *To what extent does the test length and accuracy of a CAT change compared to a non-adaptive test for students with and without SEN?*

The effectiveness of CAT is shown by a shorter test duration with the same accuracy compared to the same test without an adaptive algorithm. Accordingly, a comparison of these two test

formats shows whether CAT leads to an increase in efficiency. If the results show that CAT leads to a shorter test length without compromising measurement accuracy, it may be a more efficient and less burdensome testing method for students with SEN who have difficulty with longer tests. In addition, if CAT proves to be as accurate or more accurate than non-adaptive tests, it may provide a more equitable assessment for students with SEN who have difficulty with certain types of test items or formats.

Does the size and distribution of the item pool affect whether students with and without SEN perform differently on a CAT?

From study 2, it is already known that differently sized and distributed item pools lead to different results when used as input for CAT simulations. This study aims to investigate the extent to which this affects the performance of the test for students with and without SEN. If certain item pools are found to disadvantage students with SEN, it may be necessary to modify the item pool to ensure that the CAT accurately measures the ability of all students.

Does the use of an easier first item affect the accuracy and test length of an adaptive test for students with and without SEN? Does the size and distribution of the item pool play a role?

In study 2, it was found that connecting the subCATs based on the start rule did not add any value to the measurement. To accommodate weak students in a CAT in a different way, the added value of a modified first fixed item will be investigated. It is assumed that an easier start item could help to estimate the ability of weak students faster.

Does the ratio and number of correct and incorrect answers in the CAT differ from a non-adaptive test for students with SEN? Additionally, does the size and distribution of the item pool have an impact on this comparison?

In the previous studies of this thesis, it was shown that low-achieving students often answer a large number of items incorrectly, resulting in a high error rate per test run. This study aims to investigate this issue further. Specifically, it will examine whether the size and distribution of the item pool have an impact on the number of items solved correctly and incorrectly. If there is a significant influence of the item pool, it may be necessary to consider adapting the item pool to ensure more accurate and equitable measurement of θ.

3.3.3.2 Research Process

To better assess students' reading performance, the Pearson product-moment correlation coefficient is first used to correlate the covariates of year of school attendance, grade level, and age of students. When factors are highly correlated with each other, they provide similar information. Therefore, it is often unnecessary to evaluate all factors in such cases. Instead, the variable that is present for the most students can be used. The grade level is then selected as this differentiating factor.

A *one-way Analysis of Variance* (ANOVA, Fisher, 1921) is used to examine whether students' reading performance differs between grade levels for each subtest. An ANOVA is a statistical method used to compare the means of three or more groups to determine if there are significant differences between them. The ANOVA test calculates an *F*-statistic, which is the ratio of the variance between groups to the variance within groups. If the *F*-statistic is greater than the critical value for a given level of significance, then it is concluded that there is a significant difference between at least two of the groups. If there are significant differences, it can be assumed that the reading screening is able to differentiate between groups with different reading performances.

The next step is to examine whether students with SEN scored differently on the four subtests of the reading screening than students without SEN using an *independent samples t-test* for each subtest (Student, 1908).

A *t*-test, like an ANOVA, is a statistical method used to compare the means of groups. While an ANOVA compares three or more groups, a *t*-test compares the means of two groups and determines if there is a significant difference between them. The test is based on the *t*-distribution, which is a probability distribution that describes the likelihood of obtaining a particular sample mean given the population mean and standard deviation. The *t*-test calculates a *t*-statistic, which measures the difference between the sample means and takes into account the variability of the data within each group.

Further, the unidimensional Rasch model is calculated for each subtest, as in study 1. The estimated θ from the Rasch model are then divided based on SEN status. *M* and *SD* of the subtests for each group are used to generate θ without ceiling effects for the simulations

To evaluate the performance of adaptive tests for students with and without SEN, twelve different simulations are conducted. The first two simulations simulate 1000 students with SEN

(simulation 1) and without SEN (simulation 2) taking the reading screening as a non-adaptive test, meaning they answer all items in the four item pools. No time limit is simulated, and the four subtests are worked on until all items have been answered, representing a simulation of pure power tests.

In the third and fourth simulations, 1000 students each with (simulation 3) and without SEN (simulation 4) are simulated performing the adaptive reading screening as in study 1. The sub-CATs each start with a moderately difficult item. The same procedure is followed in the fifth and sixth simulations. First, 1000 students with (simulation 5), then 1000 students without SEN (simulation 6) are run. In contrast to the previous two simulations, however, the subCATs now start with an easy item ($\sigma = -1$).

Simulations 7 to 12 have the same settings as simulations 1 to 6. However, instead of using the real thetas of the reading screening subtests, a generated item pool with 100 items ($-3 < \sigma < 3$) is used. Accordingly, simulations 7 and 8 each simulate a non-adaptive test with a full item pool for 1000 students each with and without SEN. Simulations 9 and 10 simulate an adaptive test with a moderately difficult starting item for students with and without SEN. Simulations 11 and 12 simulate an adaptive test with an easy starting item for students with and without SEN.

Apart from the starting rule, the previously known settings for the adaptive algorithm are used. Accordingly, θ is estimated after each answer using the Bayesian Model Estimator, and the next item with the highest MFI is selected. The CAT terminates when the θ can be estimated with a $SE(\theta) = 0.5$. The final θ is estimated using ML. For each simulation, the mean test length and the correlation between true and estimated θ are analyzed. In addition, the percentage of test runs that could be successfully terminated due to the stopping rule is evaluated.

For simulations 1, 5, 7, and 11, the distributions of the number and proportion of correctly and wrongly solved tasks are examined. Thus, a non-adaptively performed reading screening (simulation 1), an adaptively performed reading screening with a simple start item (simulation 5), a non-adaptively performed test with generated item pool and simple start item (simulation 7), and an adaptively performed test with generated item pool and simple start item (simulation 11) can be compared in this context.

4 Results

This chapter presents the results obtained from three simulation studies. The studies were conducted to investigate the potential of CAT in the assessment of student groups with and without SEN on the example of a reading screening. The chapter is divided into three main sections, each reporting the findings of a different study. The first study focused on the simulation of CAT for each subtest ("subCATs") of the reading screening. The second study simulated the whole screening as CAT. Both studies examined the efficiency and accuracy of these measures. The third study compared the simulated CAT performance of students with and without SEN. The findings of each study are presented in detail, starting with the descriptive statistics and results of the simulations and then discussing the implications and limitations of the findings.

4.1 Study 1: Simulation of a Computerized Adaptive Reading Screening for Inclusive School Use

4.1.1 Results

Out of the 400 students who were measured, 393 completed all subtests as shown in Figure 1. However, nine students with SEN-I had difficulty with sentence reading during the sample task of subtest 4, so this task was omitted for them.

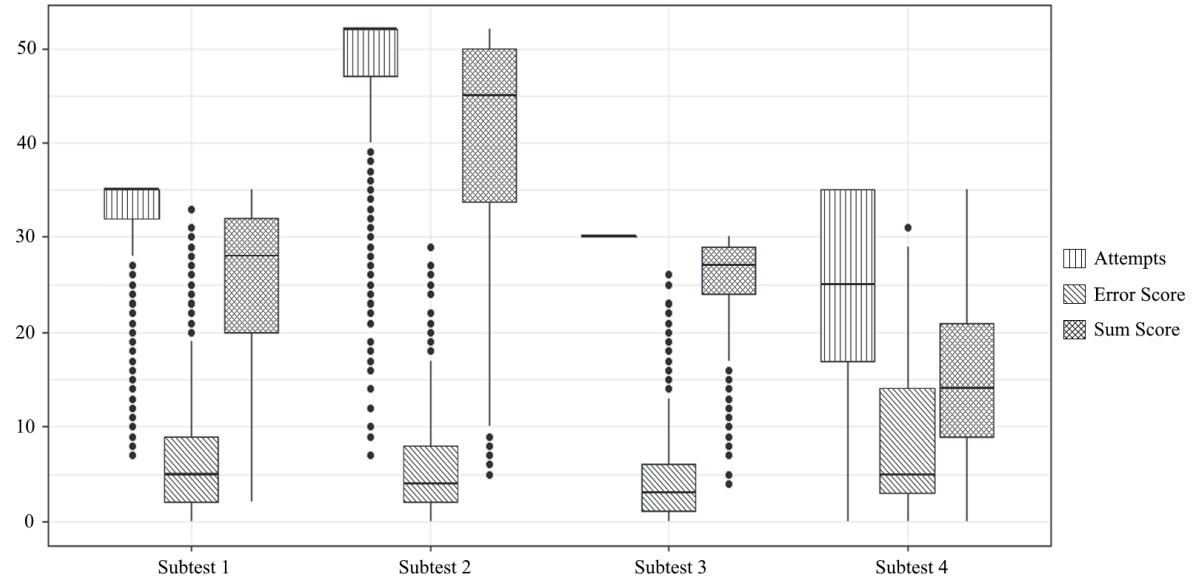
In subtest 1, students completed between 7 and the maximum number of 35 items in five minutes (*Md* = 35, *M* = 31.67, *SD* = 6.57). Their accuracy ranged between 5.7% and 100% (*Md* = 85.71, *M* = 78.02, *SD* = 20.42). In subtest 2, students completed between 7 and the maximum number of 52 items in five minutes (*Md* = 52, *M* = 46.94, *SD* = 9.93). Their accuracy ranged between 37.50% and 100% (*Md* = 91.21, *M* = 85.55, *SD* = 14.38). Subtest 3 had no time limit but was discontinued when all 30 items were answered, thus all students completed the subtest. Their accuracy ranged between 13.33% and 100% (*Md* = 90.00, *M* = 85.27, *SD* = 18.02). In subtest 4, students completed between 0 and the maximum number of 35 items within five minutes (*Md* = 25, *M* = 24.14, *SD* = 9.71). Their accuracy ranged between 11.11% and 100% (*Md* = 72.22, *M* = 66.19, *SD* = 24.22).

The number of items processed varied within each subtest. Although the majority of students were able to complete all items in subtests 1, 2, and 3, there were some outliers who completed significantly fewer items within the allotted time. This suggests a slower processing speed,

which could affect their overall score as the maximum processing time is taken into account. Furthermore, high error rates were observed across all tests, indicating that the tests were not necessarily too easy for the student population being assessed.

Figure 1

Distribution of the number of attempted items, error score and sum score for subtest 1, 2, 3, and 4.



To ensure that each subtest is a good fit of to the Rasch Model, Andersen's Likelihood Ratio tests were performed using both median and random split. Results indicated that subtest 1, 2, and 3 show no significant differences for median (subtest 1: χ^2 (69) = 58.86, p = 0.80; subtest 2: χ^2 (103) = 0, p = 1; subtest 3: χ^2 (59) = 0, p = 1) and random split (subtest 1: χ^2 (69) = 33.23, p = 1; subtest 2: χ^2 (103) = 59.22, p = 1; subtest 3: χ^2 (59) = 19.68, p = 1). Subtest 4 also shows no significance for random split (χ^2 (59) = 29.67, p = 1). However, for the median split, there is a significant difference in σ for students below and above the median (χ^2 (59) = 194.73, p < 0.001).

A graphical model test (Figure C1) revealed that seven items of subtest 4 indicated a bad fit, as their difficulty and its confidence interval were far off the line. As a result, these items were removed, leading to an item pool size of 28 items for subtest 4. Following the removal of the problematic items, the difference of σ between students above and below the median decreased and was not significant (χ^2 (55) = 56.02, p = 0.44). The revised subtest 4 therefore now better fits the Rasch Model, as the difference in σ in no longer driven by poorly calibrated items.

Table C1 shows the σ, infit, and outfit MSQ values for the subtests. As previously mentioned, the optimal values for MSQ infit and outfit statistics range between 0.5 and 1.5. Values up to 2.0 are generally acceptable and do not compromise the measurement system. However, values exceeding 2.0 can distort or degrade the measurement system, while values below 0.5 are less productive for measurement but do not cause degradation (Linacre, 2002). In this study, the infit and outfit MSQ values for the subtests are within the acceptable range. The σ values also fall within the expected range and exhibit a wide range of difficulty, which is important for adaptive testing. It is worth noting that subtest 3 has the narrowest difficulty range among the subtests.

The person-item maps depicted in Figure 2 confirm the characteristics observed in the data analysis. Moreover, these maps reveal a ceiling effect, particularly in subtests 2 and 3, as many students have correctly completed the most challenging items of these subtests. However, due to the absence of more difficult items, it's not possible to estimate the ability of these high-performing students any further, resulting in the ceiling effect.

Although this ceiling effect can be a concern in some contexts, it's not a significant issue for this screening, as the primary objective is to identify lower-performing students. This ceiling effect simply indicates that many students in the sample have already mastered the skills being tested. To perform a more realistic simulation, a new distribution of θ (n = 1000) is generated based on the *M* and *SD* of the θ distribution, as shown in Figure 3. This is done to remove the ceiling effect.

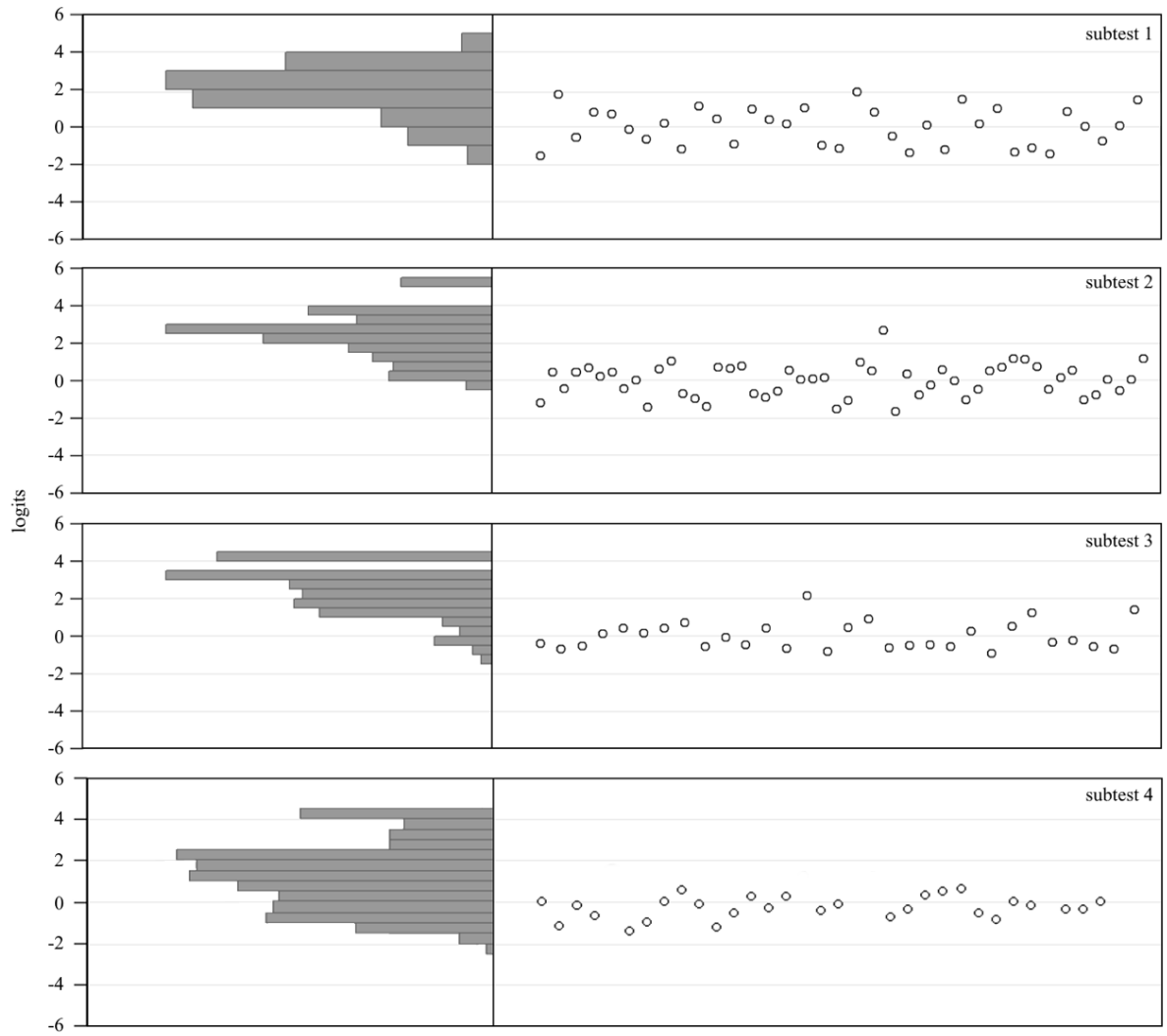
After removing the ceiling effect and generating a new distribution of θ, the results indicate a significant expansion in the distribution, particularly in subtests 1 and 3. This expansion of θ suggests that the majority of students in the sample possess a higher level of skill in these areas. Conversely, subtest 4 had the lowest measured θ, indicating that this subtest assesses more challenging skills requiring further development. In contrast, subtest 2 had the highest measured θ, implying that the skills assessed in this sub-test are relatively easier for students to master.

Although generating a new distribution of θ has resulted in broader ranges of measured abilities for each subtest, it's essential to note that it's no longer feasible to compare the abilities of individual students across all four subtests. This is because the new distribution generates a unique set of θ for each subtest, which means that a student's ability in one subtest is no longer directly

comparable to their ability in another subtest. Therefore, it's not possible to track the development of individual students' skills across all four subtests using this new distribution.

Figure 2

Person-item-maps of subtests 1, 2, 3, and 4.



Note: The histograms represent the distribution of estimated person abilities, and each point corresponds to an item and its difficulty. The points are ordered in a test run's item order.

The findings revealed that for subtest 1, the test length could be reduced by about half with $SE(\theta) = 0.5$. Similarly, for subtests 2 and 4, the test length could be reduced by more than half, and for subtest 3, the test length could be reduced by one third. The accuracy of the test becomes somewhat less accurate with increasing *SE*. Despite this, high correlations between $r = 0.87$ and $r = 0.92$ with $SE(\theta) = 0.5$ were still observed. Over all subCATs, it was found that $SE(\theta) = 0.3$

was not an effective stopping criterion for almost any person. On the other hand, 71% to 94% of the test runs could be successfully stopped with an *SE* of 0.5. The remaining test runs were stopped as soon as all items had been processed. Given that significantly shorter test lengths are only associated with slight losses in accuracy, *SE*(θ) = 0.5 was selected as the stopping criterion for further simulations.

Figure 3

Estimated and generated θ for all subtests.

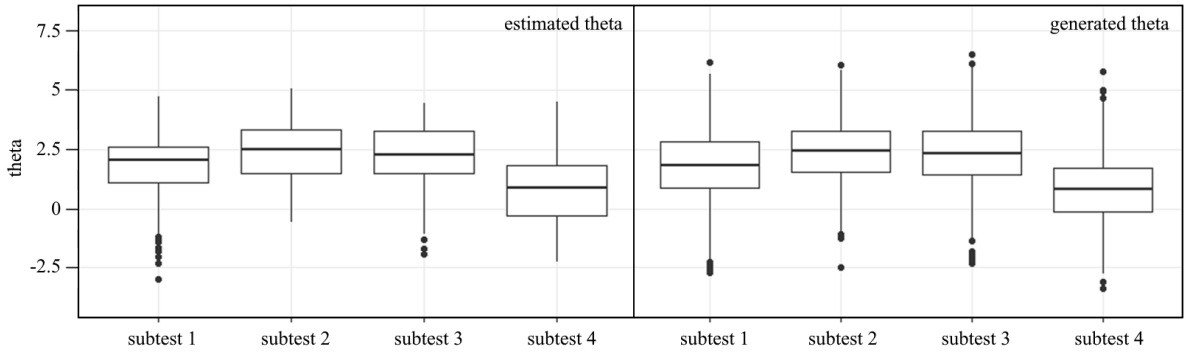


Table 3

Results of subCAT simulations comparing three different SE-based stopping rules and the resulting average test length, percentage of test runs terminated with the stopping rule and correlation between true and estimated θ.

subCAT	stopping rule	M length	amount satisfied stop	θ cor
1	SE = 0.3	35.00	0.00	0.93
	SE = 0.4	30.65	0.46	0.93
	SE = 0.5	18.26	0.83	0.91
2	SE = 0.3	51.38	0.13	0.92
	SE = 0.4	38.70	0.53	0.91
	SE = 0.5	20.34	0.92	0.87
3	SE = 0.3	30.00	0.00	0.89
	SE = 0.4	28.91	0.21	0.90
	SE = 0.5	20.79	0.71	0.87
4	SE = 0.3	28.00	0.00	0.94
	SE = 0.4	26.13	0.42	0.94
	SE = 0.5	17.48	0.81	0.92

To evaluate the accuracy of the subCATs, the correlation between the true and estimated θ was calculated. SubCAT 1 (Figure D1) accurately measures $-3 < \theta < 2$, with higher deviation for higher θ due to the ceiling effect. This effect is largely due to the lack of difficult items in the subtest, resulting in estimated θ being limited to certain values (~2.5, 3, 4) and fewer values in between. As a result, the accuracy of the subCAT is weaker for higher θ. SubCAT 2 (Figure D2) can accurately measure $-2 < \theta < 2.5$, but higher θ are again measured with higher deviation. The ceiling effect is also noticeable here, with estimated θ largely limited to values ~3 and 4 for θ higher than 2.5. SubCAT 3 (Figure D3) has the least pronounced ceiling effect and can accurately detect $-4 < \theta < 2.5$. However, only certain values are measured for high θ, such as ~4 and ~2.5. The lack of suitable items for σ ~3 may contribute to the missing estimated θ.

The subCATs also demonstrate a consistent pattern in terms of the test lengths required to achieve the desired accuracy levels. In general, the majority of individuals achieve short test lengths in the subCATs, particularly those with $\theta < 2$. For these individuals, the test runs can be reliably terminated based on the achieved accuracy. However, the test runs of individuals with higher θ are less likely to be terminated with the targeted accuracy, and as a result, longer test lengths are required. In subCAT 1 (Figure D9), for example, 80% of the individuals required less than 20 items to complete the test. In subCAT 2 (Figure D10), 90% of the subjects achieved test lengths of less than 30 items. Similarly, in subCAT 4 (Figure D12), 90% of the individuals required less than 20 items for a test run. On the other hand, subCAT 3 (Figure D11) shows a different picture, with test lengths being evenly distributed, and significant test reduction can be made for fewer individuals here.

Since the subCATs have issues estimating exact θ, particularly in the extreme areas, σ and estimated θ for $\theta = -1$, $\theta = -3$, and $\theta = 3$ are simulated as example. The extreme values of -3 and 3 are used to demonstrate the subCATs' limits. Figure 11-13 show the results of those simulations for subCAT 1 as an example, with the rest of the subCATs in the appendix (Figure D13-D19). The analysis revealed that in those extreme areas, the subCATs cannot provide sufficient easy or difficult items. Thus, the person's ability cannot be estimated accurately enough to terminate the CAT. As the current stopping rule does not apply in these cases, all items are instead selected. This results in a large number of items, all of which are either too easy or too difficult.

For extreme θ, θ and σ are shown to converge initially. As long as the two values converge, there are enough easy and hard items, respectively. At one point, however, no easier or harder

items are available. This point can be seen in Figures 5 and 6 in that the two graphs overlap and then the values move away from each other again. The estimated θ and σ of the selected items drift further and further apart. As a result, people receive many items that are too easy or too hard. Especially for people with low ability, this results in many consecutive wrong answers.

If, on the other hand, items are available with suitable σ, both σ and θ quickly settle at the expected value and the test can be terminated with the desired *SE* after a few items. This can be seen in Figure 4 of the simulation for θ = -1, which shows how the course and ratio of θ and σ should look in the optimum case. θ and σ often overlap and converge closer and closer to the true θ as the test continues. This results in a balanced ratio of easy and difficult and thus correct and incorrect answers.

Figure 4

σ of selected items and estimated θ while a simulated test run of a person with true θ = -1.

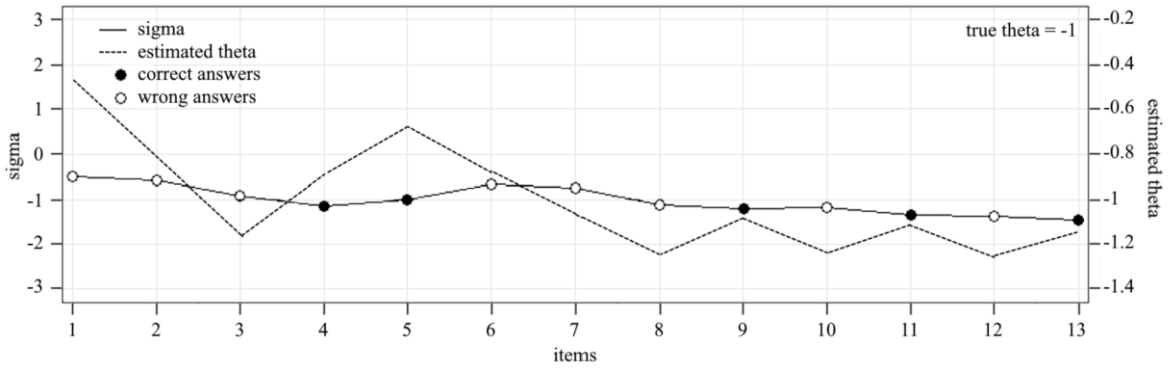


Figure 5

σ of selected items and estimated θ while a simulated test run of a person with true θ = -3.

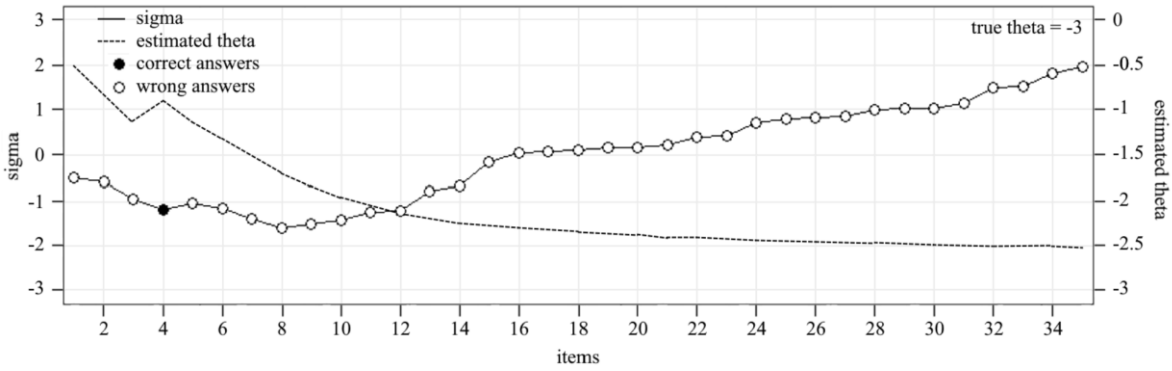
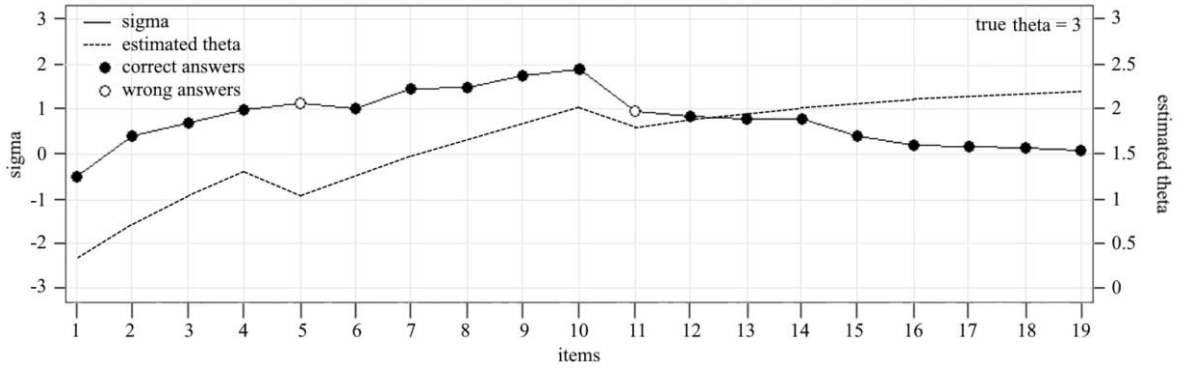


Figure 6

σ of selected items and estimated θ while a simulated test run of a person with true θ = 3.



To test how a CAT would perform for a particular group of students, an item pool with uniformly distributed σ (n = 100) for -3 < σ < 3 was generated (see Figure 7). This item pool is larger and more widely distributed than the subtests of the screening, which enables a more comprehensive and accurate measurement of students' abilities.

By using this generated item pool as a basis for further CAT simulations for the previous θ ranges of the subtests, it is possible to shorten the test for each θ range (as shown in Table 4). The results of these simulations showed that the average test length for each θ range was shortened by up to six items on average, with an average test length of about 14 to 15 items. 100% of the simulations can be stopped with a sufficient accuracy of $SE(\theta) = 0.5$, which allows up to 30% more simulations to be stopped due to this criterion. By being able to stop all simulations based on the accuracy of the measurement, it should also no longer happen that the entire item pool is pulled instead, even if it only contains items that are too hard or too easy. This also increases the accuracy of the measurement, which can be seen in the correlation between the actual and estimated θ values.

With the introduction of a new item pool with a broader and better balanced distribution, the CAT can now successfully terminate in both low and high performance ranges due to its increased accuracy. This means that students who perform particularly well or particularly poorly can now be accurately assessed with the new item pool.

Furthermore, when looking at the simulated courses of an individual student (Figure 8), it was observed that the CAT was able to terminate after just a few items. This demonstrates the efficiency of the new item pool, as the entire item pool was not drawn upon in certain situations,

despite the fact that the remaining items were either too easy or too difficult for the student being assessed.

Figure 7

σ of generated uniform item pool.

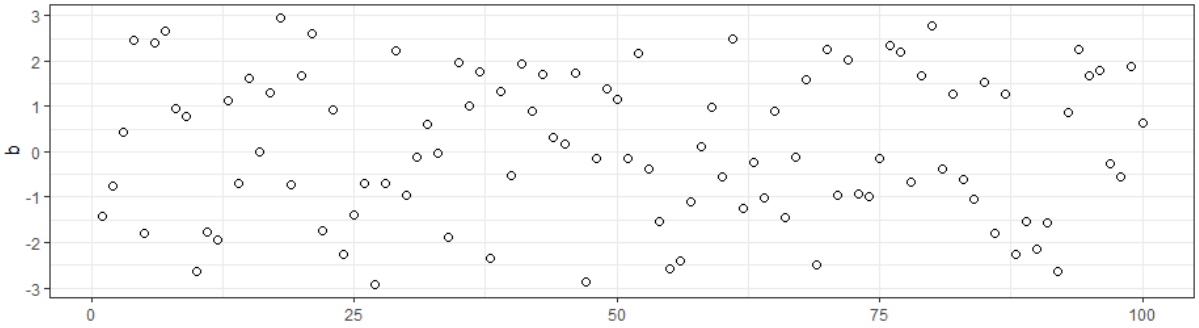


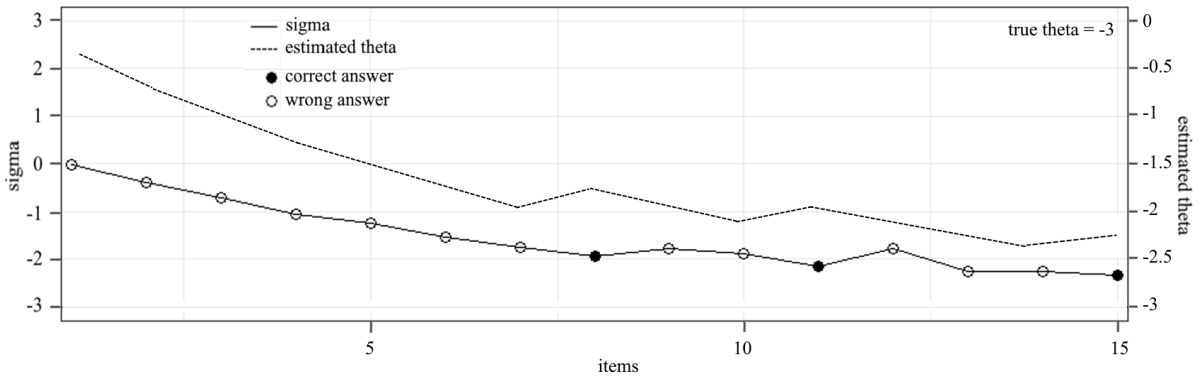
Table 4

Results of subCAT simulations comparing true and generated item pools.

θ of subtest	item pool	stopping rule	*M* length	amount satisfied stop	θ *cor*
1	True	*SE* = 0.5	18.26	0.83	0.91
	Generated	*SE* = 0.5	14.41	1.00	0.91
2	True	*SE* = 0.5	20.34	0.92	0.87
	Generated	*SE* = 0.5	14.92	1.00	0.89
3	True	*SE* = 0.5	20.79	0.71	0.87
	Generated	*SE* = 0.5	14.99	1.00	0.90
4	True	*SE* = 0.5	17.48	0.42	0.92
	Generated	*SE* = 0.5	14.10	1.00	0.93

Figure 8

σ of selected items and estimated θ while a simulated test run of a person with true θ = -3.



4.1.2 Discussion and Limitations

In the study, adaptive testing of a group of primary students with heterogeneous reading abilities, including those with and without SEN, was simulated using different item pools and stopping rules. For this purpose, four item pools from a reading screening for inclusive classes were utilized, and a fifth item pool was generated uniformly.

Heterogeneous Reading Abilities of the Sample

The results of the reading screening subtests show high heterogeneity in the students' responses, with an overall high heterogeneity of total scores across all subtests. This suggests that the sample of students with and without SEN from different grades is heterogeneous in their reading abilities, making it suitable to represent a sample in inclusive education. While some students solved many items correctly, there are also students in the low-performance range who solve only very few items correctly. Additionally, some students complete only a few items in general, which may be due to factors such as slower work pace, lack of concentration, or lack of motivation among students with SEN (Bender & Smith, 1990; Cullinan et al., 1981; Silver, 1981; Toro et al., 1990). It is important to note that these lower results suggest that the prerequisites among students with SEN could be playing a role here. Thus, it would be worth examining in a follow-up study, particularly in terms of SEN status and class comparisons, to understand the underlying factors contributing to this heterogeneity in the sample.

Subtests 1-3 show higher sum scores due to the earlier development of necessary reading skills, while weaker scores in subtest 4 indicate a need for further acquisition of sentence comprehension skills. The median sum scores of subtests 1, 2, and 3 are higher compared to subtest 4. Similarly, the percentage of correctly solved tasks is also higher in subtests 1-3 than in subtest 4. This can be attributed to the fact that subtests 1 to 3 measure skills that should already be present in the corresponding grades. Phonological awareness (Wagner & Torgesen, 1987), word knowledge (Röthlisberger et al., 2021), and reading speed (Ennemoser et al., 2013) are necessary conditions for successful fluent reading of sentences and texts. These skills are developed earlier in the reading process and are therefore addressed earlier in early childhood and school literacy education. In contrast, sentence reading is learned later and builds, among other things, on the successful acquisition of earlier skills (Coltheart et al., 2001). The lower sum scores in subtest 4 can be explained by the fact that students have not yet successfully acquired sentence comprehension skills in their reading process and need to further develop this ability.

These results were expected in the selected sample with a focus on the second to fourth grade levels.

A limiting factor is that the subtests used have not yet been validated. Although the subtests measure task types that are commonly used in reading instruction in schools, their ability to measure the constructs mentioned cannot be confirmed without systematic validation. However, since this study only uses the σ values as realistic sample data without further including the students' reading skills or the content of the items in the study and analyses, this aspect is not mainly relevant for the results of the simulation studies. It can be expected that item pools with comparable difficulty and validated items would show similar results in a simulation study.

The Impact of Item Pool Size and Difficulty Range on CAT

During the various simulations conducted, it became apparent that the size of the item pool and the range of item parameters, in particular, have an impact on the accuracy and length of the CAT. These results are in line with the assumptions and findings of Reckase (2003), Segall et al. (2000), and Ebenbeck and Gebhardt (submitted).

The range of σ and the size of the item pool are crucial factors for the accuracy and length of the measurement. This is evident from the fact that the uniformly distributed item pool performs better in CAT simulations than the four actual screening item pools. These results align with the assumptions made by Chen et al. (2000), which suggest that examinees with extreme θ ranges require a sufficient number of items from a uniformly distributed item pool. Simulations based on the generated item pool are more accurate and shorter for three out of four θ ranges (subtests 2, 3, and 4) when compared to simulations based on the real item pools. This is likely due to the wider range of σ values present in the generated item pool, which offers a greater coverage of difficulty levels. On the other hand, the real subtests have a relatively narrow σ range and contain only a few items overall, which is not ideal for CAT. Subtest 3, with the narrowest range of σ, shows the weakest results, although still achieving an accuracy of r = .87.

The effectiveness of a CAT stop rule based on accuracy is dependent on the size and range of the item pool. A stopping rule of $SE(\theta) = 0.5$ was chosen for the test, with the aim of achieving a short and accurate measurement with high efficiency. By using $SE(\theta) = 0.5$, the test is shorter, with only a minimal impact on accuracy. The loss of accuracy in correlation ranged from 0.02 to 0.05. This decrease in accuracy is less than what was observed in the results of Michiel et al. (2008) and is consistent with the accuracy ratio reported by Ebenbeck & Gebhardt (2022).

It should be noted that not every student can be tested when using the screening subtests' item pools due to the stopping rule. If the easiest item in the CAT cannot be solved correctly, it does not necessarily mean that θ can be estimated with the desired accuracy, such as $SE(\theta) = 0.5$. Consequently, the CAT cannot be stopped, and all remaining items are drawn from the item pool. However, with the generated item pool, the stopping rule can be applied to every student. Therefore, it can be concluded that the effectiveness of the stopping rule also depends on the size and σ range of the item pool.

For small item pools such as the real subtests, the stopping rule is appropriate for most students. However, it can pose a problem for learning groups in inclusive settings when, after a certain value where no easier items are available, the rest of the item pool is drawn, even though it only contains more difficult items. In such a scenario, the test would become equivalent to a regular non-adaptive test, and the difficulty level would be arranged in ascending order, leading to incorrect responses from students, which is not ideal for motivation. To address this issue, other solutions need to be explored, such as expanding the item pool to include more items of different difficulties or adding an additional stopping rule that stops testing when no easier items are available or after making, for example, three incorrect responses following the easiest item.

4.2 Study 2: Comparing Test Length and Measurement Accuracy pf Standalone and Sequentially Linked Adaptive Tests for Inclusive School Use – A Simulation Study

4.2.1 Results

The sum scores of the subtests correlate with each other with .34 < r < .55 ($p < 0.001$) (Table 5). The lowest correlation is between subtest 1 and 3. The highest correlation is between subtest 2 and 4. Thus, there is some correlation between the subtests. In particular, the successive subtests correlate with each other: subtest 1 is performed first. The correlation with the following subtest 2 is $r = .44$ ($p < 0.001$). The results of this subtest correlate with the following subtest 3 to $r = .45$ ($p < 0.001$). Subtest 3 is followed by subtest 4. The sum scores of these subtests correlate to $r = .54$ ($p < 0.001$).

Table 5

Means, standard deviations, and correlations with confidence intervals for the sum scores of the screening subtests.

Variable	*M*	*SD*	1	2	3
1. Sum_1	26.03	8.14			
2. Sum_2	41.92	10.01	.44** [.35, .53]		
3. Sum_3	26.19	4.74	.34** [.24, .44]	.45** [.36, .53]	
4. Sum_4	14.96	7.22	.40** [.30, .49]	.55** [.47, .62]	.54** [.46, .61]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

In simulation 1, a simulation of screening was conducted where each subCAT was executed sequentially by individuals with real θ, as presented in Table 5. The subCATs were initialized with an item of moderate σ, as in study 1. The average length of the subCATs ranged from 17.31 to 21.84 items, leading to an average screening length of 78.28 items. The precision criterion resulted in the termination of subCATs in 64% to 91% of cases, achieving an accuracy of $r = .88$ to $r = .92$.

Table 6

Results of screening CAT simulations with real and generated item pools and real and generated person abilities comparing different starting rules.

simulation	item pool	θ	starting rule	*M* length	% satisfied stop	θ *cor*
1	Subtest 1	Subtest 1	fixed item	17.92	0.86	0.90
	Subtest 2	Subtest 2	fixed item	21.21	0.91	0.88
	Subtest 3	Subtest 3	fixed item	21.84	0.64	0.88
	Subtest 4	Subtest 4	fixed item	17.31	0.83	0.92
2	Subtest 1	Subtest 1	fixed item	17.92	0.86	0.90
	Subtest 2	Subtest 2	Final θ of subCAT 1	27.43	0.90	0.88
	Subtest 3	Subtest 3	Final θ of subCAT 2	23.19	0.71	0.89
	Subtest 4	Subtest 4	Final θ of subCAT 3	18.98	0.84	0.93
3	Subtest 1	Fixed	fixed item	13.49	1.00	0.87
	Subtest 2	Fixed	fixed item	13.46	1.00	0.87
	Subtest 3	Fixed	fixed item	13.95	1.00	0.87
	Subtest 4	Fixed	fixed item	14.19	0.99	0.86
4	Subtest 1	Fixed	fixed item	13.49	1.00	0.87
	Subtest 2	Fixed	Final θ of subCAT 1	13.52	1.00	0.86
	Subtest 3	Fixed	Final θ of subCAT 2	14.15	1.00	0.87
	Subtest 4	Fixed	Final θ of subCAT 3	14.39	0.99	0.85
5	Generated	Subtest 1	fixed item	14.37	1.00	0.91
	Generated	Subtest 2	fixed item	14.96	1.00	0.87
	Generated	Subtest 3	fixed item	14.97	1.00	0.91
	Generated	Subtest 4	fixed item	14.01	1.00	0.93
6	Generated	Subtest 1	fixed item	14.37	1.00	0.91
	Generated	Subtest 2	Final θ of subCAT 1	14.70	1.00	0.88
	Generated	Subtest 3	Final θ of subCAT 2	14.96	1.00	0.91
	Generated	Subtest 4	Final θ of subCAT 3	14.09	1.00	0.94

In simulation 2, the subCATs were linked together by using the final estimated θ from the previous test as the starting value for the following subCAT, as shown in Table 6. This start rule did not shorten the test but rather increased the test length by approximately 1.5 to 6 items, leading to an average screening length of 87.52 items. However, the change in the start rule had minimal to no impact on the proportion of test runs that could be discontinued due to their accuracy and had little to no effect on the accuracy of the measurement.

Looking at an individual run of a real person with the θ values θ = -1.82 for subtest 1, true θ = -0.27 for subtest 2, true θ = -0.03 for subtest 3, and true θ = -2.50 for subtest 4 (Figure 9 and Figure 10), similar results can be observed. The total test length is even longer when the

subCATs are linked. Although the starting item may be different from the unlinked screening version, this starting item has little influence on the further course of the test.

Figure 9

σ of selected items and estimated θ while a simulated test run with fixed first item of a person with true θ = -1.82 for subtest 1, true θ = -0.27 for subtest 2, true θ = -0.03 for subtest 3 and true θ = -2.50 for subtest 4.

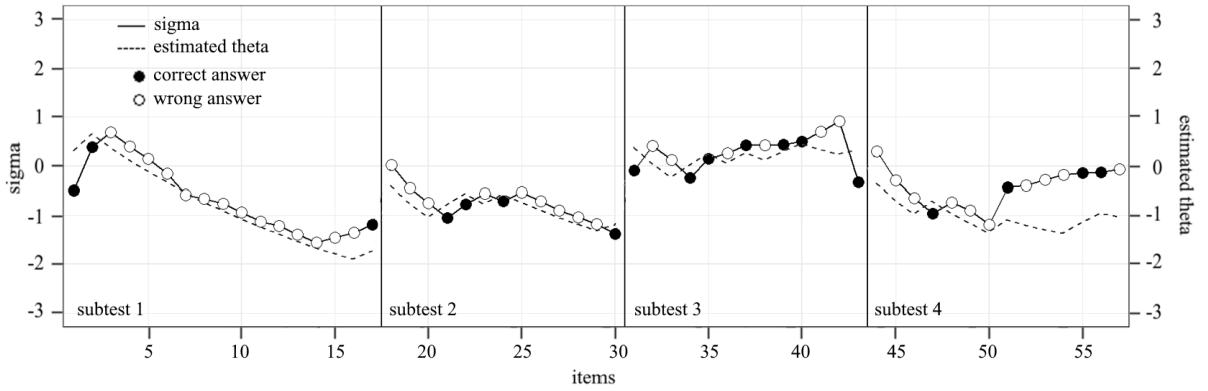
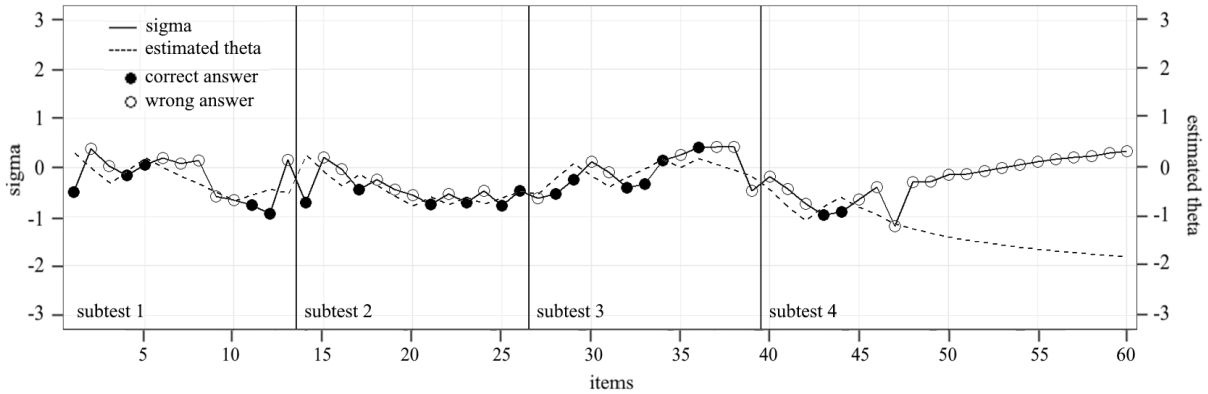


Figure 10

σ of selected items and estimated θ while a simulated test run with θ input of the previous estimated θ of a person with true θ = -1.82 for subtest 1, true θ = -0.27 for subtest 2, true θ = -0.03 for subtest 3 and true θ = -2.50 for subtest 4.

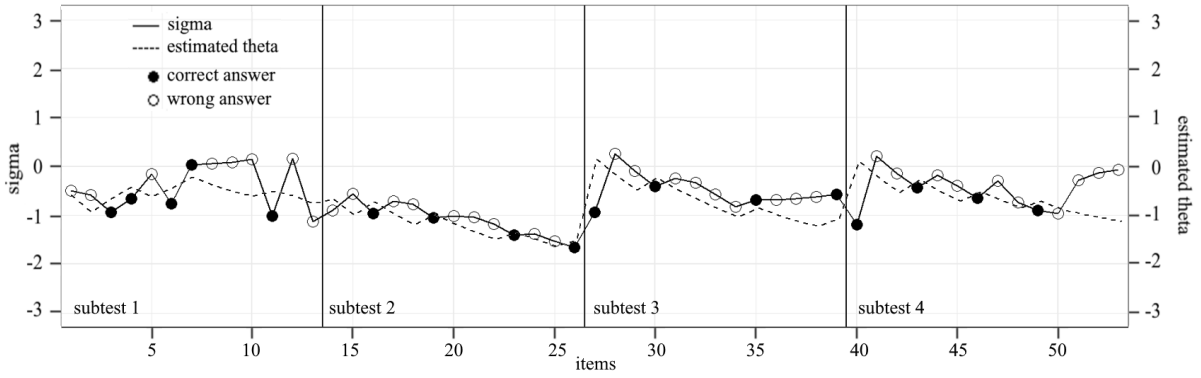


In simulation 3, individuals were simulated in separate but unconnected subCATs, each demonstrating the same ability level in each subtest, as detailed in Table 6. Due to the normally distributed θ, the average test length of the subCATs ranged from 13.46 to 14.19 items, resulting in an average screening length of 55.09 items. Almost all test runs could be discontinued due to their accuracy, with test runs measuring with an accuracy of r = .85 to r = .87.

In simulation 4, individuals were simulated in a connected screening CAT, each demonstrating the same ability level in each subtest, as presented in Table 6. The simulations showed almost no differences from simulation 3. The average test length of the subCATs ranged from 13.49 to 14.39, leading to an average screening length of 55.55 items. Almost all test runs could be discontinued due to their accuracy, achieving an accuracy of .85 to .87. An example of a person with θ = -1 in all subtests is shown in Figure 11.

Figure 11

σ of selected items and estimated θ while a simulated test run with θ input of the previous estimated θ of a person with true θ = -1 for all subtests.



Simulation 5 involved simulating separate subCATs with generated σ and true θ from the sample, as detailed in Table 6. The average test length for the subCATs ranged between 14.09 and 14.96 items, corresponding to an average screening length of 58.31 items. All test runs could be terminated due to their accuracy, with the subCATs achieving an accuracy between r = .87 and r = .93.

In the last simulation, simulation 6, generated σ and true θ were used for a connected CAT, as presented in Table 6. The average test length for the subCATs ranged between 14.09 and 14.96, corresponding to an average screening length of 61.12 items. All test runs could be terminated due to their accuracy, with the subCATs achieving an accuracy between r = .87 and r = .93.

Overall, the findings of the simulations indicate that the new stopping rule and connecting the subCATs have no effect on the length or accuracy of the CAT. These results hold true for both true and generated person parameters, which are either homogeneous or heterogeneous. Furthermore, these results hold true for the real items of the screening as well as an artificially generated item pool with more and broadly distributed items.

4.2.2 Discussion and Limitations

In this study, the aim was to investigate whether the computerized adaptive version of a modular test could result in a more accurate and shorter measurement by adjusting the start rule for sequentially administered subtests. To achieve this goal, a modular screening was simulated in six rounds, with each round using either the real item pools of a screening, a uniformly generated item pool, real θ, or uniformly generated θ as simulation input.

Potential of Test Length Shortage and Measurement Accuracy

In no simulation did connecting the subCATs have a significant effect on the length of the resulting screenings. When using the real item pool, the test length of a connected screening is on average even longer than when simulating four individual tests. There is hardly any change when using the uniformly generated item pool. Thus, although the size and distribution of σ have a positive effect on connecting subtests, it cannot contribute to a shorter test length overall. In particular, students with low abilities are likely to experience the same pattern of errors observed in study 2: if the ability cannot be estimated accurately enough when the easiest item is answered incorrectly, the remaining items that are too difficult are drawn from the item pool instead, lengthening the test duration. The modified starting rule probably has little to no influence on this issue.

This contradicts the assumptions of Weiss & Kingsbury (1984), who suggested that incorporating prior information about the examinee could lead to shorter tests. In this case, the reason for the lack of test length reduction cannot be attributed to a low correlation between the subtests, as one might suspect. Although the final estimated θ values between the subtests are only correlated with each other up to a maximum of .55. However, even for simulated individuals with perfectly correlated θ values, there is no shortening of the test length. Instead, it is likely that the starting item has significantly less influence on the test process compared to the subsequent algorithm. Another possibility is that the Bayesian estimator, used to incorporate the person's ability at the beginning of the measurement, may not be as effective in selecting an item as setting a fixed first item.

Also, there are hardly any changes in the accuracy of the CAT whether the subCATs are connected or not, regardless of the item pool used or the simulated θ. It should be noted that a low correlation between the subtests cannot be the reason for this, as otherwise differences in accuracy for the uniform θ would have been observable. Similarly, a too small or too narrow item pool cannot be responsible for this, as changes should have been observable for the uniform

item pool. Instead, it can be concluded that the remaining CAT algorithm for drawing the next item compensates for any possible bias introduced by a fixed start item, in such a way that the changed start rule does not matter. This is further supported by the fact that there are hardly any differences in the proportions of successfully stopped test runs, regardless of the start rule. This should not be seen as a negative implication for connecting CATs, but rather for the quality of the algorithm as a whole.

Practicability of Implementation

The use of a starting rule that connects the subCATs with each other is unlikely to harm the measurement, as it could potentially shorten the test for some students if the correlation between the subtests is high (Weiss & Kingsbury, 1984). However, the increased programming and computational effort required for such an implementation, as well as the reduced flexibility in everyday school life and classroom use, suggest that a cost-benefit analysis is necessary. It is important to note that individual short tests can be used more effectively, given their shorter execution time and the ability for teachers to select specific skills to measure or not measure for some students in order to better meet the needs of a group in inclusive education. Therefore, the use of a starting rule that connects subCATs with each other is not recommended.

Instead, it is important to check whether the selected starting items, with medium difficulty, are suitable for students with low abilities or students with SEN. Alternatively, starting with an easy item instead of a medium-difficulty item could be considered. This approach would prevent overwhelming students with low abilities from the start and could possibly reduce their proportion of errors. However, for this to be effective, the item pool must have sufficient easy items, or else the desired accuracy level cannot be achieved with a few items, and remaining difficult items would have to be chosen at some point. This situation should be avoided at all costs. Additionally, when selecting an easier starting item, it is necessary to consider whether it would negatively impact students with high abilities. It is likely that these students would require more items after the first one to reach their performance range, resulting in longer test runs. Such issues should be explored in a follow-up study, as addressed in study 3.

4.3 Study 3: A Simulation-Based Comparison of the Effectiveness of Adaptive Tests for Students With and Without Special Educational Needs

4.3.1 Results

The school attendance year, grade level, and age of the students in the sample are highly correlated. The correlation between grade level and age is the lowest (r = .89, p < .001), while the correlation between school attendance year and age is the highest (r = .91, p < .001). Since there are no missing values for the grade level variable, a statement can be made for each student, and therefore, this variable is used for further analysis. The sample is divided into four groups based on their grade level: grade level 2 (n = 123), grade level 3 (n = 141), grade level 4 (n = 106), and secondary level (n = 30).

There are significant differences in the sum scores of all subtests across different grade levels (subtest 1: $F(3) = 4.05$, $p < .01$; subtest 2: $F(3) = 19.08$, $p < .001$; subtest 3: $F(3) = 8.72$, $p < .001$; subtest 4: $F(3) = 26.75$, $p < .001$). Students in grade 3 on average correctly solve significantly more items than students in grade 2 in subtests 2 and 4 (subtest 2: $t(246.21) = -4.51$, $p < .001$; subtest 4: $t(239.35) = -6.71$, $p < .001$). There is no significant difference in the sum scores of all subtests between grade 3 and grade 4 students. Similarly, there is no significant difference in the sum scores of all subtests between grade 2 students and secondary level students. Students in grade 3 and grade 4 on average correctly solve significantly more items than secondary students in subtest 2 (Grade 3: $t(33.84) = 3.97$, $p < .001$; Grade 4: $t(35.00) = 4.38$, $p < .001$). In addition, students in grade 4 correctly solve more items than secondary students in subtest 4 ($t(48.36) = 3.67$, $p < .001$).

Between students with SEN and students without SEN, there are significant average differences in the sum scores for all subtests (subtest 1: $t(123.59) = 4.70$, $p < .001$; subtest 2: $t(105.06) = 5.57$, $p < .001$; subtest 3: $t(101.58) = 5.24$, $p < .001$); subtest 4: $t(132.02) = 4.71$, $p < .001$). When comparing the sum scores of students with SEN-L to students with SEN-I and SEN-S with a t-test, there is no significant difference between those three groups. As a result, these groups are combined into a single group of students with SEN in the following simulations.

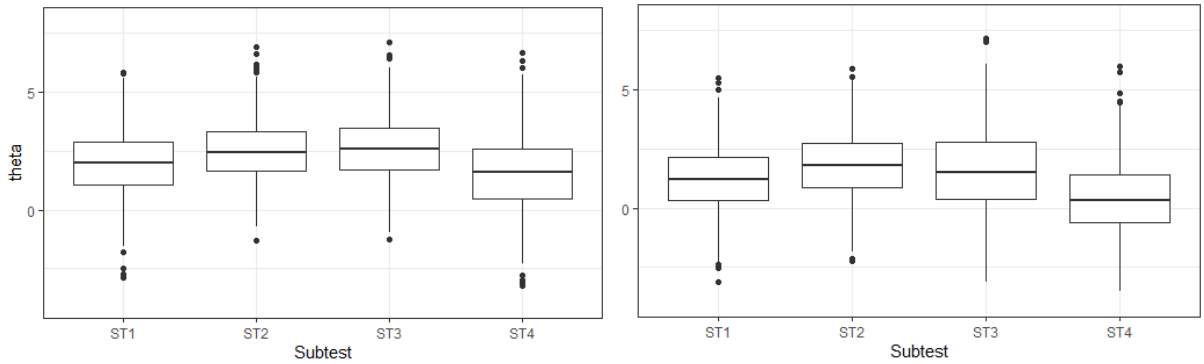
Students with SEN completed fewer items in total in subtests 1, 2, and 4 in the same amount of time as students without SEN (subtest 1: $F(3) = 11.90$, $p < .001$; subtest 2: $F(3) = 23.77$, $p < .001$; subtest 4: $F(3) = 5.188$, $p < .01$). Since subtest 3 had no maximum completion time, all students completed the same number of items. However, post-hoc analyses with direct

comparisons of student groups showed that only students with SEN-I were able to complete significantly fewer total items than students with SEN-L, SEN-S, or without SEN in subtests 1 and 2, and there was no significant difference in the total number of completed items otherwise.

Students without SEN complete significantly more tasks correctly on average than students with SEN, which also makes their θ estimated in the Rasch model differ from each other to the same extent. These θ ranges are again used to generate θ ranges per subtest for students with and without special needs (Figure 12).

Figure 12

Generated θ of students with (right) and without (left) SEN.



In the simulation of the linear screening (Table E1, simulations 1 and 2), accuracy in subtests for students with SEN ranges from r = .94 (subtest 1, 2, and 3) to r = .95 (subtest 4) and for students without SEN ranges from r = .85 (subtest 3) to r = .92 (subtest 4). Because a response pattern for the full item pool was simulated for all students, the mean test lengths did not differ; rather, each student completed the entire item pool.

In the simulation of an adaptive screening with a medium start item (Table E1, simulations 3 and 4), the accuracy for students with SEN in the subtests ranges from r = .90 (subtest 2) to r = .92 (subtest 3) and for students without SEN ranges from r = .84 (subtest 3) to r = .91 (subtest 4). The mean test length of the subtests ranged from 15.68 items (subtest 4) to 18.64 items (subtest 3) for students with SEN and from 17.86 items (subtest 4) to 22.05 items (subtest 3) for students without SEN. Thus, students with SEN would on average complete more items per test session. The percentage of test runs that were discontinued due to the target accuracy of $SE(\theta)$ = 0.5 ranged from 80% (subtest 3) to 95% (subtest 2) for students with SEN and from 65% (subtest 3) to 92% (subtest 2) for students without SEN. Accordingly, the test runs of students with SEN can be terminated more frequently due to their accuracy.

When the adaptive screening is started with an easier starting item (Table E1, simulations 5 and 6), the accuracy values change only slightly. Subtest accuracy ranges from r = .90 (subtest 1 and 2) for students with SEN and from r = .83 (subtest 3) to r = .90 (subtest 4) for students without SEN. The mean test length of the subtests nearly did not change through the easier starting item. It ranged from 15.64 items (subtest 4) to 18.86 items (subtest 3) for students with SEN and from 17.88 items (subtest 4) to 22.87 items (subtest 2) for students without SEN. Therefore, even with an easier starting item, students with SEN would still, on average, complete more items per test session. The percentages of test runs that were discontinued due to the target accuracy $SE(\theta) = 0.5$ were slightly higher with an easier starting item. They ranged from 79% (subtest 3) to 96% (subtest 2) for students with SEN and from 66% (subtest 3) to 93% (subtest 2) for students without SEN. Accordingly, also with an easier starting item, the test runs of students with SEN can be terminated more frequently due to their accuracy.

In the simulation of a linear screening with generated uniform item pool (Table E1, simulations 7 and 8), accuracy in subtests for students with SEN ranges from r = .97 (subtest 2 and 3) to r = .98 (subtest 1 and 4) and for students without SEN ranges from r = .94 (subtest 3) to r = .97 (subtest 1 and 4). Because a response pattern for the full item pool was also in this case simulated for all students, the mean test lengths did not differ; rather, each student completed the entire item pool.

In the simulation of an adaptive screening with generated uniform item pool and a medium start item (Table E1, simulations 9 and 10), the accuracy for students with SEN in the subtests ranges from r = .90 (subtest 2) to r = .93 (subtest 3) and for students without SEN ranges from r = .87 (subtest 2 and 3) to r = .92 (subtest 4). The mean test length of the subtests ranged from 13.74 items (subtest 4) to 15.13 items (subtest 3) for students with SEN and from 14.26 items (subtest 1) to 16.05 items (subtest 3) for students without SEN. Thus, students with SEN would on average complete more items per test session. The percentage of test runs that were discontinued due to the target accuracy of $SE(\theta) = 0.5$ ranged from 90% (subtest 2) to 93% (subtest 3) for students with SEN and from 87% (subtest 2 and 3) to 92% (subtest 4) for students without SEN. Accordingly, the test runs of students with SEN can be terminated more frequently due to their accuracy.

When an adaptive screening is started with a generated uniform item pool and an easier starting item (Table E1, simulations 11 and 12), the accuracy values again change only slightly. Subtest accuracy ranges from r = .91 (subtest 1 and 2) to r = .93 (subtest 3 and 4) for students with SEN

and from r = .86 (subtest 2 and 3) to r = .92 (subtest 4) for students without SEN. The mean test length of the subtests nearly did change through the easier starting item. It ranged from 13.59 items (subtest 4) to 14.24 items (subtest 3) for students with SEN and from 14.11 items (subtest 1) to 14.80 items (subtest 3) for students without SEN. So also with an easier starting item, students with SEN would on average complete more items per test session. All test runs were terminated because of their accuracy.

Table 7

Number and percentage of correct and wrong answers per subtest for different simulations of true and generated screening item pools.

simulation	θ	*n* correct *M*(*SD*)	*n* wrong *M*(*SD*)	% correct *M*(*SD*)	% wrong *M*(*SD*)
1	Subtest 1 SEN	24.52 (7.27)	10.48 (7.27)	70.05 (20.77)	29.95 (20.77)
	Subtest 2 SEN	40.39 (9.79)	11.61 (9.79)	77.68 (18.82)	22.32 (18.82)
	Subtest 3 SEN	21.82 (7.37)	8.18 (7.37)	72.73 (24.58)	27.27 (24.58)
	Subtest 4 SEN	15.82 (7.51)	12.18 (7.51)	56.50 (26.82)	43.50 (26.82)
5	Subtest 1 SEN	11.04 (7.24)	4.74 (2.21)	65.75 (17.30)	34.25 (17.30)
	Subtest 2 SEN	13.95 (9.94)	3.93 (2.10)	72.59 (17.18)	27.41 (17.18)
	Subtest 3 SEN	14.65 (9.33)	4.15 (3.05)	71.35 (22.22)	28.65 (22.22)
	Subtest 4 SEN	9.16 (6.49)	6.54 (4.65)	55.98 (23.64)	44.02 (23.64)
7	Subtest 1 SEN	65.75 (18.27)	34.25 (18.27)	65.75 (18.27)	34.25 (18.27)
	Subtest 2 SEN	72.70 (17.20)	27.30 (17.20)	72.70 (17.20)	27.30 (17.20)
	Subtest 3 SEN	68.80 (21.53)	31.20 (21.53)	68.80 (21.53)	31.20 (21.53)
	Subtest 4 SEN	53.89 (21.12)	46.11 (21.12)	53.89 (21.12)	46.11 (21.12)
11	Subtest 1 SEN	8.40 (2.67)	5.26 (2.04)	60.90 (15.31)	39.10 (15.31)
	Subtest 2 SEN	9.24 (3.05)	4.70 (2.08)	65.37 (15.52)	34.63 (15.52)
	Subtest 3 SEN	9.42 (3.98)	4.85 (2.49)	64.40 (18.58)	35.60 (18.58)
	Subtest 4 SEN	7.41 (2.86)	6.18 (2.35)	53.97 (17.05)	46.03 (17.05)

Comparing the real screening in a linear and an adaptive simulation, differences in the response behavior of students with SEN (Table 7) are evident. In the linear screening (simulation 1), more items are answered correctly and incorrectly because the length of the test is predetermined. The percentage of correctly solved items for this group of students ranges on average from 56.5% (subtest 4) to 77.68% (subtest 2). In the adaptive screening (simulation 5), the students complete fewer items overall. As a result, they also complete fewer items correctly and incorrectly. The proportion of correctly solved items is somewhat lower in comparison due to the use of the adaptive algorithm and lies on average between 55.98% (subtest 4) and 72.59% (subtest 2).

Similar results are shown for the generated item pools (Table 7). In the linear screening (simulation 7), much more items are answered correctly and incorrectly because the length of the test is predetermined, and the item pool was extended to 100 items. The percentage of correctly solved items for this group of students ranges on average from 53.89% (subtest 4) to 72.70% (subtest 2). In the adaptive screening (simulation 11), the students complete fewer items overall. As a result, they also complete fewer items correctly and incorrectly. The proportion of correctly solved items is somewhat lower in comparison due to the use of the adaptive algorithm and lies on average between 53.97% (subtest 4) and 64.40% (subtest 2).

4.3.2 Discussion and Limitations

In this study, various CATs were simulated for students with and without SEN in order to evaluate the performance of the CATs for these student groups. Based on this, the study investigated how students with SEN would perform in different CATs and how their response pattern would be. The CATs differed in the difficulty of their first item and the difficulty range and size of the item pool.

Start Item Choice

Using a simpler item with $\sigma = -1$ as the first item does not disadvantage any group of students. There are almost no performance differences or differences in the results of the CAT for students without SEN when using a simpler start item. However, for students with SEN, the use of an easier start item affects the length and accuracy of the measurement. When an easier start item is used, the measurement is slightly shorter, and more test runs can be terminated due to the accuracy.

From a practical point of view, using an easier start item facilitates entry into the measurement and allows more students to answer it correctly. This is likely to increase test motivation from the beginning. Therefore, it is recommended to use an easier start item for the development of adaptive reading screening and CATs for educational purposes in general. This approach not only ensures that no group of students is disadvantaged, but it also makes the test more accessible for students with SEN, leading to more accurate results and presumably increased motivation, what needs to be addressed in following research.

Differences between Students with SEN

In this study, it was found that students with SEN-I do not exhibit significantly lower reading skills compared to students with SEN-S or SEN-L. This finding contradicts previous research, which has suggested that students with lower intelligence tend to have lower reading performance than those with higher intelligence (Cohen et al., 2001; Di Blasi et al., 2019; Gresham et al., 1996; Levy, 2011). Given that students with SEN-I typically have a lower IQ than those with SEN-L, it was expected to observe differences in their reading profiles.

One possible explanation for this finding could be related to the sample selection process. Specifically, the sample of students with SEN-I in this study only included students with mild or moderate intellectual disabilities. This may have occurred because teachers were instructed to select students with basic reading abilities for the assessment, which are typically mastered by students with an IQ > 65 (Cohen et al., 2001). However, this approach may have excluded poor readers or those who are still in the process of learning to read from the sample. In contrast, the sample of students with SEN-L and SEN-S included those who were educated in inclusive schools. It can be assumed that this selection of students may have better literacy skills compared to those with SEN-L and SEN-S who are exclusively educated in special schools.

Another potential explanation for the findings could be related to the identification and labeling of students with SEN. This process is often not well-defined and can lead to overlapping categories. For instance, the boundaries between SEN-I, SEN-L, and SEN-S may not be clear, and classifying students solely based on their SEN status may not accurately reflect their individual strengths and weaknesses (Snell et al., 2009; Gresham et al., 1996). Moreover, the severity of disabilities among students in each group may have varied, and these differences could have offset each other in terms of their reading abilities. Therefore, in a follow-up analysis, it would be beneficial to examine more closely the varying degrees of disability severity among the students and their impact on reading skills. This could help to better understand and highlight the differences between these groups of individuals.

The aim of this study was not to provide a detailed description of the reading abilities of students with and without SEN, but rather to generate significantly different groups of students with and without SEN that can be used as input for simulation studies. In the research context, it is common to analyze and consider students with different types of SEN together (Arvidsson & Granlund, 2018; Bouck & Satsangi, 2015; Cohen et al., 2001; Nouwens et al., 2017), which is in line with the goal of this study.

Differences in the CAT Measurement of Students with and without SEN

The CAT algorithm requires fewer test items to assess the abilities of students with SEN compared to students without SEN, as shown consistently across all simulations using different item pools. However, when it comes to the reading screening CAT, the difference between students with and without SEN is somewhat larger. This is likely due to the adaptive algorithm being more sensitive to the θ ranges of students with SEN, allowing for a more accurate identification of their θ level and selection of appropriate test items.

It is reasonable that students with SEN would work on fewer items in a CAT than students without SEN. Research and results of this study show that students with SEN work more slowly than those without SEN (Ebenbeck et al., 2023; Kaznowski, 2004), leading to concentration and frustration problems and working on fewer items in the same time (Bender & Smith, 1990; Cullinan et al., 1981; Silver, 1981; Toro et al., 1990). A shorter test with fewer items, in which they also make fewer errors, can thus be seen as an adjustment to ensure fairness in the testing situation. Students without SEN may receive more items suggested by the CAT, but they also solve them more quickly. It is therefore possible that, overall, both groups of students take a similar amount of time to complete the same test. This is optimal for integrating such testing procedures into everyday school life, as a similar time frame can be assumed for testing in the classroom.

The CATs measure slightly less accurately than the non-adaptive tests. However, both the CATs and non-adaptive tests measure students with SEN more accurately. Despite the less accurate measurement with shorter testing in the CATs, students with SEN can still be measured with $r > 0.90$. Therefore, a very accurate measurement for students with SEN can be assumed despite the test being shortened. The fact that students with SEN can be measured more accurately than students without SEN is probably due to the composition of the item pool, which is more low-ability-oriented. This is also reflected in the smaller differences in accuracy between students with and without SEN in the generated item pool. Therefore, it can be concluded that expanding the item pool to include more items with a wider σ range would benefit both groups of students. Additionally, it is important to note that the purpose of the adaptive reading screening is not to compare students to one another, but rather to identify areas where they may need additional support. It is therefore more important for such tests to measure more accurately in lower performance areas, which is why this CAT performance is considered positive for both students with and without SEN.

5 Discussion

In the three studies presented, an adaptive reading screening was developed through simulation studies that built upon each other. The aim of the research project was to develop adaptive algorithms that could be applied to an existing reading screening. Furthermore, the studies sought to identify general conditions for adaptive screenings that could be applied in heterogeneous learning groups. The selected algorithms were tested for comparison on both real and generated data and were compared for different groups of students with and without SEN. The results and limitations of each study have already been discussed briefly. This chapter will provide a detailed discussion and examination of the entire research process. Additionally, the limitations of the research process will be identified, and suggestions for future research needs will be provided.

5.1 Development of an Adaptive Screening

5.1.1 Item Pool

An algorithm based adaptive test requires a suitable item pool based on the IRT, as noted by Bock and Gibbons (2021, p. 245). The performance of a CAT is impacted by the size of the item pool (Reckase, 2003). Therefore, various item pool sizes were compared to develop a CAT for students with SEN. The reading screening used in this study consists of four subtests, each with a different item pool size. Additionally, an item pool with 100 items and a uniform σ distribution was generated. All item pools were based on a unidimensional Rasch model.

The size and σ distribution of the item pool have an impact on several aspects of the adaptive test. Across all studies, it was found that a larger item pool with an evenly distributed σ results in a shorter test length for the same individuals and potentially increases the accuracy of the measurement. The shorter and more accurate measurement is due to the availability of more items in extreme difficulty areas, which better covers the examinees' θ through the distribution of the item pool.

In addition, a larger item pool provides sufficient items of suitable difficulty to draw several items of similar difficulty to an examinee during adaptive testing. This drawing is based on the examinee's response to previous items, and it helps to efficiently estimate their ability level. A larger and more evenly distributed item pool also increases the probability of an examinee

receiving an item that can provide a significant amount of information about their ability. These results aligns with the works of Segall (2005) and Chen et al. (2000), which highlight the importance of having a large and diverse item pool in adaptive testing. A comprehensive item pool can ensure that the adaptive test accurately measures an individual's ability and provides a more precise estimation of their performance level.

However, having a large item pool alone is not sufficient for accurate adaptive testing. The σ distribution must also be taken into account to ensure that there are enough items at each difficulty level. If there are not enough items at a certain difficulty level, even the easiest or hardest item may not be sufficient to estimate the θ of the examinee with the desired accuracy. In this scenario, depending on the stop rule, all further items may be drawn from the item pool until no more items are available for testing. As shown in studies 1 and 2, these items do not add much value to the estimation of the examinee's θ and only prolong the measurement. Therefore, this case should be avoided as much as possible. To ensure an accurate estimation of an examinee's ability level, it is essential to have a balanced item pool with a diverse range of difficulty levels. This can be achieved through careful item selection and calibration during the test development process. Additionally, the σ distribution should be checked to identify potential gaps in the item pool that could impact the accuracy of the adaptive test. By addressing these issues, the adaptive test can provide a more precise measurement of an individual's ability and improve the overall effectiveness of the screening process.

A large item pool with a wide range of item difficulties also benefits students in inclusive classrooms. Class-based tests are conducted together in the classroom, rather than individually with each student. Such tests should therefore ideally be suitable for all students in a class. In the context of adaptive testing, this means that the item pool should ideally contain enough items for each performance level. This allows all students in a class, regardless of their abilities, to be measured accurately and effectively. In the context of DBDM, it is particularly important that there are enough easy items in the item pool. To provide data-based resources and appropriate instructions, weak students who need these resources must be identified early and accurately. Assessments that aim to achieve this must therefore be particularly sensitive in the low-performance range. An item pool that contains enough easy items is helpful here and expands the measuring instrument to lower performance levels without having to resort to test modifications.

In the case of the reading screening example, item pools consisting of 30 to 52 items were used, which may be considered small compared to the recommendations of Stocking (1994) and Way (2005). However, it is important to note that this screening CAT was designed for a one-time use, rather than as a formative assessment. In line with DBDM, a screening can be employed to identify weak students as early as possible. Subsequently, an appropriate step-by-step support should be assigned that corresponds as closely as possible to the measured areas of ability. Progress monitoring measurements can be used during the support period as a complement to the screening process, in order to further track the learning progress of weaker students identified by the screening and to make necessary adjustments to the support provided. In situations where tests are frequently repeated and items should not be repeated, larger item pools may be more beneficial. A larger item pool can provide more options for selecting items of appropriate difficulty levels and prevent items from being repeated. However, creating a larger item pool may require additional resources and effort. Therefore, the decision to use a smaller or larger item pool depends on the specific needs and goals of the adaptive test. In the case of the reading screening, a smaller item pool was deemed appropriate for a one-time use, and the item pool was carefully selected and calibrated to ensure accurate measurement of students' reading abilities.

Since the reading screening can also be conducted as a non-adaptive test, it is not problematic if all items are used in this case. In the non-adaptive screening, students frequently complete all items in the pool successfully. If the length of the screening allows for non-adaptive measurement, it would also be acceptable for adaptive testing if all items in the pool were used, resulting in the same test length. However, this becomes more problematic when using a significantly larger item pool with more than 100 items, as exemplified by the generated item pool. If a pool is large but not appropriately distributed, it is likely that the difficulty of the items in the pool would not be sufficient to measure all individuals with the desired accuracy (Reckase, 2003; Ebenbeck & Gebhardt, submitted). If the easiest or hardest item cannot provide enough information, there would be no more items that could give additional insight. Instead, the entire remaining item pool would need to be used again, resulting in a considerably longer measurement with a larger item pool, which may be difficult to justify in everyday teaching. Therefore, it is necessary to create not only a large but also a well-distributed item pool and to simulate in advance whether the pool is suitable for testing with the desired accuracy within a reasonable length.

From a pedagogical standpoint, drawing the entire item pool because the desired test accuracy cannot be achieved is problematic, especially for students who have lower performance. If they struggle with the easiest item and a few of the following items, then all remaining items in the pool will be drawn, and these items will inevitably be even more challenging, making it unlikely that they will be answered correctly. This large number of items that are too difficult, and becoming increasingly difficult without the possibility of quitting, is not conducive to use in schools due to motivation issues. On the other hand, students with high abilities may receive many items that are too easy, which may also be detrimental to their learning. The same issue of the item pool and the associated adaptive drawing - i.e., the lack of items at the appropriate difficulty level - would lead to a high proportion of correctly solved items for high-ability students and a high proportion of incorrectly solved items or aborted testing for low-ability students. This demonstrates the necessity of involving the target group in the development of an assessment in general and an adaptive test in particular. Therefore, it is important to create a well-designed and appropriately distributed item pool that meets the needs of all students, regardless of their ability level.

All three studies have demonstrated that using a large item pool with evenly distributed item difficulties can address the issue of item selection and improve the accuracy of test results. Therefore, a large item pool is more suitable for test administration from both methodological and pedagogical perspectives.

However, developing a large item pool presents challenges, especially when designing tests for primary students and those with SEN. Obtaining a large number of items requires more instructional and research time, which may not be feasible for some schools and students (Mills & Stocking, 1996). Additionally, some students may be unable to complete the entire item pool due to their individual circumstances, which could lead to a loss of focus and potentially distort the results. One solution is to work on only parts of the item pool in several sessions or to work with a much larger sample and work on parts of the item pool at a time. In general, it is essential to strike a balance between research effort, feasibility for schools and students, and the desired test outcomes.

5.1.2 Starting the CAT

The adaptive test begins by presenting the first item to the student, and while the specific item chosen has only a minor influence on the rest of the test, it can result in slight variations in the average length of the resulting CAT, as demonstrated by the findings from study 2 and study 3. Typically, a common approach is to use an item with a severity of $\sigma = 0$ as the first item. study 2's results show that this choice results in shorter tests compared to using Bayesian estimators that incorporate prior information. However, study 3 found that using an easier first item with a σ of -1 compared to a medium difficulty item also leads to slightly shorter measurements. These results have been shown to be consistent across different item pools and groups of subjects.

Despite initial expectations based on the work of Weiss and Kingsbury (1984), incorporating previously known information about a student may not improve the accuracy of subtest measurements. The reasons for this outcome are not clear. Study 2 ruled out the possibility that the lack of added value was due to a low correlation between subtests, as there was no additional test shortening or improved accuracy, even when an individual had identical θ scores in all four subtests. The use of a larger item pool with evenly distributed item difficulties also cannot account for this result, as study 2 did not show any differences between the various item pools. Therefore, it is unclear in what circumstances using a Bayesian estimator to select the first item could enhance accuracy over using a fixed first item. Based on the findings, this approach has been abandoned for the adaptive reading screening. However, it is likely that the algorithm determining the next items has a much greater influence on the test than the first item, so the starting rule has less impact on the course of testing.

As an alternative to using previous information, it is possible to use a simpler first item for developing a CAT suitable for heterogeneous student groups from both methodological and pedagogical perspectives. Methodologically, using an easier first item could shorten the length of almost every subtest (except subtest 3) while maintaining accuracy. The first item is fixed from the start, so its suitability can be specifically checked and selected. This approach provides a common starting point for all students, enabling later examination of the branches of the CAT. From a pedagogical perspective, starting with an easier item can lead to a more positive test experience compared to starting with a moderately difficult item. This is because the probability of correctly answering the first item is higher with an easier item. Therefore, students are more

likely to have a sense of achievement at the beginning and approach the rest of the test with a positive attitude.

In general, there are no benefits to using a moderately difficult first item over an easier one. Implementing and developing a CAT with an easier first item is less complex compared to using a moderately difficult first item. Unlike using a Bayesian estimator at the start of a test, selecting an easier first item can be easily implemented without saving test results for use in the subsequent subtests.

5.1.3 Stopping the CAT

During the simulations conducted in studies 1 to 3, it became evident that the termination of the CAT is influenced not only by the assumed stopping rule itself but also by the item pool of the CAT. To determine the stopping rule, a precision criterion was used, which was based on a target *SE*, following the theoretical foundation provided by Dodd et al. (1993) and Stafford et al. (2019). The findings of Michiel et al. (2008) and Ebenbeck and Gebhardt (2022) are consistent with the current study, as a higher *SE* resulted in a shorter test length but also led to less accurate measurements. Therefore, it is crucial to strike a balance between test length and measurement accuracy when selecting a stopping rule.

Michiel et al. (2008) used a small item pool similar to the one used in this study. With $SE(\theta) < 0.5$, they achieved an average test length reduction of 81% (29 items less). In comparison, for the adaptive reading screening, test length reductions between 30% (9.21 items, subtest 3) and 63% (34.66 items, subtest 2) were achieved. However, with a larger and evenly distributed item pool of 100 items, on average, up to 85% of the test length could be reduced. This demonstrates that the potential for test reduction is dependent on the original item pool and its distribution. Michiel et al. (2008) achieved an accuracy of $r = .895$ with their stopping rule. In comparison, both the adaptive reading screening and the generated item pool perform better in most cases, with accuracies ranging from $r = 0.90$ to $r = 0.94$. Therefore, the balance between test length and test accuracy should be carefully considered. Nonetheless, using a generated item pool, a correlation of $r > 0.90$ was achieved for students with SEN, despite a very lenient stopping rule of $SE(\theta) = 0.5$. This suggests that while a higher *SE* leads to shorter testing and affects accuracy, the average loss of accuracy is small enough that it is more beneficial to argue for a shorter test length rather than a lower *SE*.

However, when dealing with large item pools, an additional stopping rule may be considered. The accuracy rule of $SE(\theta) = 0.5$ alone may not be sufficient to prevent the possibility of an insufficient item pool for very low or high θ values. As mentioned earlier, in such cases, the entire remaining item pool is drawn upon, which is not ideal from a pedagogical perspective. Hence, there is a need to explore various solutions to address this issue, such as:

On the one hand, introducing a maximum length for the CAT can be a potential solution for larger item pools. Ebenbeck and Gebhardt (2022) and Ludewig et al. (2022) have demonstrated the effectiveness of this approach. The advantage of using a fixed-length CAT is that it allows for better planning in the school environment (Ludewig et al., 2022). Teachers would know in advance that the test will stop after a predetermined number of items. A suitable number of items can be selected to ensure that the student is not overburdened. If the test is too easy or too difficult, or if there are not enough easy or difficult items available for measurement, it will stop automatically after reaching the predetermined maximum length instead of presenting all items. However, the disadvantage of this approach is that the maximum length must be defined in advance, and there are no uniform criteria to determine it. Ludewig et al. (2022), for example, used the available time in a school lesson and the percentage of tasks that can be successfully completed by 80% of students in a school lesson as criteria, as well as the number of dimensions within the test. Such criteria could be used to define a maximum length for a reading screening test as well. On the other hand, introducing a maximum length does not prevent students with weak abilities from being overwhelmed by the items before reaching the maximum length and still having to work through a certain number of items. Therefore, the maximum length should not be set too low, as it may impede the adaptive selection algorithm and the accuracy rule from working effectively on the test.

Additionally, a minimum criterion could be introduced to the CAT to address particularly low-performing students in a class. This criterion would come into play if a student cannot answer the e.g. three easiest items in the item pool correctly. In this scenario, the test would be automatically terminated without any further items being drawn. The minimum criterion would act as an additional stop for students with very low abilities who may find the subtest too difficult. For instance, if a student is struggling to answer the easiest items, it's highly unlikely they would be able to answer more difficult items accurately. From a scoring perspective, this minimum criterion would have no negative consequences. The student would only be scored on the items they were able to answer, and the test would end early, reducing the burden on the student.

However, it's important to ensure that the minimum criterion is set at an appropriate level that does not unfairly disadvantage students with certain abilities.

The purpose of the screening CAT is to identify students who require additional support based on their performance regardless of their SEN status. This is particularly important for students with low abilities, where the specific value of their ability may be less important than the fact that they require additional support. In such cases, it may be more relevant to define a threshold at which students should be assigned further support. If the minimum criterion is applied, it should be assumed that the student in question needs additional support, regardless of their exact ability level. However, incorporating this stop criterion was not possible in this study due to technical limitations with the R packages used. Nonetheless, it is an important consideration for future implementation of the reading screening. By setting a minimum criterion, educators can quickly identify students who require further support, without having to go through the entire CAT process. This can be particularly helpful for students who may become overwhelmed by the CAT.

In summary, it is recommended to set a fixed starting item with $\sigma \sim -1$ for each subtest in the adaptive version of the reading screening LES-IN-DIG. To estimate θ after each student's response, it is recommended to use the proven combination of ML and subsequent Bayesian estimator, once no fully correct or incorrect response patterns are left. The next item should be selected based on MFI. The CAT should be stopped if one of the following rules is met: achieving a CAT accuracy of $SE(\theta) = 0.5$, reaching a maximum length (the specific length per subtest needs to be determined through further research), or when no more easy items can be drawn. It is important to expand the item pools with more items before implementing the screening, especially for subtest 3. This should not only include more items generated by the existing difficulty-generating item characteristics but also potentially more easy and difficult items. Ideally, this would result in a subtest-specific item pool with evenly distributed σ. This approach will improve the accuracy and efficiency of the screening process, providing a fair and reliable assessment of the students' reading abilities, which can be used to assign appropriate support and intervention.

5.2 Chances of Adaption for Students with SEN

5.2.1 Computerized Adaptive Testing

Early intervention is crucial for providing support to students with SEN. Therefore, a key goal of special education and inclusive education is to identify individual students' problems before they develop into persistent difficulties. Standardized measurement instruments, particularly in the context of DBDM, provide reliable information on students' strengths and weaknesses (Lai & Schildkamp, 2013), which is essential for students who have difficulties in certain academic areas and require additional support to develop further.

For students with SEN, assessments therefore play a particularly important role, and screenings must accurately and reliably identify students with problems so that appropriate support can be provided based on the results (Jacobson, 1999). Against this background, the potential benefits and limitations of adaptive tests for students with SEN can be discussed.

Adaptive tests are particularly reliable and suitable for students with SEN as they can adjust to the students' abilities. The results of the studies in this work show that the adaptive algorithm works for this student group, and the difficulty level of the proposed items gradually adjusts to the student's ability level. As the test progresses, students are more likely to receive items that match their ability level, rather than increasingly difficult items, as is the case with many other testing procedures.

Also, CAT addresses the areas where students with SEN have difficulties. The results of these and other studies (e.g. Boudreau & Hedberg, 1999; Di Blasi et al. 2019; Gilmour et al., 2019; Gresham et al., 1996) show that students with SEN often exhibit weaker academic performance than their peers without SEN. They make more mistakes and work more slowly, which means they complete fewer tasks in the same amount of time (Ebenbeck et al. 2023). Adaptive tests shorten the testing time for students with SEN more than for students without SEN. This could be considered as an additional support for their special needs and working behavior. The required concentration time is thus also reduced, and frustration during the measurement could presumably remain relatively low, since the test ideally should not become much too difficult, but rather gradually adapts to the student's ability. Since the testing time is shortened and students with SEN complete fewer items overall than in a non-adaptive test, they also make fewer mistakes overall. Study 3 shows that up to two-thirds fewer items are answered incorrectly in total. Although the percentage of incorrectly answered items does not change, since fewer items

are processed overall, fewer items are answered incorrectly. Therefore, the student spends as little time as possible on items they cannot or can only solve with difficulty, and the measurement remains as efficient as possible.

Adaptive tests lead to fairer measurements. Research shows that teachers may assess the performance of low-performing students less accurately than the performance of high-performing students (Begeny et al., 2008; Coladarci, 1986, p. 144). CBTs, in general, can improve psychometric quality and thus lead to fairer measurements for students (Liebers et al., 2019). CATs are conducted digitally and therefore have all the advantages of CBTs. However, teachers are even more guided in the implementation, as the algorithm takes over the adaptation of the test difficulty. The studies of this work show that measurements of $r > 0.9$ for students with SEN can often be observed. The accuracy of the measurements in a CAT can therefore be considered very good, allowing teachers to rely on the results and use them for further development planning. Since the reading CAT measurements can be implemented as class-based testing on tablets, testing can also be time- and resource-saving. The teacher can measure several students in the same amount of time. Since students work on tablets and do not require individual support, the teacher can also provide general support during the measurement and switch between students for advice, support, and assistance.

Overall, the simulated CATs show that they do not disadvantage students with SEN, but instead offer more advantages in terms of accuracy, test length, administration, and fairness. With suitable item pools, CAT measurement has the potential to measure and identify such students who need to be identified accurately and quickly. As with all CBTs, it must be clarified in advance to what extent the students are familiar with working on a PC or tablet. In general, technical equipment and technical knowledge are no longer a problem for students in a CAT.

5.2.2 Machine Learning Based Adaptive Testing

Taken together, the results of these studies show the high utility of adaptive testing systems for students with SEN and low performance levels in general. CAT primarily refers to an adaptation of item difficulty. The immediate previous answers of the examinees and the pre-defined and calibrated item parameters serve as the basis for adaptation. These studies have found that CAT provides added value for assessment. Therefore, it is likely that further developments of CAT can offer a similar, if not higher, added value.

With CAT, adaptation is essentially influenced by the two aforementioned factors. Therefore, the type of adaptation is naturally limited. In comparison, AI-based tools and procedures are becoming increasingly easy to implement and more practical for everyday use. Machine learning stands out particularly in the context of educational technology and adaptive assessments (Hilbert et al. 2021, Korkmaz & Correia, 2019, Nafea, 2017, Zheng et al. 2020). Unlike CAT, machine learning works not only with the immediately given answers of the examinees and statistical ability estimation, but also with (response) patterns. Thus, entire answer patterns can serve as a basis for adaptations. AI-based adaptive tests can also learn from answer patterns of examinees who have taken the test before. This knowledge can be incorporated into adaptations to achieve more accurate adjustments (McCusker, 2013, Shute & Zapata-Rivera, 2007). It is conceivable that not only correctly and incorrectly answered tasks are included in the adaptation, but also other characteristics of the student, such as their previous experiences, age, or SEN status.

Another addition provided by AI-based adaptive testing is the provision of individual feedback for the examinees in learning environments (Afzaal et al., 2021, Asthana & Hazela, 2020) as well as the prediction of response patterns (Böhme et al. 2022). This cannot be achieved by CAT, but could be particularly valuable for low-performing students. An expansion of CAT with machine learning would therefore also be suitable for enabling, improving, and integrating adaptive testing for students with SEN into classroom instruction. Furthermore, such individual feedback can also provide added value for teachers. Based on the response patterns of the students, the adaptive test can derive support recommendations and suggest existing support materials for the teacher. The results of the support materials can in turn be used as input for further measurements. This creates a symbiosis between assessment and instruction that can be used for further data-based support decisions.

5.2.3 Need for Digitalization

In order to implement and integrate adaptive testing of any form in the classroom, a progressive digitization of schools is necessary. CATs can be implemented as software or web-based applications. For the reading screening discussed in this work, a web-based implementation is planned. However, digital devices are necessary to use the adaptive test along with the CAT algorithm. Ideally, there are enough high-performance tablets in the school to enable the use of adaptive tests.

The use of computers and tablets in assessment is no longer a hurdle for students (Jeong, 2014; McClelland & Cuevas, 2020). Additionally, touch-based implementations on tablets can make the process easier and more intuitive and barrier-free for students while ensuring higher psychometric quality and less experimenter effects (Liebers et al., 2019, Schaper, 2009, p. 26; Walter & Schuhfried, 2004). However, it is necessary for teachers to have comprehensive digital competencies. They must be able to handle digital devices to assist when necessary. They should also be trained to conduct and interpret the adaptive test. A basic understanding of how the adaptive test works additionally facilitates the implementation and interpretation of results.

In summary, digital methods offer a great added value for assessment and DBDM in schools. To take advantage of these benefits, a technical upgrade of schools and corresponding training of teachers and student teachers is necessary to be able to also use new developments in the field, such as CAT or AI, productively.

5.3 Limitations

The limitations of this study concern the study design as a simulation study, the selected sample of students with SEN from different grades, the chosen IRT model, and the final scoring of the CAT.

In CAT studies, it is common to use simulations to ensure a resource-efficient development and exploration. However, simulation studies have inherent limitations because their quality depends on the quality of the input data, and they cannot account for unforeseen factors. They make simplifying assumptions and cannot predict, for example, how a student would behave in a testing atmosphere. Therefore, simulation studies can only make assumptions, limiting the generalizability of the results and not predicting the accurate behavior of an individual in the real world. Additionally, in these studies, only the estimated θ in the Rasch model was selected as input, and answer patterns were estimated based on it, which adds another factor of uncertainty. This approach had to be chosen due to missing values in the data. However, this means that the actual answer patterns of the students are missing in the simulations. With a complete survey of all items, simulations with existing patterns would also be possible, which would represent the likely performances of the students in a CAT with even higher predictability. Despite these limitations, simulation studies offer a cost-effective, risk-free, fast, and controllable opportunity for data-based prediction and are therefore often used in the context of CAT.

This study is the first to address CAT in inclusive education, specifically focusing on students with SEN-L, SEN-I and SEN-S. This student population was chosen partly due to its high prevalence in inclusive settings and partly due to its disadvantage in conventional testing systems. Limitations of the study, therefore, arise from the targeted student body: students with other SEN, such as physical-motor disabilities, blindness or deafness, were not included in the studies. The author acknowledges that this student body also plays a significant role in inclusive classes and must be considered for a test as accessible as possible.

However, it can be assumed that CAT also offers opportunities for students with other special educational needs. Şenel and Kutlu (2017) have already shown that the technology can also be used for assessment for blind students. Since CAT primarily refers to adaptation based on item parameters, students with performance deficits in academic areas are more likely to be the target group - this also applies to individuals with physical special needs, depending on the level of their performance spectrum. However, there are also additional questions that arise, especially in the area of assistive technologies, to make testing with CAT on a computer profitable and without disadvantages in implementation.

This also includes students with severe intellectual disabilities. In order to carry out a CAT, a few points need to be clarified in advance for students with severe intellectual disabilities. Conducting the test on the computer could pose challenges for students with severe intellectual disabilities that need to be individually assessed and addressed. However, in such cases, it would be possible to assist the student in operating the computer or use the computer as an assistant. The extent to which such assistance would affect test quality would need to be clarified on a case-by-case basis. In addition, as with any test, it must be clarified in advance whether the skills being measured by the test can actually be measured for students with severe intellectual disabilities. This is evident in the example of reading screening: the CAT technology itself does not pose a limitation for students with severe intellectual disabilities, but rather the ability being measured. However, a fully inclusive instrument should be able to assess all students. Therefore, expanding the measured areas would be worth pursuing in further research.

The sample includes students from different grade levels. Although their reading abilities show significant differences in the analog reading screening, it was not separately investigated to what extent they would perform differently in the CAT. It can be assumed that a similar pattern would emerge as in the differentiation between students with and without SEN. Lower grade students with lower performance would also be more accurately and briefly assessed than

higher grade students. Students in secondary school classes for intellectual development show a very heterogeneous pattern and would, therefore, receive different CATs based on their individually developed reading abilities. An analysis by grade level was not conducted as the broad composition of grade levels aimed to represent a high degree of performance heterogeneity among students in the sample.

One further limitation of the studies in this work is their generalizability to CATs based on other IRT models. In this case, only the adaptation of an item pool with one item parameter was used. Besides item difficulty, no other parameter was included and manipulated as intended in the Rasch model. The range of IRT models is wide, and the unidimensional Rasch model can be considered the psychometrically simplest model. Other IRT models include additional dimensions or item parameters per item. Both the dimensions and additional item parameters would need to be taken into account if such a model were to be used as input for CAT. The studies presented here can provide statements on CATs in heterogeneous learning groups based on the Rasch model. However, they cannot make generalized statements about CATs based on other IRT models.

One further limitation of this work is that no statements can be made about an appropriate scoring of the CATs. As justified in the theory, it is possible to represent the students' results in terms of their θ values on a logit scale, but teachers would likely need to be guided in their interpretation. However, other possibilities for scoring and representation were not explored in these studies.

5.4 Future Research Possibilities

This work provides a basis for demonstrating more accurate and shorter testing through CAT for students with SEN. From these results, limitations, and discussion points, further opportunities for future research arise to shed more light and explore the field of adaptive testing in inclusive and special education. Future research in this field should focus on addressing these limitations and exploring new avenues for the application of adaptive testing in special education. By continuing to develop and refine adaptive testing methods and tools, one can provide more accurate, efficient, and personalized assessments and interventions for students with SEN, thereby promoting their academic and social success.

5.4.1 Implementation of the Adaptive Reading Screening

The results of the simulation studies should be validated in practice to ensure their reliability. To achieve this, the digital reading screening tool LES-IN-DIG (Ebenbeck et al., 2023), which was used as input for the CAT simulation studies in this work, will be further developed into an adaptive reading screening tool. Possible adaptations of the item pool for this step have already been discussed. The screening will be equipped with a suitable adaptive algorithm and published on the online platform www.levumi.de after these adaptations. Teachers can use the screening with their students freely through the platform.

After the implementation of the adaptive screening, the screening itself can be studied in practice. Different grade levels of students should be tested with the adaptive screening to verify its test quality. The performance of the screening in the lower performance range of students with SEN, especially students with SEN-L and SEN-I, is particularly interesting. These students represent the lower end of the ability range being measured and therefore are the target group for the adaptive screening. A separate study based on these results is necessary to examine the performance of the newly developed instrument. In this study, not only the accuracy and test length but also the response patterns per subtest of the screening should be evaluated to obtain detailed information about the performance of CAT as an assessment for students with SEN in practice.

Furthermore, it is important to examine test motivation and test anxiety, especially among students with SEN (Stone & Davey, 2011), as these topics were not yet addressed. While the results indicate no effect on psychological aspects for the majority of students, a follow-up study should investigate whether this also applies to students with SEN, who generally have more emotional well-being problems. Finally, there is potential to test and explore the findings of this study on CATs for students with SEN using additional IRT models, as well as for other learning areas in early childhood, school, and extracurricular contexts, to provide a comprehensive picture of this topic.

5.4.2 Adaptive Testing in the Classroom

In addition to examining the test quality for different student groups, the use of adaptive tests in the classroom should be investigated. Of particular interest is how adaptive testing is

integrated into DBDM and what conclusions can be drawn from the link between adaptive testing and tailored support.

Closely linked to this is the consideration of adaptive teaching, in which learning content and processes are tailored to the individual needs of each student. Adaptive tests can be a part of adaptive teaching, as they provide a foundation for DBDM in this teaching setting. The results of adaptive tests can be used by the teacher to adjust their teaching, learning content, and teaching methods to benefit all students. The success of these adjustments can be verified and monitored again with adaptive formative or summative tests.

6 Conclusion

From its early days as a theoretical concept to its current role as a powerful tool for educators, CAT has the potential to revolutionize assessments in inclusive education. This innovative approach to testing allows educators to assess students' knowledge and abilities quickly and accurately in a way that is both engaging and effective.

As demonstrated in the literature and confirmed by the results of this study, CAT leads to shorter assessments for students with and without SEN, with students with SEN particularly benefiting from this type of testing. With a suitable item pool, students with SEN may be measured more accurately in a shorter time than students without SEN. Therefore, this technology is particularly suitable for assessments in heterogeneous learning groups in inclusive and special education, where traditional one-size-fits-all assessments may not accurately capture students' abilities.

Moreover, adaptive testing has the potential to provide a more personalized approach to teaching and learning. By identifying students' strengths and weaknesses, educators can tailor their instruction to meet the needs of individual learners. This can lead to more effective instruction and improved academic outcomes for all students, especially those with SEN.

In addition, adaptive testing can help support the development of tools that provide timely and useful feedback to both students and educators. By leveraging the data generated by adaptive testing, developers can create feedback tools that provide targeted and actionable information to help guide students' learning and support their academic growth. For example, educators can use data from adaptive tests to identify specific areas where students may need additional support or instruction, or to highlight areas where students have excelled and should be challenged further.

However, for the potential of CAT to be fully realized, several prerequisites must be met. First and foremost, schools must have adequate technological resources to conduct CAT. This includes hardware, software, and network infrastructure that are compatible with CAT systems. Additionally, teachers must be willing to embrace new technologies and overcome any fear or aversion to them. Teacher education programs can help support educators in learning how to use and implement adaptive testing effectively.

On the other hand, CAT systems must be designed to be accessible (e.g. as Open Educational Resource), flexible, and cost-effective. They should be customizable to meet the specific needs of different subject areas, grade levels, age groups, and abilities. CAT developers should also ensure that their systems are evidence-based and scientifically validated to produce accurate and reliable results. Students with SEN should be considered during this development process to ensure the fairness and suitability of the CAT for all students. This will help to increase their acceptance and adoption by teachers and principals and ensure that they provide a meaningful value-add to the assessment process.

In conclusion, CAT has the potential to provide a more accurate, efficient, and engaging way to assess students' abilities.

Its implementation requires careful consideration and further research to ensure its accessibility, fairness, and effectiveness for all students, especially those with SEN. With the right conditions and support, CAT may help to create more inclusive and equitable learning environments that promote academic achievement and success for all students.

7 References

Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., ... & Weegar, R. (2021). Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation. Frontiers in Artificial Intelligence, 4, 723447.

Akhtar, H., Vekety, B., & Kovacs, K. (2022). The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis. *Assessment*, *0*(0). https://doi.org/10.1177/10731911221100995

Amelang, M., & Zielinski, W. (1997). *Psychologische Diagnostik und Intervention* (2., korrigierte, aktualisierte und überarbeitete Auflage). *Springer-Lehrbuch*. Springer. https://doi.org/10.1007/978-3-662-22370-3

American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders: Dsm-5-TR™* (Fifth edition, text revision). American Psychiatric Association Publishing.

An, X., & Yung, Y. F. (2014). *Item Response Theory: What It Is And How You Can Use The IRT Procedure to Apply It.*

Andersen, E. B. (1973). A Goodness of Fit Test for the Rasch Model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180

Arishi, A., Boyle, C., & Lauchlan, F. (2017). Inclusive Education and the Politics of Difference: Considering the Effectiveness of Labelling in Special Education. *Educational and Child Psychology*. https://ore.exeter.ac.uk/repository/handle/10871/29377

Arvidsson, P., & Granlund, M. (2018). The Relationship Between Intelligence Quotient and Aspects of Everyday Functioning and Participation for People Who Have Mild and Borderline Intellectual Disabilities. *Journal of Applied Research in Intellectual Disabilities*, *31*(1), e68-e78. https://doi.org/10.1111/jar.12314

Asbury, K., Fox, L., Deniz, E., Code, A., & Toseeb, U. (2021). How is COVID-19 Affecting the Mental Health of Children with Special Educational Needs and Disabilities and Their Families? *Journal of Autism and Developmental Disorders*, *51*(5), 1772–1780. https://doi.org/10.1007/s10803-020-04577-2

Asthana, P., & Hazela, B. (2020). Applications of machine learning in improving learning environment. Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions, 417-433.

Babcock, B., & Weiss, D. J. (2009). Termination Criteria in Computerized Adaptive Tests: Variable-length CATs are not Biased. In D. J. Weiss (Chair), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Badian, N. A. (1999). Persistent Arithmetic, Reading, or Arithmetic and Reading Disability. *Annals of Dyslexia*, *49*(1), 43–70. https://doi.org/10.1007/s11881-999-0019-8

Baghaei, P. (2008). Local Dependency and Rasch Measures. *Rasch Measurement Transactions*, *21*(3), 1105–1106. https://www.rasch.org/rmt/rmt213b.htm

Beaujean, A. A., Benson, N. F., McGill, R. J., & Dombrowski, S. C. (2018). A Misuse of IQ Scores: Using the Dual Discrepancy/Consistency Model for Identifying Specific Learning Disabilities. *Journal of Intelligence*, *6*(3). https://doi.org/10.3390/jintelligence6030036

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' Perceptions of Students' Reading Abilities: An Examination of the Relationship Between Teachers' Judgments and Students' Performance Across a Continuum of Rating Methods. *School Psychology Quarterly*, *23*(1), 43–55. https://doi.org/10.1037/1045-3830.23.1.43

Bender, W. N., & Smith, J. K. (1990). Classroom Behavior of Children and Adolescents with Learning Disabilities: A Meta-Analysis. *Journal of Learning Disabilities, 23*(5), 298–305. https://doi.org/10.1177/002221949002300509

Bennett, R. E. (2002). Inexorable and Inevitable: The Continuing Story of Technology and Assessment. *Journal of Technology, Learning, and Assessment, 1*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1667/1513

Betz, N. E. (1977). Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance. *Applied Psychological Measurement, 1*(2), 259–266. https://doi.org/10.1177/014662167700100212

Blumenthal, S., Blumenthal, Y [Y.], Lembke, E. S., Powell, S. R., Schultze-Petzold, P., & Thomas, E. R. (2021). Educator Perspectives on Data-Based Decision Making in Germany and the United States. *Journal of Learning Disabilities, 54*(4), 284–299. https://doi.org/10.1177/0022219420986120

Bock, R., & Gibbons, R. (2021). *Item Response Theory* (1st edition). Wiley, Safari.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and Performance Differences among Computer-Based and Paper-Pencil Tests. *Journal of Educational Computing Research, 31*(1), 51–60. https://doi.org/10.2190/GRQQ-YT0F-7LKB-F033

Böhme, R., Coors, S., Oster, P., Munser-Kiefer, M., & Hilbert, S. (2022). Machine learning for spelling acquisition-How accurate is the prediction of specific spelling errors in German primary school students?.

Botting, N. (2005). Non-Verbal Cognitive Development and Language Impairment. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 46*(3), 317–326. https://doi.org/10.1111/j.1469-7610.2004.00355.x

Bouck, E. C., & Satsangi, R. (2015). Is There Really a Difference? Distinguishing Mild Intellectual Disability from "Similar" Disability Categories. *Education and Training in Autism and Developmental Disabilities, 50*(2), 186–198.

Boudett Parker, K., City, E. A., & Murnane, R. J. (2006). The "Data Wise" Improvement Process. Eight Steps for Using Test Data to Improve Teaching and Learning. *Harvard Education Letter, 22*(1). https://www.nesacenter.org/uploaded/conferences/FTI/2011/handouts/RussellDataWiseArticle.pdf

Boudreau, D. M., & Hedberg, N. L. (1999). A Comparison of Early Literacy Skills in Children With Specific Language Impairment and Their Typically Developing Peers. *American Journal of Speech-Language Pathology, 8*(3), 249–260. https://doi.org/10.1044/1058-0360.0803.249

Bowen, S. K., & Rude, H. A. (2006). Assessment and Students with Disabilities: Issues and Challenges with Educational Reform. *Rural Special Education Quarterly, 25*(3), 24–30.

Brown, J. M., & Weiss, D. J. (1977). *An Adaptive Testing Strategy for Achievement Test Batteries.* Minnesota University, Minneapolis Departement of Psychology.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The Four Generations of Computerized Educational Measurement. *ETS Research Report Series, 1988*(1), i-148. https://doi.org/10.1002/j.2330-8516.1988.tb00291.x

Bundschuh, K., & Winkler, C. (2019). *Einführung in die Sonderpädagogische Diagnostik* (9., überarbeitete Auflage). *utb Sonderpädagogik/Pädagogische Psychologie: Vol. 999.* Ernst Reinhardt Verlag. https://doi.org/Konrad

Calhoon, M. B., Fuchs, L. S., & Hamlett, C. L. (2000). Effects of Computer-Based Test Accommodations on Mathematics Performance Assessments for Secondary Students

with Learning Disabilities. *Learning Disability Quarterly*, *23*(4), 271–282. https://doi.org/10.2307/1511349

Carrington, S., Tangen, D., & Beutel, D. (2019). Inclusive Education in the Asia Indo-Pacific Region. *International Journal of Inclusive Education*, *23*(1), 1–6.

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, *71*(5). https://doi.org/10.18637/jss.v071.i05

Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H. (2000). A Comparison of Item Selection Rules at the Early Stages of Computerized Adaptive Testing. *Applied Psychological Measurement*, *24*(3), 241–255. https://doi.org/10.1177/01466210022031705

Choppin, B. (1982). *A Fully Conditional Estimation Procedure for Rasch Model Parameters*. https://eric.ed.gov/?id=ed228267

Claessen, M., Leitão, S., Kane, R., & Williams, C. (2013). Phonological Processing Skills in Specific Language Impairment. *International Journal of Speech-Language Pathology*, *15*(5), 471–483. https://doi.org/10.3109/17549507.2012.753110

Cohen, D., Rivière, J. P., Plaza, M., Thompson, C., Chauvin, D., Hambourg, N., Lanthier, O., Mazet, P., & Flament, M. (2001). Word Identification in Adults with Mild Mental Retardation: Does IQ Influence Reading Achievement? *Brain and Cognition*, *46*(1-2), 69–73. https://doi.org/10.1016/S0278-2626(01)80037-3

Coladarci, T. (1986). Accuracy of Teacher Judgments of Student Responses to Standardized Test Items. *Journal of Educational Psychology*, *78*(2), 141–146. https://doi.org/10.1037/0022-0663.78.2.141

Coloma, C. J., Silva, M., Palma, S., & Holtehuer, C. (2015). Reading Comprehension in Children With Specific Language Impairment: An Exploratory Study of Linguistic and Decoding Skills. *Psykhe*, *24*(2), 1–8. https://doi.org/10.7764/psykhe.24.2.763

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). Drc: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037/0033-295x.108.1.204

Conole, G., & Warburton, B. (2005). A Review of Computer-Assisted Assessment. *Research in Learning Technology*, *13*(1), 17–31. https://doi.org/10.1080/0968776042000339772

Cornoldi, C., Giofrè, D., Orsini, A., & Pezzuti, L. (2014). Differences In The Intellectual Profile of Children with Intellectual vs. Learning Disability. *Research in Developmental Disabilities*, *35*(9), 2224–2230. https://doi.org/10.1016/j.ridd.2014.05.013

Cortiella, C., & Horowitz, S. H. (2014). *The State of Learning Disabilities: Facts, Trends and Emerging Issues*. National Center for Learning Disabilities.

Cullinan, D., Epstein, M. H., & Lloyd, J. (1981). School Behavior Problems of Learning Disabled and Normal Girls and Boys. *Learning Disability Quarterly*, *4*(2), 163–169. https://doi.org/10.2307/1511001

Demetriou, K. (2020). Special Educational Needs Categorisation Systems: To Be Labelled or Not? *International Journal of Disability, Development and Education*, *69*(5), 1772–1794. https://doi.org/10.1080/1034912X.2020.1825641

Department for Education and Science. (1978). *Special Educational Needs: Report of the Committee of Enquiry into the Education of Handicapped Children and Young People (The Warnock Report)*. Her Majesty's Stationary Office.

Di Blasi, F. D., Buono, S., Cantagallo, C., Di Filippo, G., & Zoccolotti, P. (2019). Reading Skills in Children with Mild to Borderline Intellectual Disability: A Cross-Sectional

Study on Second to Eighth Graders. *Journal of Intellectual Disability Research*, *63*(8), 1023–1040. https://doi.org/10.1111/jir.12620

Dodd, B. G., Koch, W. R., & Ayala, R. J. (1993). Computerized Adaptive Testing Using the Partial Credit Model: Effects Of Item Pool Characteristics and Different Stopping Rules. *Educational and Psychological Measurement*, *53*(1), 61–77. https://doi.org/10.1177/0013164493053001005

Dworschak, W., & Kölbl, S. (2022). Adaptives Verhalten. Zur Bedeutung eines (zu) wenig beachteten Konstrukts im Kontext geistiger Behinderung aus diagnostischer Sicht. In M. Gebhardt, D. Scheer, & M. Schurig (Eds.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung. Version 1.0* (pp. 175–190). Universität Regensburg. https://doi.org/10.5283/epub.53149

Earle, F. S., Gallinat, E. L., Grela, B. G., Lehto, A., & Spaulding, T. J. (2017). Empirical Implications of Matching Children With Specific Language Impairment to Children With Typical Development on Nonverbal IQ. *Journal of Learning Disabilities*, *50*(3), 252–260. https://doi.org/10.1177/0022219415617165

Ebenbeck, N. (2023). Computerized Adaptive Testing in Inclusive Education. Data and Syntax. https://doi.org/10.17605/OSF.IO/7JEUS

Ebenbeck, N., & Gebhardt, M. (submitted). Development of Computerized Adaptive Tests for Low-Performing School Students.

Ebenbeck, N., & Gebhardt, M. (2022). Simulating Computerized Adaptive Testing in Special Education based on Inclusive Progress Monitoring Data. *Frontiers in Education*, *7*, Article 945733. https://doi.org/10.3389/feduc.2022.945733

Ebenbeck, N., Jungjohann, J., Mühling, A., & Gebhardt, M. (2023). Die Bearbeitungsgeschwindigkeit von Kindern mit Lernschwierigkeiten als Grundlage für die Testentwicklung von Lernverlaufsdiagnostik. *Zeitschrift Für Heilpädagogik*, *74*, 29–37. https://doi.org/10.5283/epub.53484

Ebenbeck, N., Rieser, J., Jungjohann, J., & Gebhardt, M. (2022). How The Existence of Special Schools Affects the Placement of Students with Special Needs in Inclusive Primary Schools. *Journal of Research in Special Educational Needs*, *22*(3), 274–287. https://doi.org/10.1111/1471-3802.12565

Ebrahimi, M. R., Hashemi T., S. M., & Shahbazi, V. (2019). Score Equivalence, Gender Difference, and Testing Mode Preference in a Comparative Study between Computer-Based Testing and Paper-Based Testing. *International Journal of Emerging Technologies in Learning*, *14*(7), 128–143. https://doi.org/10.3991/ijet.v14i07.10175

Eggen, T. J. H. M. (2000). On the Loss of Information in Conditional Maximum Likelihood Estimation of Item Parameters. *Psychometrika*, *65*(3), 337–362. https://doi.org/10.1007/BF02296150

Eigner, B. (2022). Diagnostik im Kontext geistiger Behinderung: Komplexität, Herausforderungen, Strategien. In M. Gebhardt, D. Scheer, & M. Schurig (Eds.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung. Version 1.0* (pp. 421–434). Universität Regensburg.

Ennemoser, M., Kuhl, J., & Pepouna, S. (2013). Evaluation des Dialogischen Lesens zur Sprachförderung bei Kindern mit Migrationshintergrund. *Zeitschrift für Pädagogische Psychologie*, *27*(4), 229–239. https://doi.org/10.1024/1010-0652/a000109

Farquharson, K., Centanni, T. M., Franzluebbers, C. E., & Hogan, T. P. (2014). Phonological and Lexical Influences on Phonological Awareness in Children with Specific Language Impairment and Dyslexia. *Frontiers in Psychology*, *5*, 838. https://doi.org/10.3389/fpsyg.2014.00838

Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., & Clemens, N. H. (2018). Data-Based Decision Making in Reading Interventions: A Synthesis and Meta-Analysis of the Effects for Struggling Readers. *The Journal of Special Education*, *52*(3), 174–187. https://doi.org/10.1177/0022466918790001

Fisher, R. A. (1921). On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron.

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & Beurs, E. (2016). Simulating Computer Adaptive Testing with the Mood and Anxiety Symptom Questionnaire. *Psychological Assessment*, *28*(8), 953–962. https://doi.org/10.1037/pas0000240

Fletcher, J. M., Coulter, W. A., Reschly, D. J., & Vaughn, S. (2004). Alternative Approaches to the Definition and Identification of Learning Disabilities: Some Questions and Answers. *Annals of Dyslexia*, *54*(2), 304–331. https://doi.org/10.1007/s11881-004-0015-y

Forkmann, T., Boecker, M., Norra, C., Eberle, N., Kircher, T., Schauerte, P., Mischke, K., Westhofen, M., Gauggel, S., & Wirtz, M. (2009). Development of an Item Bank for the Assessment of Depression in Persons with Mental Illnesses and Physical Diseases Using Rasch Analysis. *Rehabilitation Psychology*, *54*(2), 186–197. https://doi.org/10.1037/a0015612

Francis, D. J., Fletcher, J. M., Shaywitz, B. A., Shaywitz, S. E., & Rourke, B. P. (1996). Defining Learning and Language Disabilities. *Language, Speech, and Hearing Services in Schools*, *27*(2), 132–143. https://doi.org/10.1044/0161-1461.2702.132

Franz, D. J., Richter, T., Lenhard, W., Marx, P., Stein, R., & Ratz, C. (2023). The Influence of Diagnostic Labels on the Evaluation of Students: A Multilevel Meta-Analysis. *Educational Psychology Review*, *35*(1). https://doi.org/10.1007/s10648-023-09716-6

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, Why and How Valid Is It? *Reading Reasearch Quarterly*, *41*(1). https://www.jstor.org/stable/pdf/4151803.pdf

Fuchs, D., & Fuchs, L. S. (2011). Responsiveness to Intervention: Multilevel Assessment and Instruction as Early Intervention and Disability Identification. *The Reading Teacher*, *63*(3), 250–252. https://doi.org/10.1598/RT.63.3.10

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-Intervention: Definitions, Evidence, and Implications for the Learning Disabilities Construct. *Learning Disabilities Research and Practice*, *18*(3), 157–171. https://doi.org/10.1111/1540-5826.00072

Fuchs, L. S., Fuchs, D., & Bishop, N. (1992). Instructional Adaptation for Students at Risk. *The Journal of Educational Research*, *86*(2), 70–84.

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics Screening and Progress Monitoring at First Grade: Implications for Responsiveness to Intervention. *Exceptional Children*, *73*(3), 311–330. https://doi.org/10.1177/001440290707300303

Fuchs, L. S., Fuchs, D., Compton, D. L., Wehby, J., Schumacher, R. F., Gersten, R., & Jordan, N. C. (2015). Inclusion Versus Specialized Intervention for Very-Low-

Performing Students. *Exceptional Children*, *81*(2), 134–157. https://doi.org/10.1177/0014402914551743

Gardner, H. (2005). *The Development and Education of Mind*. *World Library of Educationalists*. Taylor & Francis.

Gebhardt, M., Krammer, M., Schwab, S., Rossmann, P. & Gasteiger-Klicpera, B. (2013). What is Behind the Diagnosis of Learning Disability in Austrian Schools? An Empirical Evaluation of the Results of the Diagnostic Process. International Journal of Special Education, 28(2), 147–153. http://www.internationalsped.com/

Gebhardt, M. (2023). *Inklusiv- und sonderpädagogische Pädagogik im Schwerpunkt Lernen. Eine Einführung*. Universität Regensburg. https://doi.org/10.5283/epub.45609

Gelderblom, G., Schildkamp, K., Pieters, J., & Ehren, M. (2016). Data-Based Decision Making for Instructional Improvement in Primary Education. https://discovery.ucl.ac.uk/id/eprint/1535422/1/Ehren_Data_based_decision_making_for_instructional_improvement_20052016.pdf

Gerald, B. (2018). A Brief Review of Independent, Dependent and One Sample t-test. *International Journal of Applied Mathematics and Theoretical Physics*, *4*(2), 50. https://doi.org/10.11648/j.ijamtp.20180402.13

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a Computerized Adaptive Test for Depression. *Archives of General Psychiatry*, *69*(11), 1104–1112. https://doi.org/10.1001/archgenpsychiatry.2012.14

Gilmour, A. F., Fuchs, D., & Wehby, J. H. (2019). Are Students With Disabilities Accessing the Curriculum? A Meta-Analysis of the Reading Achievement Gap Between Students With and Without Disabilities. *Exceptional Children*, *85*(3), 329–346. https://doi.org/10.1177/0014402918795830

Goldan, J., & Schwab, S. (2020). Measuring Students' and Teachers' Perceptions of Resources in Inclusive Education – Validation of a Newly Developed Instrument. *International Journal of Inclusive Education*, *24*(12), 1326–1339. https://doi.org/10.1080/13603116.2018.1515270

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review. *Applied Sciences*, *11*(12), 5467. https://doi.org/10.3390/app11125467

Green, S., Davis, C., Karshmer, E., Marsh, P., & Straight, B. (2005). Living Stigma: The Impact of Labeling, Stereotyping, Separation, Status Loss, and Discrimination in the Lives of Individuals with Disabilities and Their Families. *Sociological Inquiry*, *75*(2), 197–215. https://doi.org/10.1111/j.1475-682X.2005.00119.x

Greenspan, S., & Woods, G. W. (2014). Intellectual Disability as a Disorder of Reasoning and Judgement: The Gradual Move Away From Intelligence Quotient-Ceilings. *Current Opinion in Psychiatry*, *27*(2), 110–116. https://doi.org/10.1097/YCO.0000000000000037

Gresham, F. M., MacMillan, D. L., & Bocian, K. M. (1996). Learning Disabilities, Low achievement, and Mild Mental Retardation: More Alike Than Different? *Journal of Learning Disabilities*, *29*(6), 570–581. https://doi.org/10.1177/002221949602900601

Grünke, M. & Cavendish, W. M. (2016). Learning disabilities around the globe: Making sense of the heterogeneity of the different viewpoints. Learning Disabilities: A Contemporary Journal, 14(1), 1–8.

Grynova, M., & Kalinichenko, I. (2018). Trends in Inclusive Education in the USA and Canada. *Comparative Professional Pedagogy*, *8*(2), 28–34.

Gulliford, R., & Upton, G. (2002). *Special Educational Needs*. Routledge.

Gulliksen, H. (1950). The Reliability of Speeded Tests. *Psychometrika*, *15*(3), 259–269. https://doi.org/10.1007/BF02289042

Gunderson, L., & Siegel, L. S. (2001). The Evils of the Use of IQ Tests to Define Learning Disabilities in First- and Second-Language Learners. *The Reading Teacher*, *55*(1), 48–55. https://www.jstor.org/stable/20205010

Hakkarainen, A., Holopainen, L., & Savolainen, H. (2013). Mathematical and Reading Difficulties as Predictors of School Achievement and Transition to Secondary Education. *Scandinavian Journal of Educational Research*, *57*(5), 488–506. https://doi.org/10.1080/00313831.2012.696207

Han, K. T. (2012). SimulCAT: Windows Software for Simulating Computerized Adaptive Test Administration. *Applied Psychological Measurement*, *36*(1), 64–66. https://www.umass.edu/remp/software/simcata/simulcat/

Harlen, W., & James, M. (1997). Assessment and Learning: Differences and Relationships Between Formative and Summative Assessment. . *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365–379.

Hasbrouck, J., & Tindal, G. A. (2006). Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *The Reading Teacher*, *59*(7), 636–644. https://doi.org/10.1598/RT.59.7.3

Hayes, A. M., & Bulat, J. (2017). *Disabilities Inclusive Education Systems and Policies Guide for Low- and Middle-Income Countries*. RTI Press. https://files.eric.ed.gov/fulltext/ED581498.pdf

He, W., & Reckase, M. D. (2014). Item Pool Design for an Operational Variable-Length Computerized Adaptive Test. *Educational and Psychological Measurement*, *74*(3), 473–494. https://doi.org/10.1177/0013164413509629

Heine, J. H. (2016). *The Pairwise-Method for Parameter Estimation Within the Ordinal Rasch-Model*. http://www.heinewiki.de/jhh/pairwise_files/presentation_pairwise_method.pdf

Heine, J. H. (2022). *pairwise: Rasch Model Parameters by Pairwise Algorithm. R package version 0.6.0-0.* https://cran.r-project.org/package=pairwise

Heine, J. H., & Tarnai, C. (2015). Pairwise Rasch Model Item Parameter Recovery Under Sparse Data Conditions. *Psychological Test and Assessment Modeling*, *57*(1), 3–36. http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2015_20150327/01_heine.pdf

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a Concept-Grounded, Curriculum-Based Measure in Mathematics to Predict Statewide Test Scores for Middle School Students with LD. *The Journal of Special Education*, *36*(2), 102–112. https://doi.org/10.1177/00224669020360020501

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., ... & Stachl, C. (2021). Machine learning for the educational sciences. Review of Education, 9(3), e3310.

Hoogland, I., Schildkamp, K., van der Kleij, F., Heitink, M., Kippers, W., Veldkamp, B., & Dijkstra, A. M. (2016). Prerequisites for Data-Based Decision Making in the Classroom: Research Evidence and Practical Illustrations. *Teaching and Teacher Education*, *60*, 377–386. https://doi.org/10.1016/j.tate.2016.07.012

Hsu, Y.-J. (2019). *A Comparative Study of Optimal Pool Design Methods in Computerized Adaptive Testing: Doctoral Dissertation.* https://doi.org/10.17077/etd.qudr3885

Isoaho, P., Kauppila, T., & Launonen, K. (2016). Specific Language Impairment (SLI) and Reading Development in Early School Cears. *Child Language Teaching and Therapy*, *32*(2), 147–157. https://doi.org/10.1177/0265659015601165

Jacobson, C. (1999). How Persistent is Reading Disability? Individual Growth Curves in Reading. *Dyslexia*, *5*(2), 78–93. https://doi.org/10.1002/(SICI)1099-0909(199906)5:2%3C78::AID-DYS127%3E3.0.CO;2-8

Jeong, H. (2014). A Comparative Study of Scores on Computer-Based Tests and Paper-Based Tests. *Behaviour & Information Technology*, *33*(4), 410–422. https://doi.org/10.1080/0144929X.2012.710647

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2015). From Response to Intervention to Multi-Tiered Systems of Support: Advances in the Science and Practice of Assessment and Intervention. In S. R. Jimerson, M. K. Burns, & A. VanDerHeyden (Eds.), *Handbook of Response to Intervention: The Science and Practice of Multi-Tiered Systems of Support* (2nd ed., pp. 1–6). Springer-Verlag. https://doi.org/10.1007/978-1-4899-7568-3_1

Jones, R. L. (1972). Labels and Stigma in Special Education. *Exceptional Children*, *38*(7), 553–564. https://doi.org/10.1177/001440297203800705

Jungjohann, J. (2022). Komplexe Nebensätze, Kohärenz- oder Inferenzbildung: Unterschiede im satzübergreifenden Leseverständnis von Jugendlichen mit sonderpädagogischem Unterstützungsbedarf im Bereich Sprache. *Forschung Sprache*, *10*(2), 19–33. https://doi.org/10.5283/epub.53198

Jungjohann, J., DeVries, J. M., Gebhardt, M., & Mühling, A. (2018). Levumi: A Web-Based Curriculum-Based Measurement to Monitor Learning Progress in Inclusive Classrooms. In K. Miesenberger & G. Kouroupetroglou (Eds.), *Computers Helping People with Special Needs* (Vol. 10896, pp. 369–378). Springer International Publishing. https://doi.org/10.1007/978-3-319-94277-3_58

Jungjohann, J., Ebenbeck, N., Diehl, K., Liebers, K., Gebhardt, M. (submitted). Das Lesescreening LES-IN für inklusive Grundschulklassen: Entwicklung und psychometrische Prüfung einer Paper-Pencil-Version als Basis für computerbasiertes adaptives Testen (CAT).

Kaznowski, K. (2004). Slow Learners: Are Educators Leaving Them Behind? *NASSP Bulletin*, *88*(641), 31–45. https://doi.org/10.1177/019263650408864103

Kelderman, H. (1984). Loglinear Rasch Model Tests. *Psychometrika*, *49*(2), 223–245. https://ris.utwente.nl/ws/files/6876750/Kelderman_10.1007_BF02294174.pdf

Kelso, K., Fletcher, J., & Lee, P. (2007). Reading Comprehension in Children with Specific Language Impairment: An Examination of Two Subgroups. *International Journal of Language & Communication Disorders*, *42*(1), 39–57. https://doi.org/10.1080/13682820600693013

Kintz, B. L., Delprato, D. J., Mettee, D. R., Persons, C. E., & Schappe, R. H. (1965). The Experimenter Effect. *Psychological Bulletin*, *63*, 223–232. https://doi.org/10.1037/h0021718

Kline, P. (2015). *A Handbook of Test Construction: Introduction to Psychometric Design*. *Psychology Revivals Ser*. Routledge.

KMK (2019). Empfehlungen zur schulischen Bildung, Beratung und Unterstützung von Kindern und Jugendlichen im sonderpädagogischen Schwerpunkt LERNEN (Beschluss

der Kultusministerkonferenz vom 14.03.2019). https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2019/2019_03_14-FS-Lernen.pdf
Korkmaz, C., & Correia, A. P. (2019). A review of research on machine learning in educational technology. Educational Media International, 56(3), 250-267.
Kovaleski, J., VanDerHeyden, A., & Shapiro, E. S. (2013). The RTI Approach to Evaluating Learning Disabilities. *The Guilford Press*. http://tarn.kansasmtss.org/pdf/BookStudies/RTI-Approach-to-Evaluating-LD.pdf
Krammer, G. (2018). The Andersen Likelihood Ratio Test with a Random Split Criterion Lacks Power. *Journal of Modern Applied Statistical Methods, 17*(2), Article jmasm.eP2685. https://doi.org/10.22237/jmasm/1555594442
Kubinger, K. D. (2003). Adaptives Testen. In K. D. Kubinger & R. S. Jäger (Eds.), *Schlüsselbegriffe der psychologischen Diagnostik: Handbuch* (1st ed., pp. 1–9). Beltz.
Kubinger, K. D. (2007). Towards Economic Wechsler-Like Testing: Adaptive Intelligence Diagnosticum (AID 2). In M. A. Lange & A. M. Annus (Eds.), *Leading-Edge Psychological Tests and Testing Research* (pp. 173–182). Nova Science Publ.
Kubinger, K. D. (2017). Neue Konzepte und Belege zu den Einsatzmöglichkeiten des AID in der Entwicklungs- und Pädagogischen Psychologie. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *49*(3), 115–126. https://doi.org/10.1026/0049-8637/a000174
Lai, M. K., & Schildkamp, K. (2013). Data-Based Decision Making: An Overview. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Studies in Educational Leadership: Vol. 17. Data-Based Decision Making in Education: Challenges and Opportunities* (pp. 9–21). Springer. https://doi.org/10.1007/978-94-007-4816-3_2
Latu, E., & Chapman, E. (2002). Computerised Adaptive Testing. *British Journal of Educational Technology*, *33*(5), 619–622. https://doi.org/10.1111/1467-8535.00296
Levy, Y. (2011). Iq Predicts Word Decoding Skills in Populations with Intellectual Disabilities. *Research in Developmental Disabilities*, *32*(6), 2267–2277. https://doi.org/10.1016/j.ridd.2011.07.043
Liebers, K., Kanold, E., & Junger, R. (2019). Digitale Lernstandsanalysen in der inklusiven Grundschule? In S. Bartusch, C. Klektau, T. Simon, & Teumer, S. Weidermann, A. (Eds.), *Lernprozesse Begleiten* (pp. 209–221). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21924-6_16
Linacre, J. M. (2000). *Computer-Adaptive Testing: A Methodology Whose Time Has Come.* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d41985bb9c543b94c60fc7dc6ab5e0ca31b8362f
Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.
Linacre J. M., & Wright B. D. (1989). The "Length" of a Logit. *Rasch Measurement Transactions*, *3*(2), 54–55. https://www.rasch.org/rmt/rmt32b.htm
Lindsay, G. (2003). Inclusive Education: A Critical Perspective. *British Journal of Special Education*, *30*(1), 3–12. https://doi.org/10.1111/1467-8527.00275
Lindsay, G., Wedell, K., & Dockrell, J. (2020). Warnock 40 Years on: The Development of Special Educational Needs Since the Warnock Report and Implications for the Future. *Frontiers in Education*, *4*, Article 164. https://doi.org/10.3389/feduc.2019.00164
Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lord, F. M. (1968). Some Test Theory for Tailored Testing. *ETS Research Bulletin Series*, *1968.2*(1968), i-62.

Lord, F. M. (1980). *Applications of Item Response Theory To Practical Testing Problems*. Taylor and Francis.

Lovett, B. J., & Lewandowski, L. J. (2006). Gifted Students with Learning Disabilities: Who are They? *Journal of Learning Disabilities*, *39*(6), 515–527. https://doi.org/10.1177/00222194060390060401

Luckin, R. (2017). Towards Artificial Intelligencebased Assessment Systems. *Nature Human Behavior*, *1*, Article 0028. https://www.researchgate.net/profile/Rosemary-Luckin/publication/314088884_Towards_artificial_intelligence-based_assessment_systems/links/5b523a9f0f7e9b240ff26e23/Towards-artificial-intelligence-based-assessment-systems.pdf

Ludewig, U., Trendtel, M., Schlitter, T., & McElvany, N. (2022). Adaptives Testen von Textverständnis in der Grundschule. *Diagnostica*, *68*(1), 39–50. https://doi.org/10.1026/0012-1924/a000279

Lunz, M. E. (2010). Using The Very Useful Wright Map. *Measurement Resarch Associates Test Insights*. https://www.rasch.org/mra/mra-01-10.htm

Lutz, S., Boschner, S., & Gebhardt, M. (2022). Data‐Based Decision Making (DBDM) in der inklusiven Diagnostik und Förderplanung. In M. Gebhardt, D. Scheer, & M. Schurig (Eds.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung. Version 1.0* (pp. 33–42). Universität Regensburg.

Lyon, G. R. (1989). Iq Is Irrelevant to the Definition of Learning Disabilities: A Position in Search of Logic and Data. *Journal of Learning Disabilities*, *22*(8), 504. https://eric.ed.gov/?id=ej400599

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywith, B. A., Torgesen, J. K., Wood, F. B., & Olsen, R. (2001). *Rethinking Learning Disabilities.*

Madalyn, N. (2021). Improving the Identification and Diagnosis of Learning Disabilities. *Capstone Projects and Master's Theses*(1185). https://digitalcommons.csumb.edu/caps_thes_all/1185

Magis, D., & Barrada, J. R. (2017). Computerized Adaptive Testing with R : Recent Updates of the Package catR. *Journal of Statistical Software*, *76*(Code Snippet 1). https://doi.org/10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, *48*(8), 1–31.

Magis, D., Yan, D., & Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR. Use R!* Springer. https://doi.org/10.1007/978-3-319-69218-0

Maki, K. E., & Adams, S. R. (2019). A Current Landscape of Specific Learning Disability Identification: Training, Practices, and Implications. *Psychology in the Schools*, *56*(1), 18–31. https://doi.org/10.1002/pits.22179

Martin, R. (2008). New Possibilities and Challenges for Assessment Through the Use of Technology. *Towards a Research Agenda on Computer-Based Assessment*, 6–9.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An Examination of the Equivalence between Non-Adaptive Computer-Based and Traditional Testing. *Journal of Educational*

Computing Research, *24*(1), 29–39. https://doi.org/10.2190/9EPM-B14R-XQWT-WVNL

Mayer, A., & Motsch, H. J. (2016). Förderschwerpunkt Sprache. *Sonderpädagogische Förderschwerpunkte in NRW*, *28*.

McClelland, T., & Cuevas, J. A. (2020). A Comparison of Computer-Based Testing and Paper and Pencil Testing in Mathematics Assessment. *The Online Journal of New Horizons in Education*, *10*(2), 78–89. https://www.tojned.net/journals/tojned/articles/v10i02/v10i02-01.pdf

McCusker, K. A., Harkin, J., Wilson, S., & Callaghan, M. (2013, October). Intelligent assessment and content personalisation in adaptive educational systems. In 2013 12th International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 1-7). IEEE.

Meijer, R. R., & Nering, M. L. (1999). Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement*, *23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Menzies, H. M., Mahdavi, J. N., & Lewis, J. L. (2008). Early Intervention in Reading. *Remedial and Special Education*, *29*(2), 67–77. https://doi.org/10.1177/0741932508315844

Michiel H. A., Vorst, H. C. M., & Mellenbergh, G. J. (2008). Computerized Adaptive Testing of Personality Traits. *Journal of Psychology*, *216*(1), 12–21.

Mills, C. N., & Stocking, M. L. (1996). Practical Issues in Large-Scale Computerized Adaptive Testing. *Applied Measurement in Education*, *9*(4), 287–304. https://doi.org/10.1207/s15324818ame0904_1

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and Evaluating a Computerized Adaptive Testing Version of the Word Part Levels Test. *Language Testing*, *36*(1), 101–123. https://doi.org/10.1177/0265532217725776

Mojarrad, H. (2013). Computer-Based Assessment (CBA) vs. Paper/Pencil-Based Assessment (PPBA): An Investigation Into the Performance and Attitude of Iranian EFL Learners' Reading Comprehension. *International Journal of Language Learning and Applied Linguistics World*, *4*(4), 418–428.

Molenaar, I. W. (1995). Estimation of Item Parameters. In G. H. Fischer (Ed.), *Rasch Models: Foundations, Recent Developments, and Applications* (1st ed., pp. 39–51). Springer New York. https://doi.org/10.1007/978-1-4612-4230-7_3

Moosbrugger, H., & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“). In H. Moosbrugger & A. Kelava (Eds.), *Lehrbuch. Testtheorie und Fragebogenkonstruktion* (3rd ed., pp. 13–38). Springer. https://doi.org/10.1007/978-3-662-61532-4_2

Moosbrugger, H., Schermelleh-Engel, K., Gäde, J. C., & Kelava, A. (2020). Testtheorien im Überblick. In H. Moosbrugger & A. Kelava (Eds.), *Lehrbuch. Testtheorie und Fragebogenkonstruktion* (3rd ed., pp. 251–273). Springer. https://doi.org/10.1007/978-3-662-61532-4_12

Myung, I. J. (2003). Tutorial on Maximum Likelihood Estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. https://doi.org/10.1016/S0022-2496(02)00028-7

Nafea, I. T. (2018). Machine learning in educational technology. Machine learning-advanced techniques and emerging applications, 175-183.

Naglieri, J. A., & Reardon, S. M. (1993). Traditional IQ is Irrelevant to Learning Disabilities - Intelligence is Not. *Journal of Learning Disabilities*, *26*(2), 127–133. https://doi.org/10.1177/002221949302600205

National Center for Learning Disabilities. (2017a). *Identifying Struggling Students*. https://www.ncld.org/research/state-of-learning-disabilities/identifying-struggling-students/

National Center for Learning Disabilities. (2017b). *Understanding Learning and Attention Issues*. https://www.ncld.org/news/state-of-learning-disabilities/understanding-learning-and-attention-issues/

Ngwacho, G. A. (2022). Utilization of Digital Technologies to Enhance Assessments, Practices, and Equity in Inclusive Education. In J. Keengwe (Ed.), *Handbook of Research on Digital-Based Assessment and Innovative Practices in Education* (pp. 295–317). IGI Global. https://doi.org/10.4018/978-1-6684-2468-1.ch016

Norwich, B. (2009). Dilemmas of Difference and the Identification of Special Educational Needs/Disability: International Perspectives. *British Educational Research Journal*, *35*(3), 447–467. https://doi.org/10.1080/01411920802044446

Nouwens, P. J. G., Lucas, R., Smulders, N. B. M., Embregts, P. J. C. M., & van Nieuwenhuizen, C. (2017). Identifying Classes of Persons with Mild Intellectual Disability or Borderline Intellectual Functioning: A Latent Class Analysis. *BMC Psychiatry*, *17*(1), 257. https://doi.org/10.1186/s12888-017-1426-8

Nydick, S. (2022). Simulate IRT-Based Computerized Adaptive Tests. https://cran.r-project.org/web/packages/catIrt/catIrt.pdf

Papuga, M. O., Dasilva, C., McIntyre, A., Mitten, D., Kates, S., & Baumhauer, J. F. (2018). Large-scale Clinical Implementation of PROMIS Computer Adaptive Testing with Direct Incorporation into the Electronic Medical Record. *Health Systems*, *7*(1), 1–12. https://doi.org/10.1057/s41306-016-0016-1

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative Item Types for Computerized Testing. In W. J. van der Linden & G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 129–148). Springer Netherlands. https://doi.org/10.1007/0-306-47531-6_7

Partanen, M., & Siegel, L. S. (2014). Long-Term Outcome of the Early Identification and Intervention of Reading Disabilities. *Reading and Writing*, *27*(4), 665–684. https://doi.org/10.1007/s11145-013-9472-1

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25-45.

Petersen, M. A., Groenvold, M., Aaronson, N. K., Chie, W.-C., Conroy, T., Costantini, A., Fayers, P., Helbostad, J., Holzner, B., Kaasa, S., Singer, S., Velikova, G., & Young, T. (2010). Development of Computerised Adaptive Testing (CAT) for the EORTC QLQ-C30 Dimensions - General Approach and Initial Results for Physical Functioning. *European Journal of Cancer*, *46*(8), 1352–1358. https://doi.org/10.1016/j.ejca.2010.02.011

Piaw Chua, Y. (2012). Effects of Computer-Based Testing on Test Performance and Testing Motivation. *Computers in Human Behavior*, *28*(5), 1580–1586. https://doi.org/10.1016/j.chb.2012.03.020

Pomplun, M., Frey, S., & Becker, D. F. (2002). The Score Equivalence of Paper-and-Pencil and Computerized Versions of a Speeded Test of Reading Comprehension. *Educational and Psychological Measurement*, *62*(2), 337–354. https://doi.org/10.1177/0013164402062002009

Preston, A. I., Wood, C. L., & Stecker, P. M. (2016). Response to Intervention: Where It Came From and Where It's Going. *Preventing School Failure: Alternative Education*

for Children and Youth, *60*(3), 173–182. https://doi.org/10.1080/1045988X.2015.1065399

Puranik, C. S., Petscher, Y., Al Otaiba, S., Catts, H. W., & Lonigan, C. J. (2008). Development of Oral Reading Fluency in Children with Speech or Language Impairments: A Growth Curve Analysis. *Journal of Learning Disabilities*, *41*(6), 545–560. https://doi.org/10.1177/0022219408317858

R Core Team. (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rabinowitz, S., & Brandt, T. (2001). *Computer-Based Assessment: Can It Deliver on Its Promise? Knowledge Brief.* WestEd. https://files.eric.ed.gov/fulltext/ED462447.pdf

Rasch, G. (1960). *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. https://psycnet.apa.org/record/1962-07791-000

Reckase, M. D. (2003). *Item Pool Design for Computerized Adaptive Tests.* Annual meeting of the National Council on Measurement in Education. http://iacat.org/sites/default/files/biblio/re03-01.pdf

Redecker, C., & Johannessen, Ø. (2013). Changing Assessment — Towards a New Assessment Paradigm Using ICT. *European Journal of Education*, *48*(1).

Reschly, D. J. (1996). Identification and Assessment of Students with Disabilities. *The Future of Children*, 40–53.

Restori, A. F., Katz, G. S., & Lee, H. B. (2009). A Critique of the IQ / Achievement Discrepancy Model for Identifying Specific Learning Disabilities. *Europe's Journal of Psychology*, *5*(4). https://doi.org/10.5964/ejop.v5i4.244

Richter, T., Naumann, J., Isberner, M.-B., Neeb, Y., & Knoepke, J. (2012). *ProDi-L. Prozessbezogene Diagnostik von Lesefähigkeiten im Grundschulalter.* hogrefe.

Ricketts, J. (2011). Research Review: Reading Comprehension in Developmental Disorders of Language and Communication. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *52*(11), 1111–1123. https://doi.org/10.1111/j.1469-7610.2011.02438.x

Rispens, J., van Yperen, T. A., & van Duijn, G. A. (1991). The Irrelevance of IQ to the Definition of Learning Disabilities: Some Empirical Evidence. *Journal of Learning Disabilities*, *24*(7), 434–438. https://doi.org/10.1177/002221949102400709

Rizzo Meneghetti, D. (2016). catsim - Computerized Adaptive Testing Simulator. Advance online publication. https://doi.org/10.5281/zenodo.46420

Rooney, K., Polloway, E. A., & Hallahan, D. P. (1985). The Use of Self-Monitoring Procedures with Low IQ Learning Disabled Students. *Journal of Learning Disabilities*, *18*(7), 384–389. https://doi.org/10.1177/002221948501800703

Röthlisberger, M., Schneider, H., & Juska-Bacher, B. (2021). Lesen von Kindern mit Deutsch als Erst- und Zweitsprache – Wortschatz als limitierender Faktor. *Zeitschrift für Grundschulforschung*, *14*(2), 259–374. https://phrepo.phbern.ch/id/eprint/67

RStudio Team. (2020). *RStudio: Integrated Development for R.* RStudio, Inc., Boston, MA. http://www.rstudio.com/

Rulison, K. L., & Loken, E. (2009). I've Fallen and I Can't Get Up: Can High Ability Students Recover From Early Mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83–101. https://doi.org/10.1177/0146621608324023

Saddler, B., & Asaro-Saddler, K. (2013). Response to Intervention in Writing: A Suggested Framework for Screening, Intervention, and Progress Monitoring. *Reading & Writing Quarterly*, *29*(1), 20–43. https://doi.org/10.1080/10573569.2013.741945

Sands, W. A., & Gade, P. A. (1983). An Application of Computerized Adaptive Testing in U.S. Army Recruiting. *Journal of Computer-Based Instruction*, *10*(3-4), 87–89.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized Adaptive Testing: From Inquiry to Operation*. American Psychological Association. https://doi.org/10.1037/10244-000

Schalock, R. L., Luckasson, R., & Tassé, M. J. (2021). An Overview of Intellectual Disability: Definition, Diagnosis, Classification, and Systems of Supports (12th ed.). *American Journal on Intellectual and Developmental Disabilities*, *126*(6), 439–442. https://doi.org/10.1352/1944-7558-126.6.439

Schaper, N. (2009). Online-Tests aus diagnostisch-methodischer Sicht. In H. Steiner (Ed.), *Online-Assessment* (pp. 17–36). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78919-2_2

Scheiblechner, H. H. (2009). Rasch and Pseudo-Rasch Models: Suitableness for Practical Test Applications. *Psychology Science Quarterly*, *51*(2), 181–194. https://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/2-2009/05_Scheiblechner.pdf

Schildkamp, K. (2019). Data-Based Decision-Making for School Improvement: Research Insights and Gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Schildkamp, K., Poortman, C. L., & Sahlberg, P. (2019). Data-Based Decision Making in Developing Countries: Balancing Accountability Measures and Improvement Efforts. *Journal of Professional Capital and Community*, *4*(3), 166–171. https://doi.org/10.1108/JPCC-07-2019-037

Schurig, M., & Gebhardt, M. (2022). Theoretische Grundlagen von Messungen und Tests. In M. Gebhardt, D. Scheer, & M. Schurig (Eds.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung. Version 1.0* (pp. 233–246). Universität Regensburg.

Schwartz, R. M. (2005). Literacy Learning of At-Risk First-Grade Students in the Reading Recovery Early Intervention. *Journal of Educational Psychology*, *97*(2), 257–267. https://doi.org/10.1037/0022-0663.97.2.257

Segall, D. O. (2005). *Computerized Adaptive Testing.* Encyclopedia of Social Measurement. http://www.iacat.org/sites/default/files/biblio/se04-01.pdf

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2022). *Sonderpädagogische Förderung in allgemeinen Schulen (ohne Förderschulen) 2021/2022*. https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Dokumentationen/Aus_SoPae_Int_2021.pdf

Şenel, S., & Kutlu, Ö. (2017). Comparison of two test methods for VIS: Paper-pencil test and CAT. *European Journal of Special Needs Education*, 1–15. https://doi.org/10.1080/08856257.2017.1391014

Shifrer, D. (2013). Stigma of a Label: Educational Expectations for High School Students Labeled with Learning Disabilities. *Journal of Health and Social Behavior*, *54*(4), 462–480. https://doi.org/10.1177/0022146513503346

Shute, V. J., & Zapata-Rivera, D. (2010). Educational measurement and intelligent systems. of the International Encyclopedia of Education. Oxford, UK: Elsevier Publishers.

Siegel, L. S. (1989). Iq is Irrelevant to the Definition of Learning Disabilities. *Journal of Learning Disabilities*, *22*(8), 469-78, 486. https://doi.org/10.1177/002221948902200803

Silver, L. B. (1981). The Relationship Between Learning Disabilities, Hyperactivity, Distractibility, and Behavioral Problems: A Clinical Analysis. *Journal of the American Academy of Child Psychiatry*, *20*(2), 385–397. https://doi.org/10.1016/S0002-7138(09)60996-1

Simkin, Z., & Conti-Ramsden, G. (2006). Evidence of Reading Difficulty in Subgroups of Children with Specific Language Impairment. *Child Language Teaching and Therapy*, *22*(3), 315–331. https://doi.org/10.1191/0265659006ct310xx

Slavin, R. E., Karweit, N. L., & Madden, N. A. (1989). *Effective Programs for Students at Risk.* Allyn and Bacon.

Snell, M. E., Luckasson, R., Borthwick-Duffy, W. S., Bradley, V., Buntinx, W. H. E., Coulter, D. L., Craig, E. P. M., Gomez, S. C., Lachapelle, Y., Reeve, A., Schalock, R. L., Shogren, K. A., Spreat, S., Tassé, M. J., Thompson, J. R., Verdugo, M. A., Wehmeyer, M. L., & Yeager, M. H. (2009). Characteristics and needs of people with intellectual disability who have higher IQs. *Intellectual and Developmental Disabilities*, *47*(3), 220–233. https://doi.org/10.1352/1934-9556-47.3.220

Sorrel, M. A., Nájera, P., & Abad, F. J. (2021). cdcatR: An R Package for Cognitive Diagnostic Computerized Adaptive Testing. *Psych*, *3*(3), 386–403. https://doi.org/10.3390/psych3030028

Spulber, D. (2015). *Inclusive Education in Different East and West European Countries*. https://ec.europa.eu/programmes/erasmus-plus/project-result-content/34cd5e51-7598-49e1-a1bd-4e50fdeba30b/Inclusive%20Education%20in%20Different%20East%20and%20West%20European%20Countries.pdf

Stafford, R. E., Runyon, C. R., Casabianca, J. M., & Dodd, B. G. (2019). Comparing Computer Adaptive Testing Stopping Rules Under the Generalized Partial-Credit Model. *Behavior Research Methods*, *51*(3), 1305–1320. https://doi.org/10.3758/s13428-018-1068-x

Stahle, L., & Wold, S. (1989). Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, *6*(4), 259–272. https://doi.org/10.1016/0169-7439(89)80095-4

Stecker, P. M., Fuchs, D., & Fuchs, L. S. (2008). Progress Monitoring as Essential Practice within Response to Intervention. *Rural Special Education Quarterly*, *27*(4), 10–17. https://doi.org/10.1177/875687050802700403

Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The Predictive Value of IQ. *Merill-Palmer Quarterly*, *47*(1), 1–41. https://www.jstor.org/stable/23093686

Stocking, M. L. (1994). Three Practical Issues for Modern Adaptive Testing Item Pools. *ETS Research Report Series*, *1994*(1), i-34. https://doi.org/10.1002/j.2333-8504.1994.tb01578.x

Stoiber, K. C., & Gettinger, M. (2016). Multi-Tiered Systems of Support and Evidence-Based Practices. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of Response to Intervention* (pp. 121–141). Springer US. https://doi.org/10.1007/978-1-4899-7568-3_9

Stone, E., & Davey, T. (2011). Computer-Adaptive Testing for Students with Disabilities: A Review of the Literature. *ETS Research Report Series*, *2011*(2), i-24. https://doi.org/10.1002/j.2333-8504.2011.tb02268.x

Student. (1908). The probable error of a mean. Biometrika, 6(1), 1-25.

Sugai, G., & Simonsen, B. (2012). *Positive Behavioral Interventions and Supports: History, Defining Features, and Misconceptions.* https://www.hbgsd.us/cms/lib/PA50000648/Centricity/Domain/288/PBIS_.pdf

Taherbhai, H., Seo, D., & Bowman, T. (2012). Comparison of Paper–Pencil and Online Performances of Students with Learning Disabilities. *British Educational Research Journal*, *38*(1), 61–74. https://doi.org/10.1080/01411926.2010.526193

Tarver, S. G., & Hallahan, D. P. (1974). Attention Deficits In Children With Learning Disabilities. *Journal of Learning Disabilities*, *7*(9), 560–569. https://doi.org/10.1177/002221947400700906

Tassé, M. J., Schalock, R. L., Balboni, G., Bersani, H., Borthwick-Duffy, S. A., Spreat, S., Thissen, D., Widaman, K. F., & Zhang, D. (2012). The Construct of Adaptive Behavior: Its Conceptualization, Measurement, and Use in the Field of Intellectual Disability. *American Journal on Intellectual and Developmental Disabilities*, *117*(4), 291–303. https://doi.org/10.1352/1944-7558-117.4.291

Thelwall, M. (2000). Computer-Based Assessment: A Versatile Educational Tool. *Computers & Education*, *34*, 37–49. http://www.upv.es/gie/repositorioIEMA/Thelwall%202000%20Computer-based%20assessment%20a%20versatile%20educational%20tool.pdf

Thompson, N. A., & Weiss, D. A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research and Evaluation*, *16*(Article 1). https://doi.org/10.7275/wqzt-9427

Torgesen, J. K. (1989). Why IQ is Relevant to the Definition of Learning Disabilities. *Journal of Learning Disabilities*, *22*(8), 484–486. https://doi.org/10.1177/002221948902200806

Toro, P. A., Weissberg, R. P., Guare, J., & Liebenstein, N. L. (1990). A Comparison of Children With and Without Learning Disabilities on Social Problem-Solving Skill, School Behavior, and Family Background. *Journal of Learning Disabilities*, *23*(2), 115–120. https://doi.org/10.1177/002221949002300207

Trautwein, J., & Schroeder, S. (2019). WOR-TE: Ein Ja / Nein-Wortschatztest für Kinder verschiedener Altersgruppen. *Diagnostica*, *65*(1), 37–48. https://doi.org/10.1026/0012-1924/a000212

United Nations. (2008). *Convention on the Rights of Persons with Disabilities*. https://www.un.org/disabilities/documents/convention/convoptprot-e.pdf

van der Linden, W. J. (2016). *Handbook of Item Response Theory, Volume One*. *Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Ser*. CRC Press.

van der Linden, W. J., & Pashley, P. J. (2000). Item Selection and Ability Estimation in Adaptive Testing. In C. A. W. Glas & W. J. van der Linden (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 1–25). Kluwer Academic. https://doi.org/10.1007/0-306-47531-6_1

van der Linden, W. J., & Pashley, P. J. (2010). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 3–30). Springer. https://doi.org/10.1007/978-0-387-85461-8_1

van Geel, M., Keuning, T., Visscher, A. J., & Fox, J.-P. (2016). Assessing the Effects of a School-Wide Data-Based Decision-Making Intervention on Student Achievement Growth in Primary Schools. *American Educational Research Journal*, *53*(2), 360–394. https://doi.org/10.3102/0002831216637346

van Ooijen, P. M. A., Darzidehkalani, E., & Dekker, A. (2022). *AI Technical Considerations: Data Storage, Cloud usage and AI Pipeline*. https://arxiv.org/pdf/2201.08356.pdf

Vaughn, S., & Schumm, J. S. (1995). Responsible Inclusion for Students with Learning Disabilities. *Journal of Learning Disabilities*, *28*(5), 264-70, 290. https://doi.org/10.1177/002221949502800502

Vaughn, S., & Wanzek, J. (2014). Intensive Interventions in Reading for Students with Reading Disabilities: Meaningful Impacts. *Learning Disabilities Research and Practice*, *29*(2), 46–53. https://doi.org/10.1111/ldrp.12031

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing Item Pools for Computerized Adaptive Testing. In C. A. W. Glas & W. J. van der Linden (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 149–162). Kluwer Academic. https://doi.org/10.1007/0-306-47531-6_8

Veldkamp, B. P., & van der Linden, W. J. (2010). Designing Item Pools for Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 231–245). Springer. https://doi.org/10.1007/978-0-387-85461-8_12

Verhelst, N. (2001). Testing the Unidimensionality Assumption of the Rasch model. *Methods of Psychological Research Online*, *6*(3).

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual Differences and Test Administration Procedures: A Comparison of Fixed-Item, Computerized-Adaptive, and Self-Adapted Testing. *Applied Measurement in Education*, *7*(1), 53–79. https://doi.org/10.1207/s15324818ame0701_5

Vormittag, I. (2011). *Investigation of Examiner Effects on Test Takers in Standardized Achievement Tests with Special Regard to Gender*. Freie Universität Berlin. https://doi.org/10.17169/refubium-5035

Voß, S., & Blumenthal, Y [Yvonne] (2019). Impacts of the Response-to-Intervention Approach on German Elementary Students. *Internation Journal of Technology and Education*, *8*(1), 1347–1355.

Wagner, R. K., & Torgesen, J. K. (1987). The Nature of Phonological Processing and its Causal Role in the Acquisition of Reading Skills. *Psychological Bulletin*, *101*(2), 192–212. https://doi.org/10.1037/0033-2909.101.2.192

Walter, T., & Schuhfried, G. (2004). Computergestützte psychologische Diagnostik. In G. Mehta (Ed.), *Springer eBook Collection Humanities, Social Science. Die Praxis der Psychologie: Ein Karriereplaner* (pp. 265–272). Springer Vienna. https://doi.org/10.1007/978-3-7091-0571-9_21

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K–12 Reading Assessments. *Educational and Psychological Measurement*, *68*(1), 5–24. https://doi.org/10.1177/0013164407305592

Wang, T., & Kolen, M. J. (2001). Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement*, *38*(1), 19–49. https://doi.org/10.1111/j.1745-3984.2001.tb01115.x

Wang, T., & Vispoel, W. P. (1998). Properties of Ability Estimation Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *35*(2), 109–135. https://doi.org/10.1111/j.1745-3984.1998.tb00530.x

Way, W. D. (2005). *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments: Research Report 05-03*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=96c8892e542ac8dd46c532aa6400271e25b1c9d3

Weinberg, R. A. (1989). Intelligence and IQ: Landmark Issues and Great dDbates. *American Psychologist*, *44*(2), 98–104. https://doi.org/10.1037/0003-066x.44.2.98

Weiss, D. J., & Betz, N. E. (1973). *Ability Measurement: Conventional or Adaptive?* Minnesota University, Minneapolis Departement of Psychology.

Weiss, D. J [David J.], & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Werfel, K. L., & Krimm, H. (2017). A Preliminary Comparison of Reading Subtypes in a Clinical Sample of Children With Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, *60*(9), 2680–2686. https://doi.org/10.1044/2017_JSLHR-L-17-0059

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilcox, G., Fernandez C., C., & K. A. (2021). Using Evidence-Based Practice and Data-Based Decision Making in Inclusive Education. *Education Sciences*, *11*(129), 1–11. https://doi.org/10.3390/educsci11030129

Wyse, A. E., & Albano, A. D. (2015). Considering the Use of General and Modified Assessment Items in Computerized Adaptive Testing. *Applied Measurement in Education*, *28*(2), 156–167. https://doi.org/10.1080/08957347.2014.1002921

Zentel, P., Sansour, T., Engelhardt, M., Krämer, T., & Marzini, M. (2019). Mensch und/oder Maschine?. *Schweizerische Zeitschrift für Heilpädagogik*, *25*(11-12), 35-42.

Zheng, Y., Cheon, H., & Katz, C. M. (2020). Using machine learning methods to develop a short tree-based adaptive classification test: Case study with a high-dimensional item pool and imbalanced data. Applied psychological measurement, 44(7-8), 499-514.

Zucker, S. H., & Polloway, E. A. (1987). Issues in Identification and Assessment in Mental Retardation. *Education and Training in Mental Retardation*, *22*(2), 69–76. https://www.jstor.org/stable/23878332

Zwinderman A. H. (1995). Pairwise Parameter Estimation in Rasch Models. *Applied Psychological Measurement*, *19*(4), 369–375.

Zydney, J., Hord, C., & Koenig, K. (2020). Helping Students with Learning Disabilities Through Video-Based, Universally Designed Assessment. *ELearn*, Article 3403395.3397820. Advance online publication. https://doi.org/10.1145/3403395.3397820

Appendices

A. Example Items of the Subtests

Figure A1

Example item of subtest 1: The title says "Where can you hear the letter?". The graphic in the left box represents the German word for 'book' ("Buch"). In the right box, the letter, that has to be identified within the word, is represented ("B"). Buttons are labelled as beginning ("Anfang"), middle ("Mitte"), end ("Ende"), and nowhere ("Nirgends").

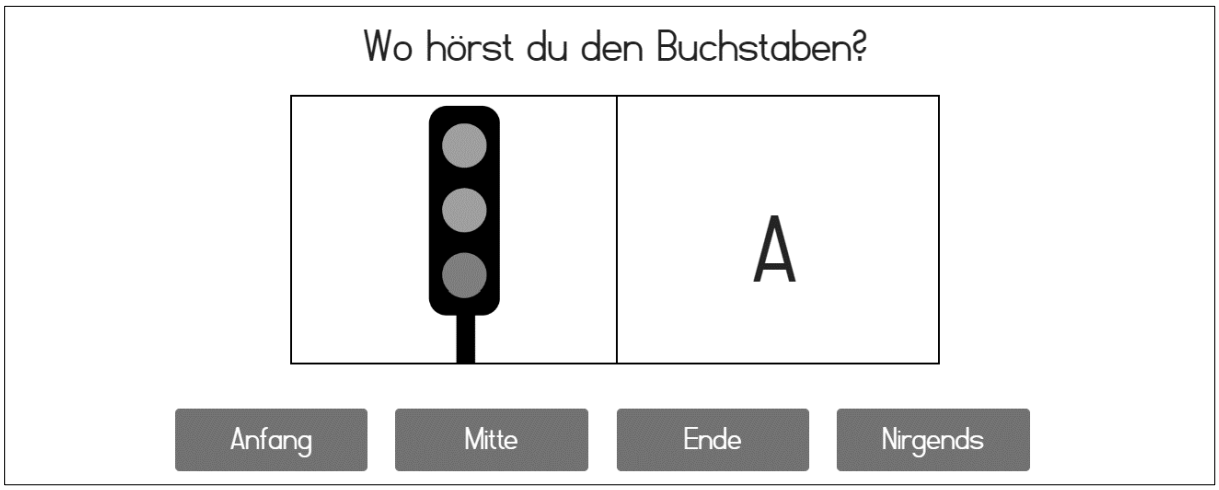


Figure A2

Example item of subtest 2: The title says "Does the word exist?". The word in this example is the real existing German word for 'different' ("anders"). Buttons are labelled as 'Does exist' ("Gibt es") and 'Does not exist' ("Gibt es nicht").

Figure A3

Example item of subtest 3: The title says "Which word did you see?". The flashed word was "April". Buttons are labelled with different words, under which the student choses the correct word.

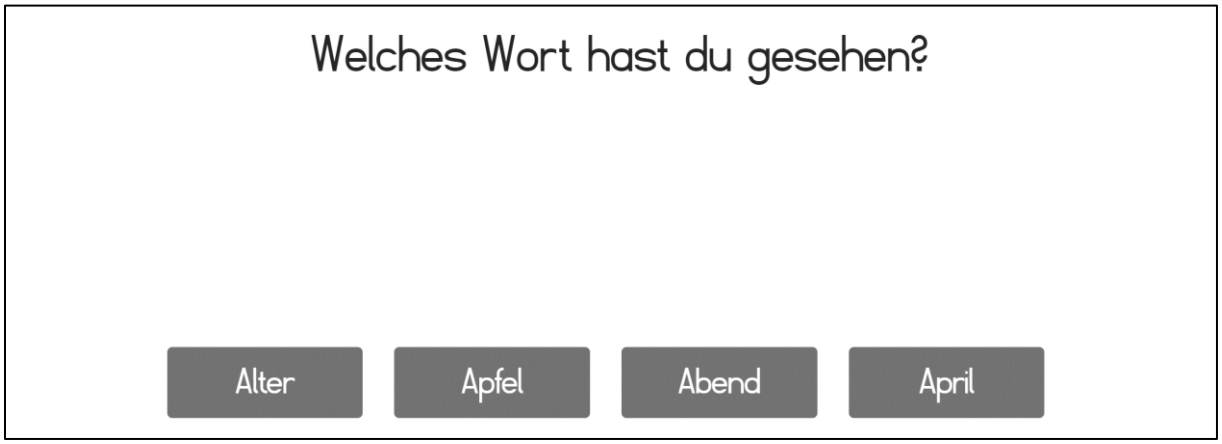


Figure A4

Example item of subtest 4: The given sentence says "A face has two… .". The buttons are labelled with the different choices "Finger" (AE finger), "Bücher" (AE books), "Augen" (AE eyes), and "Autos" (AE cars).

B. Simulation Process Charts

Figure B1

First part of the simulation process in study 1.

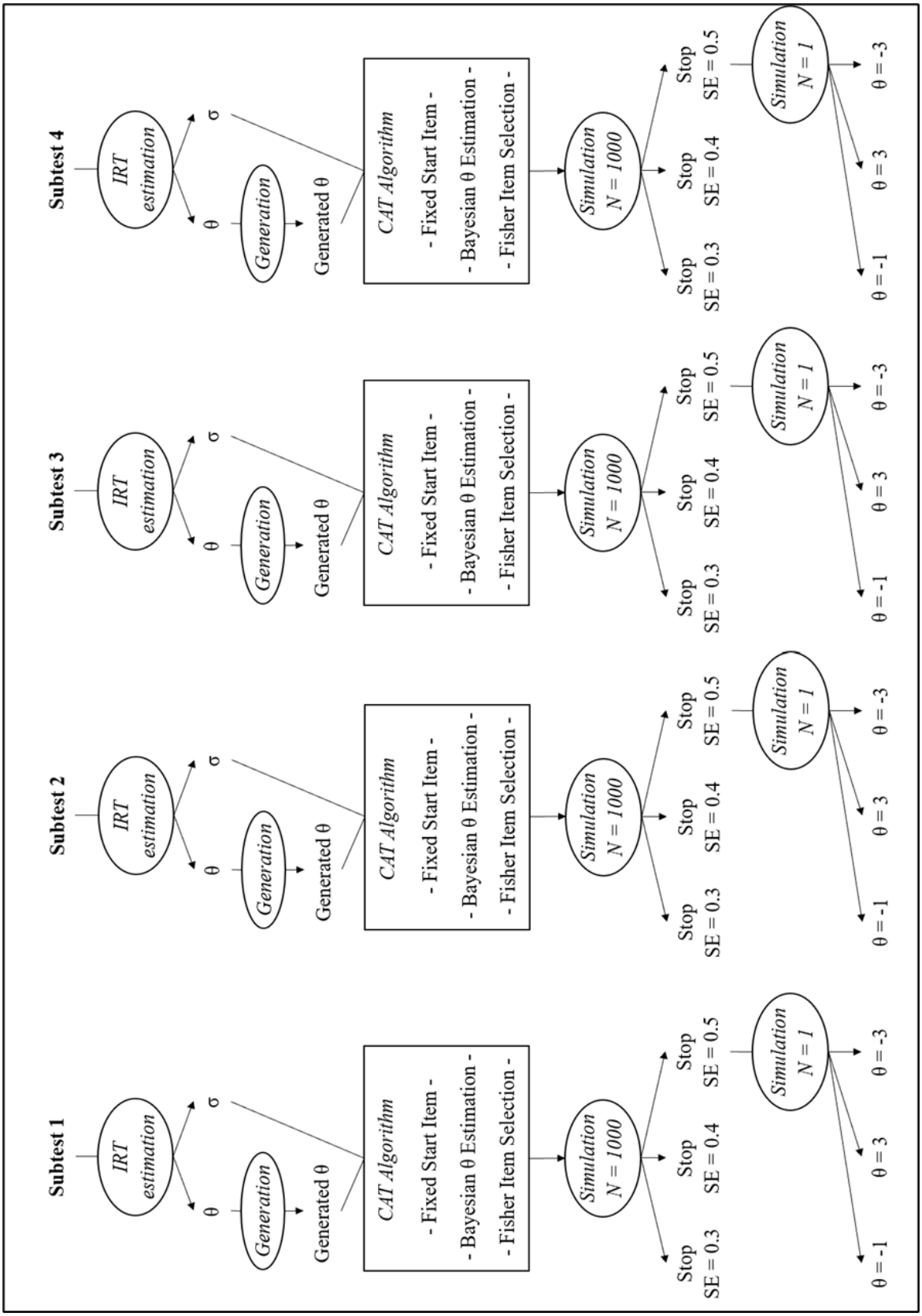


Figure B2

Second part of the simulation process in study 1.

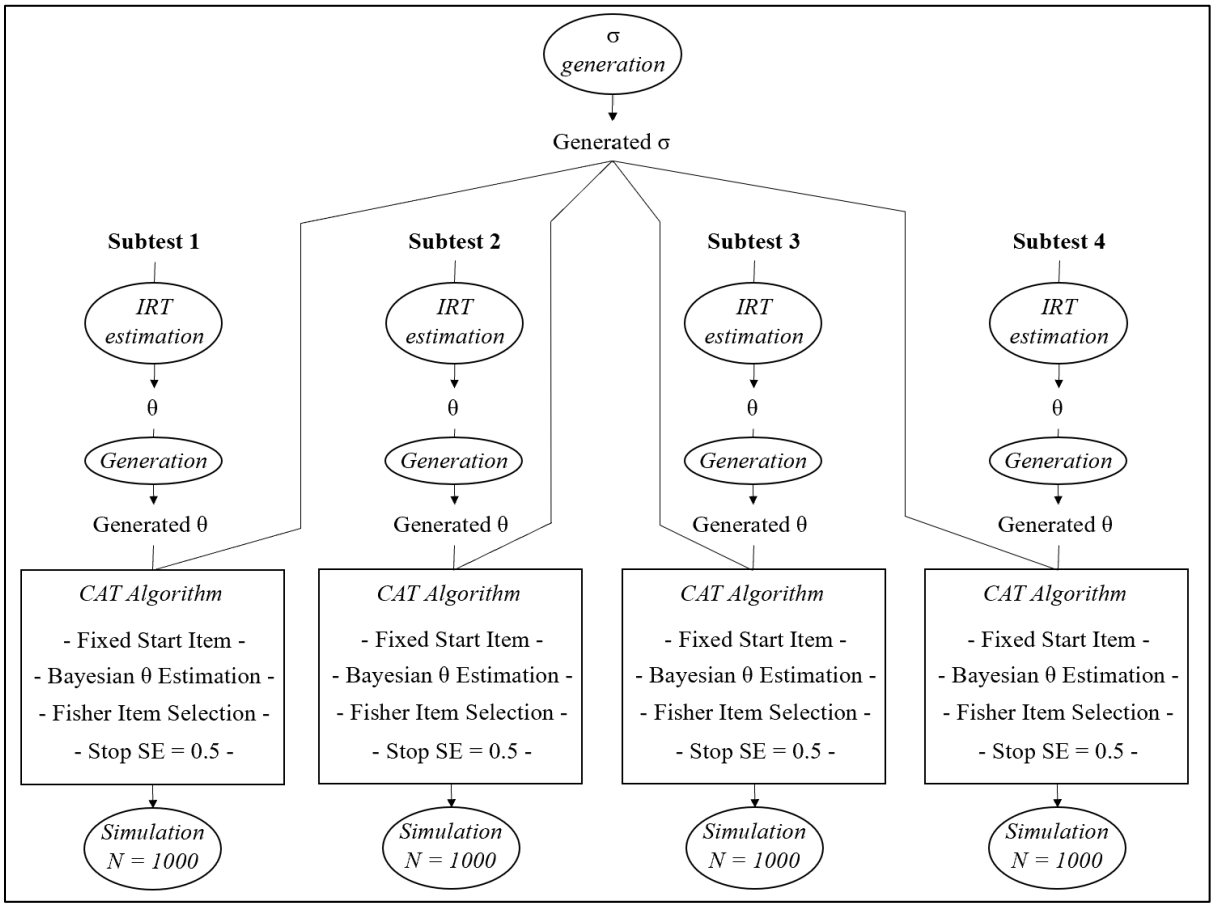


Figure B3

Simulation Process of Simulation 1 without Subtest Linking, and Simulation 2 with Subtest Linking.

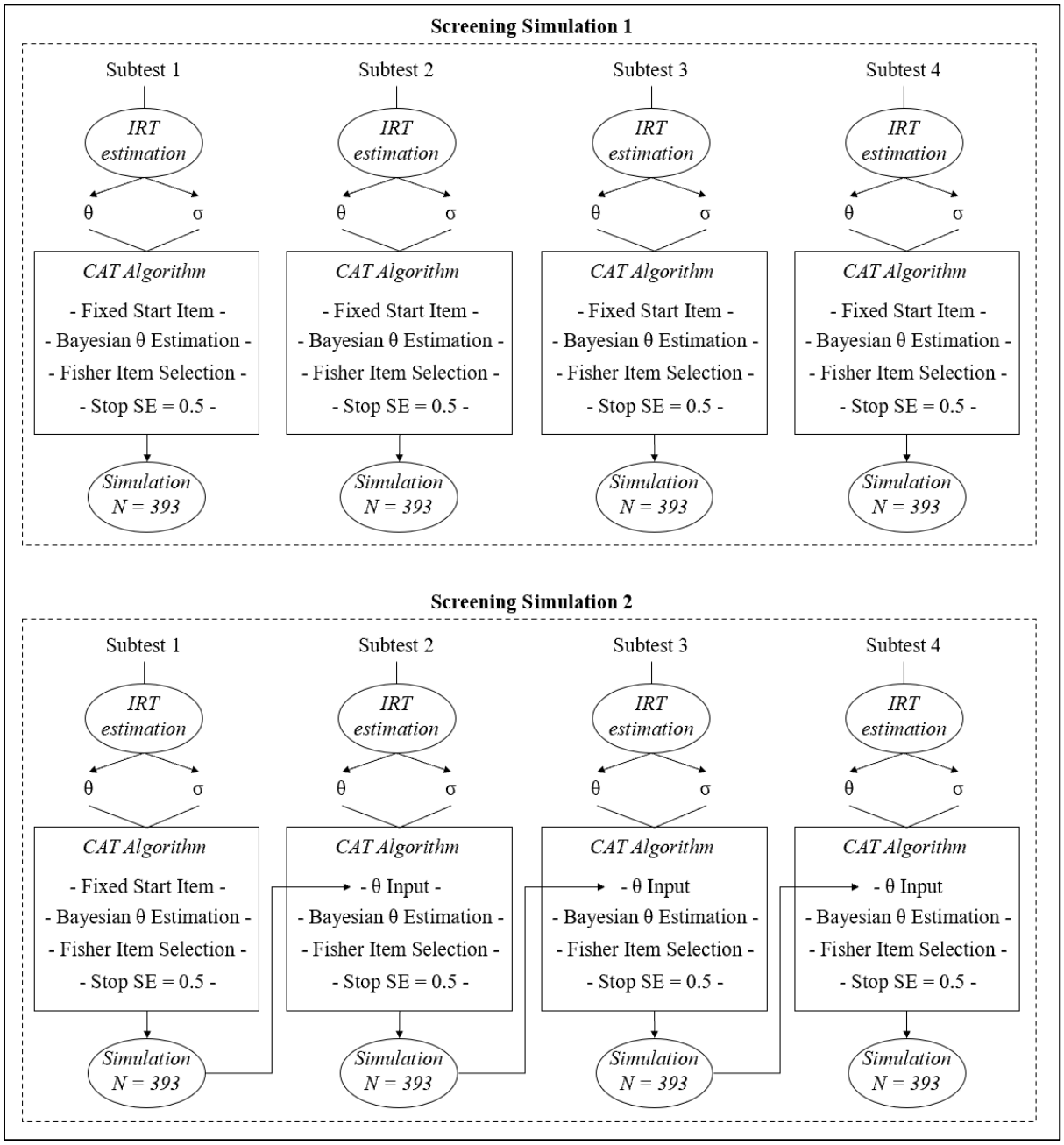


Figure B4

Simulation Process of Simulation 3 without Subtest Linking, and Simulation 4 with Subtest Linking.

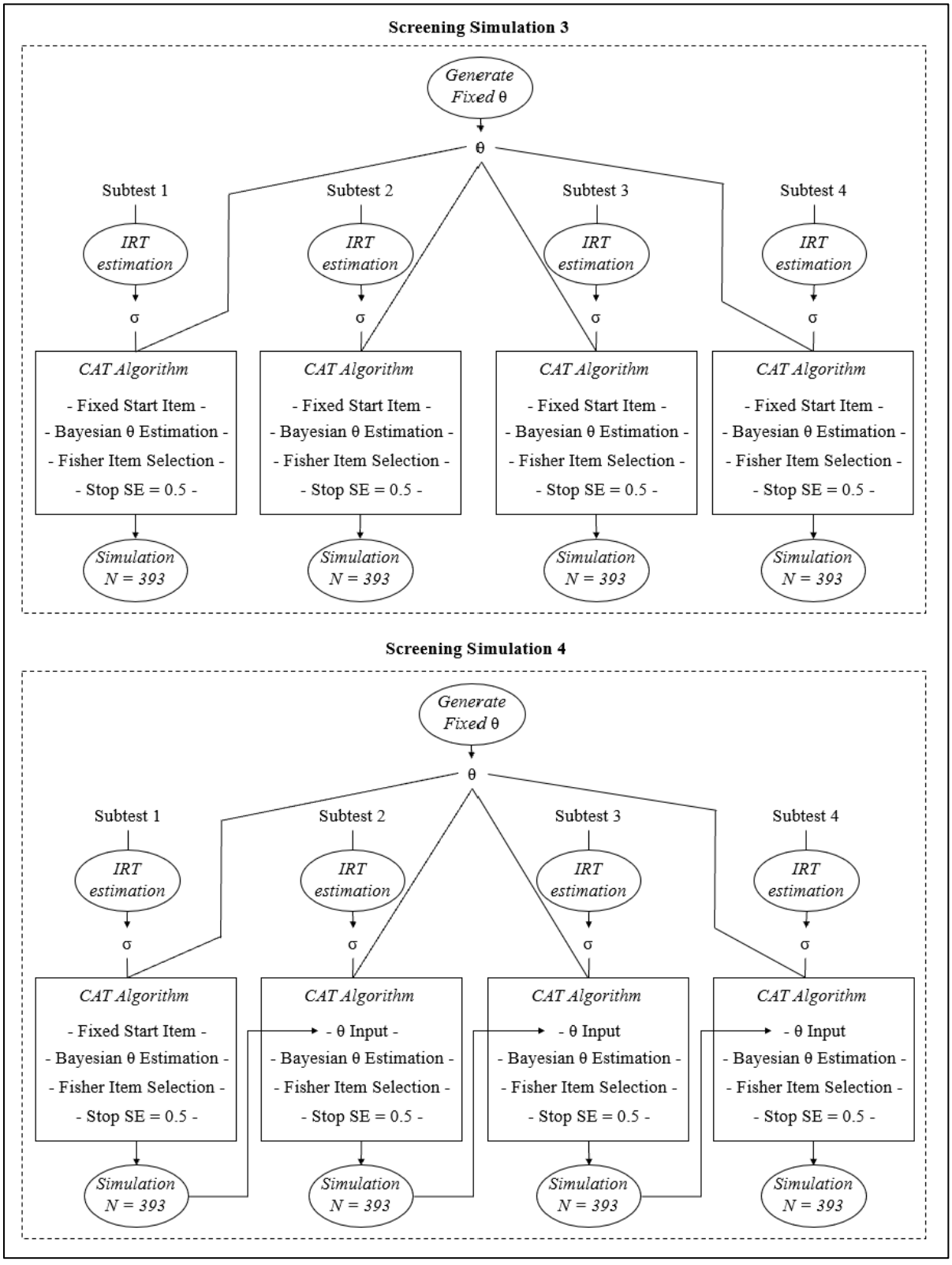
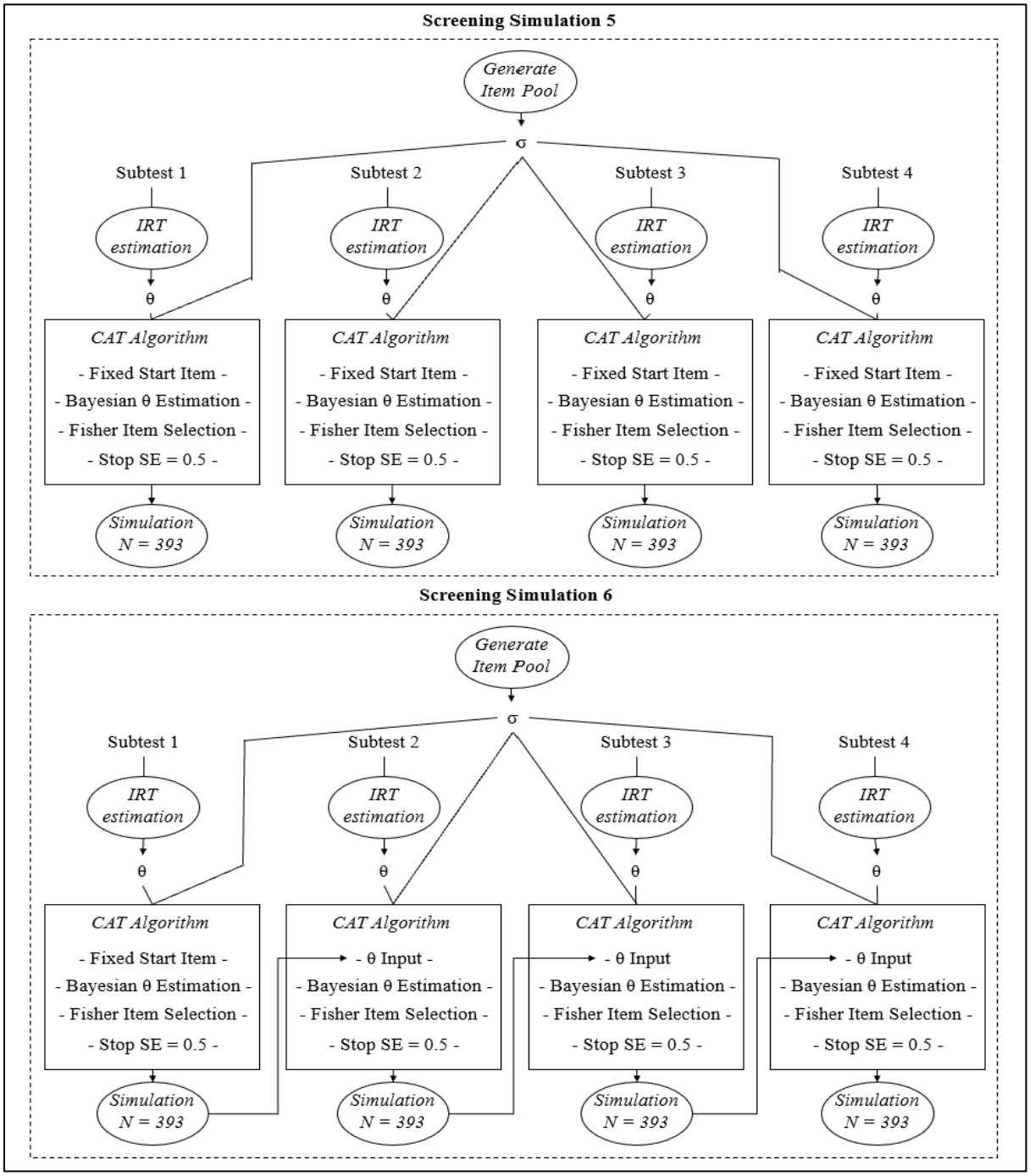


Figure B5

Simulation Process of Simulation 5 without Subtest Linking, and Simulation 6 with Subtest Linking.



C. Psychometric Analysis

Figure C1

Graphical Model Check with median split for subtest 4 before removing bad fitting items.

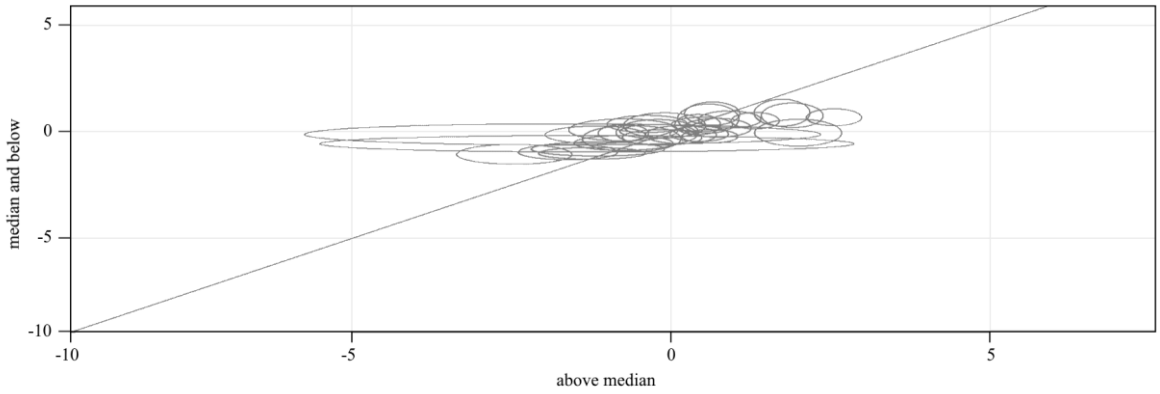


Table C1

Item Difficulties (Sigma), Outfit and Infit MSQ of all Subtests each as one-dimensional Rasch Model.

		Subtest 1	**Subtest 2**	**Subtest 3**	**Subtest 4**
Sigma	Min	-1.56	-1.66	-0.94	-1.18
	Max	1.87	2.67	2.17	0.99
	Md	0.08	0.09	-0.29	0.03
	M	0	0	0	0
	SD	1.01	0.86	0.75	0.57
Outfit MSQ	Min	0.33	0.34	0.35	0.46
	Max	1.58	1.34	1.42	1.18
	Md	0.85	0.67	0.62	0.70
	M	0.80	0.72	0.63	0.75
	SD	0.27	0.22	0.21	0.18
Infit MSQ	Min	0.66	0.73	0.68	0.72
	Max	1.26	1.28	1.31	1.16
	Md	0.91	0.87	0.84	0.90
	M	0.91	0.91	0.86	0.92
	SD	0.15	0.11	0.12	0.12

D. SubCAT Simulations

Figure D1

Scatterplot of true and estimated ability levels of subCAT 1 with SE = 0.5.

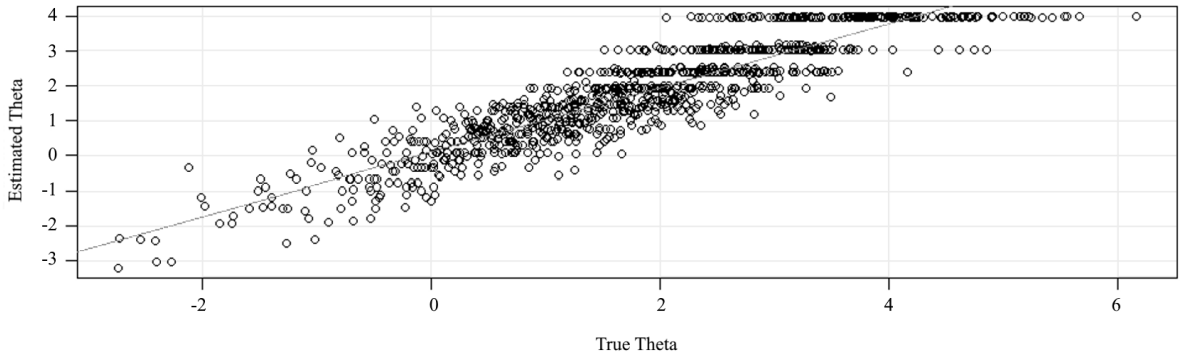


Figure D2

Scatterplot of true and estimated ability levels of subCAT 2 with SE = 0.5.

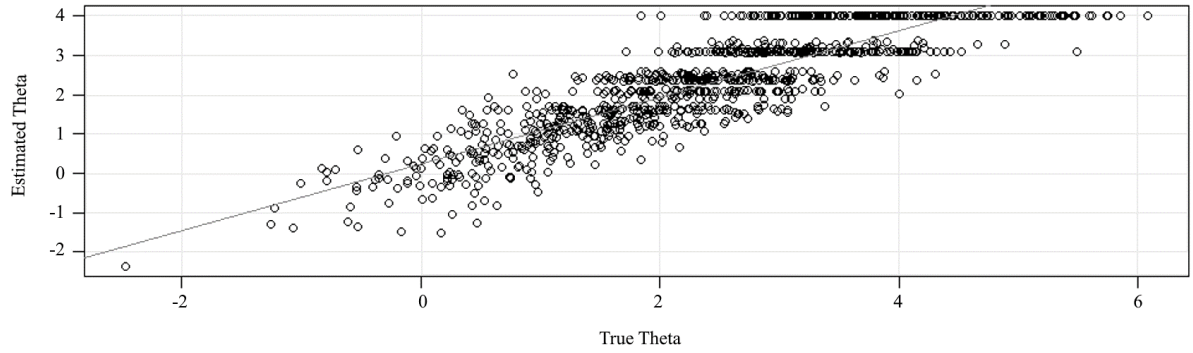


Figure D3

Scatterplot of true vs. estimated ability levels of subCAT 3 with SE = 0.5.

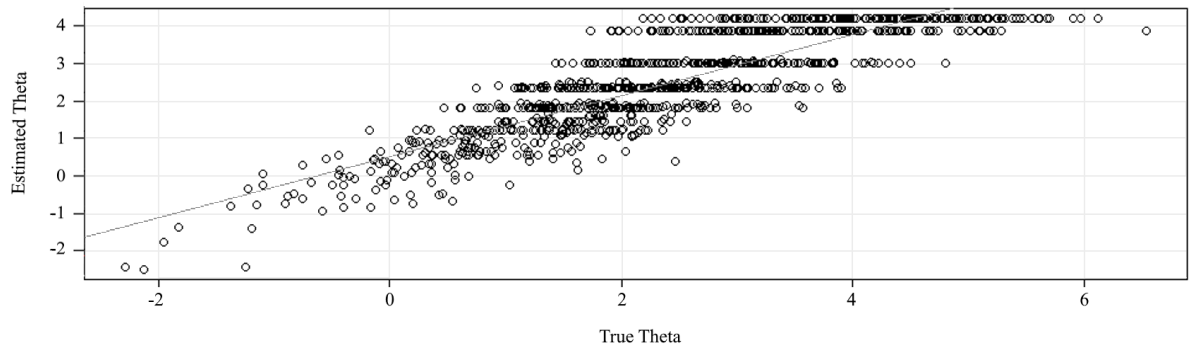


Figure D4

Scatterplot of true vs. estimated ability levels of subCAT 4 with SE = 0.5.

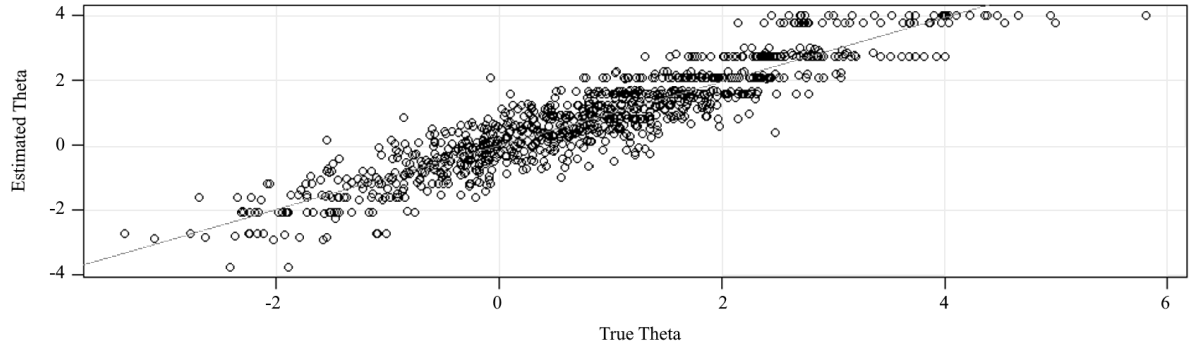


Figure D5

Conditional proportions of test runs of subCAT 1 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true thetas.

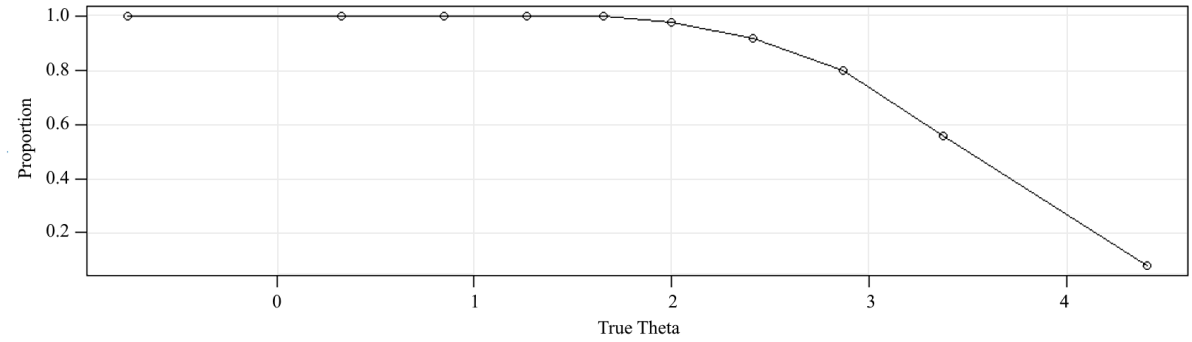


Figure D6

Conditional proportions of test runs of subCAT 2 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true ability levels.

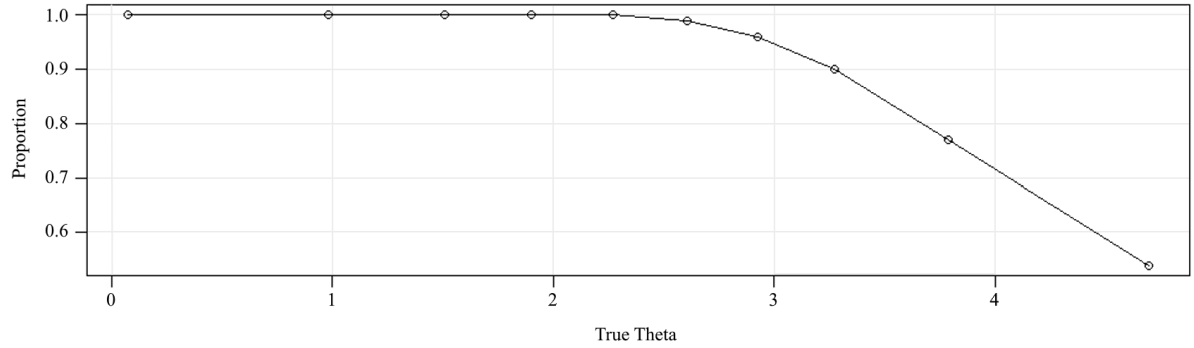


Figure D7

Conditional proportions of test runs of subCAT 3 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true ability levels.

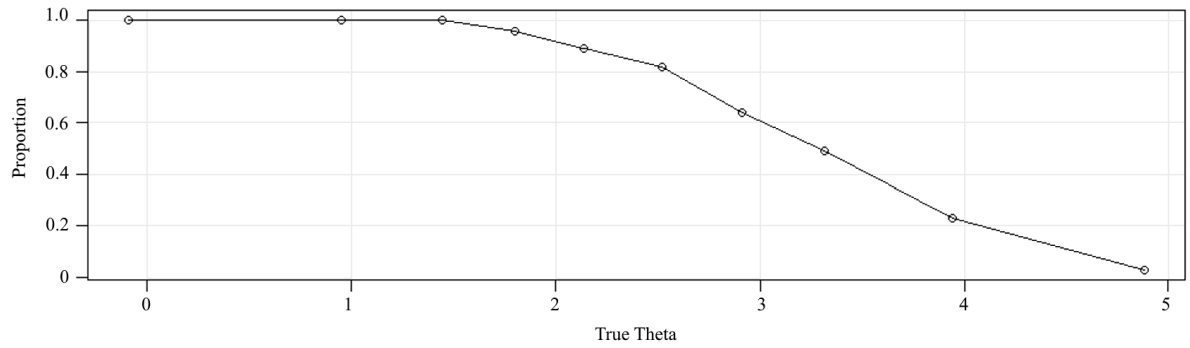


Figure D8

Conditional proportions of test runs of subCAT 4 satisfying the stopping rule SE = 0.5, as a function of the deciles of the true ability levels.

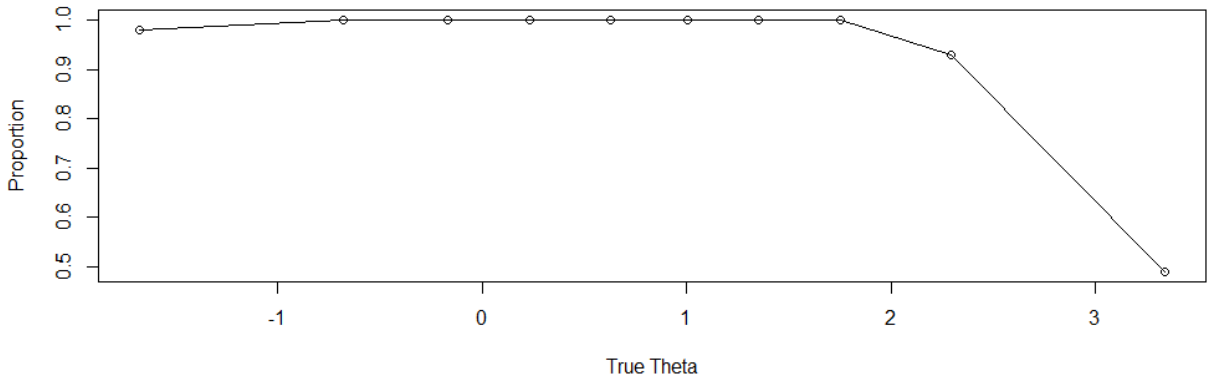


Figure D9

Test length of SubCAT 1 as a function of cumulative percent of examinees.

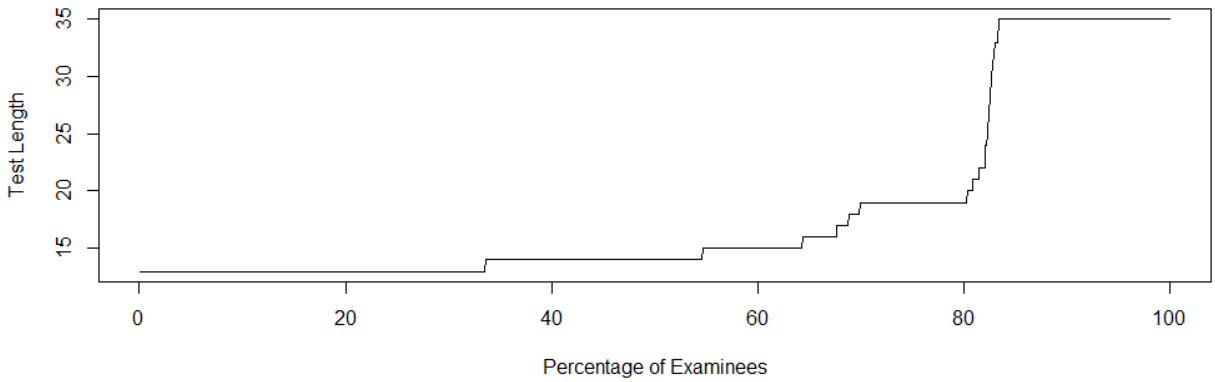


Figure D10

Test length of SubCAT 2 as a function of cumulative percent of examinees.

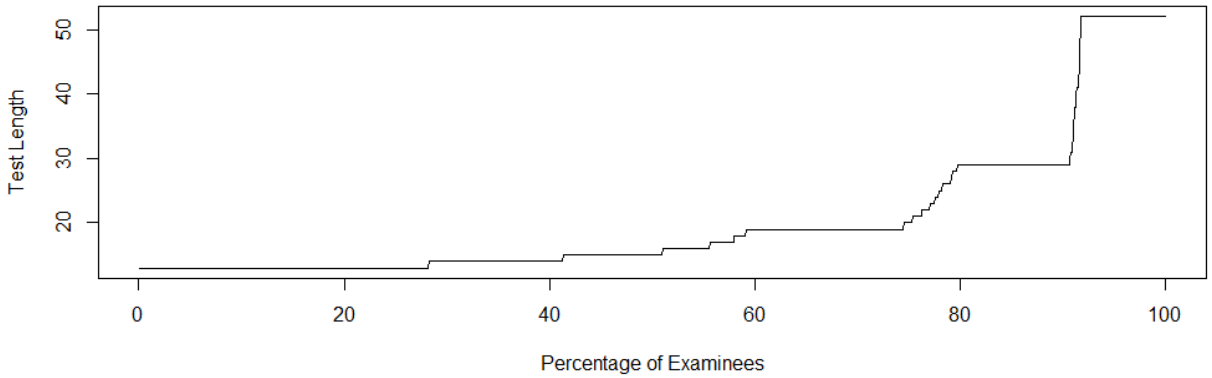


Figure D11

Test length of SubCAT 3 as a function of cumulative percent of examinees.

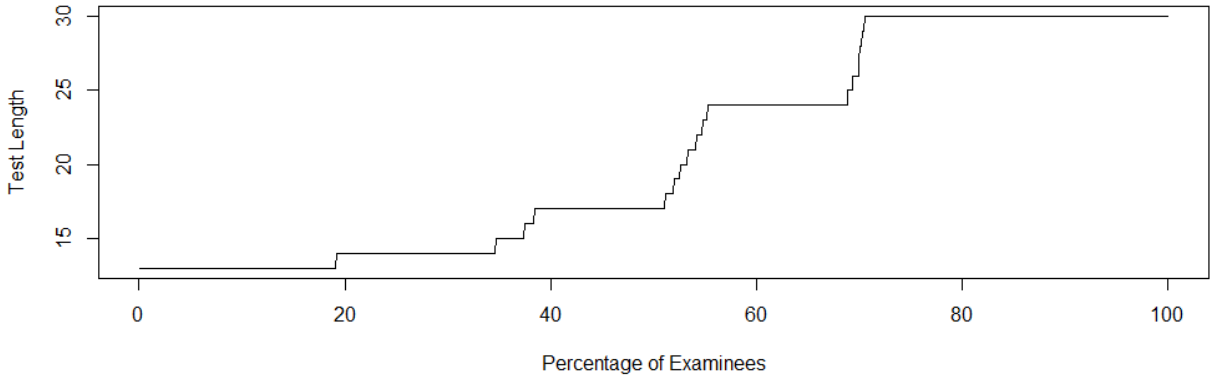


Figure D12

Test length of SubCAT 4 as a function of cumulative percent of examinees.

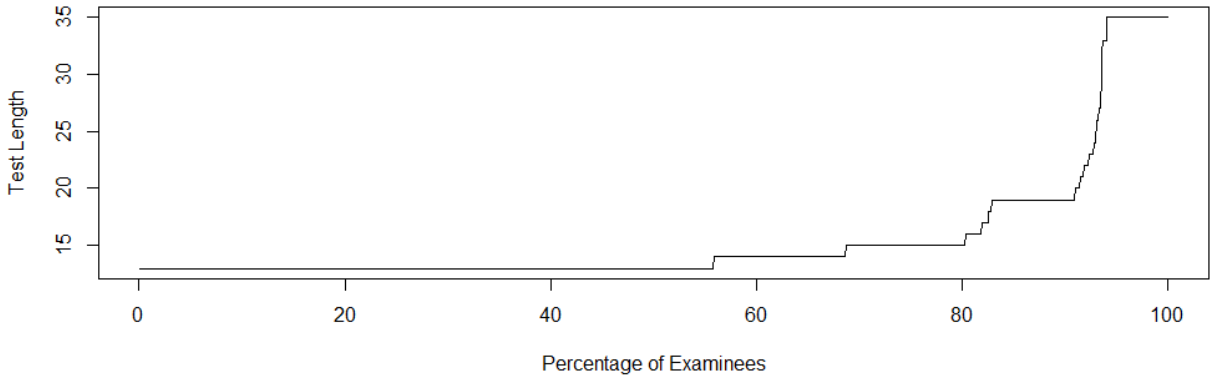


Figure D13

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = -1 for sub-test 2 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

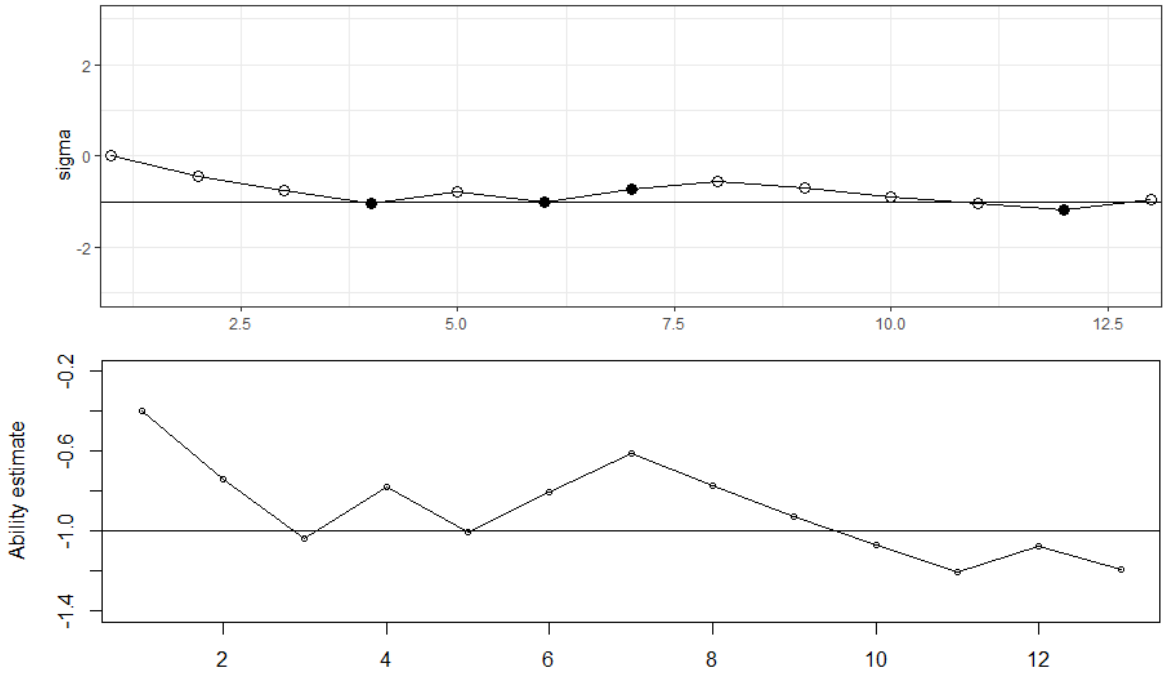


Figure D14

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = -1 for sub-test 3 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

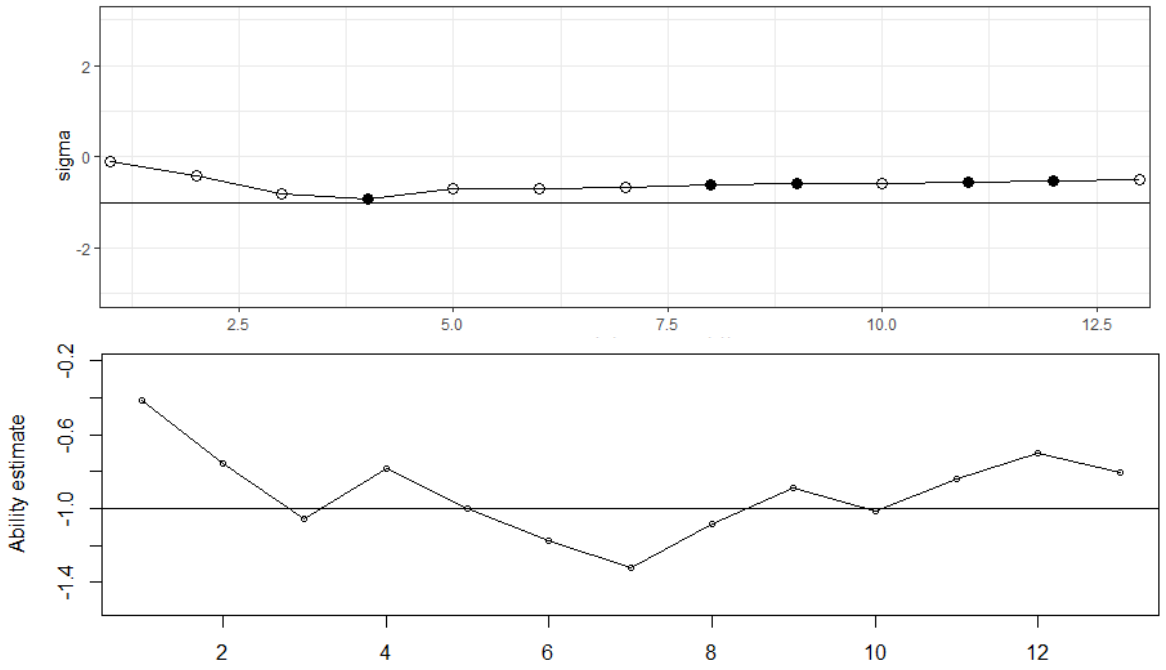


Figure D15

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = -1 for sub-test 4 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

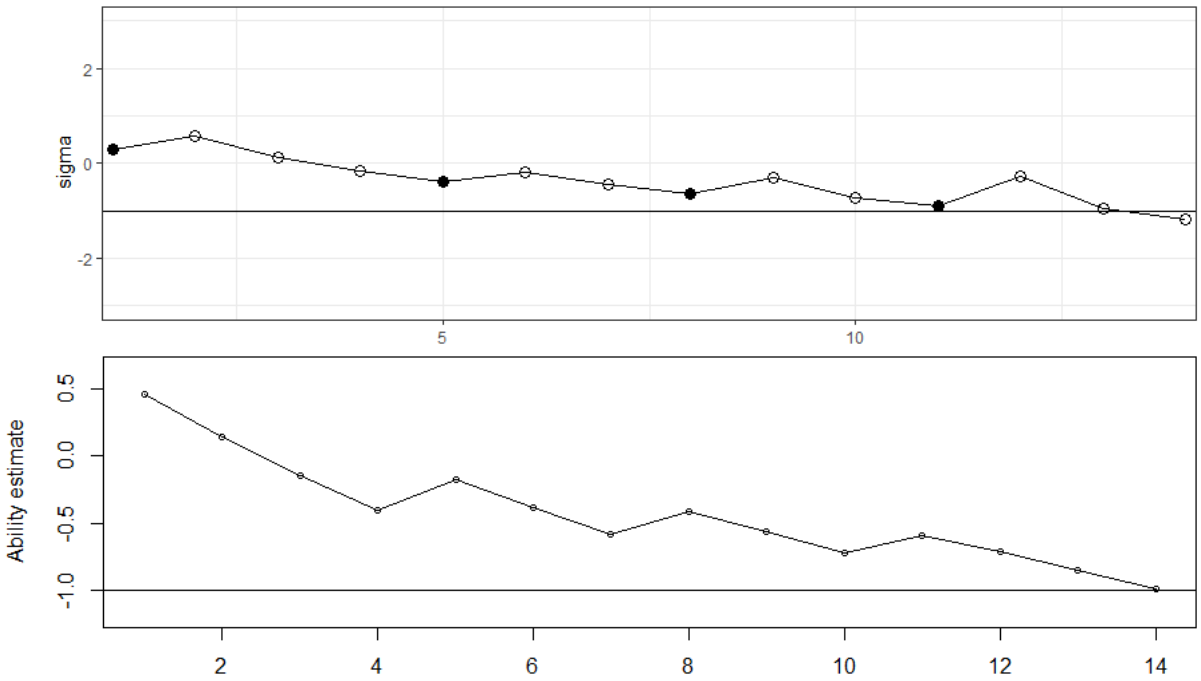


Figure D16

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = -3 for sub-test 2 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

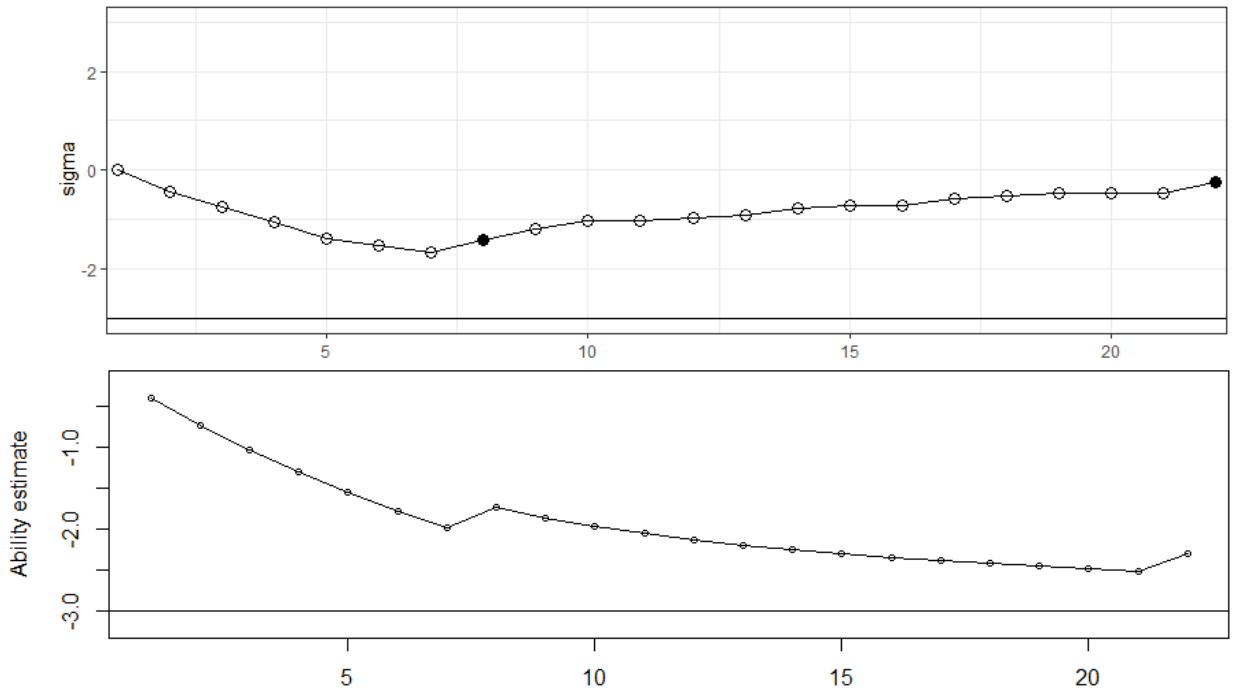


Figure D17

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = -3 for sub-test 3 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

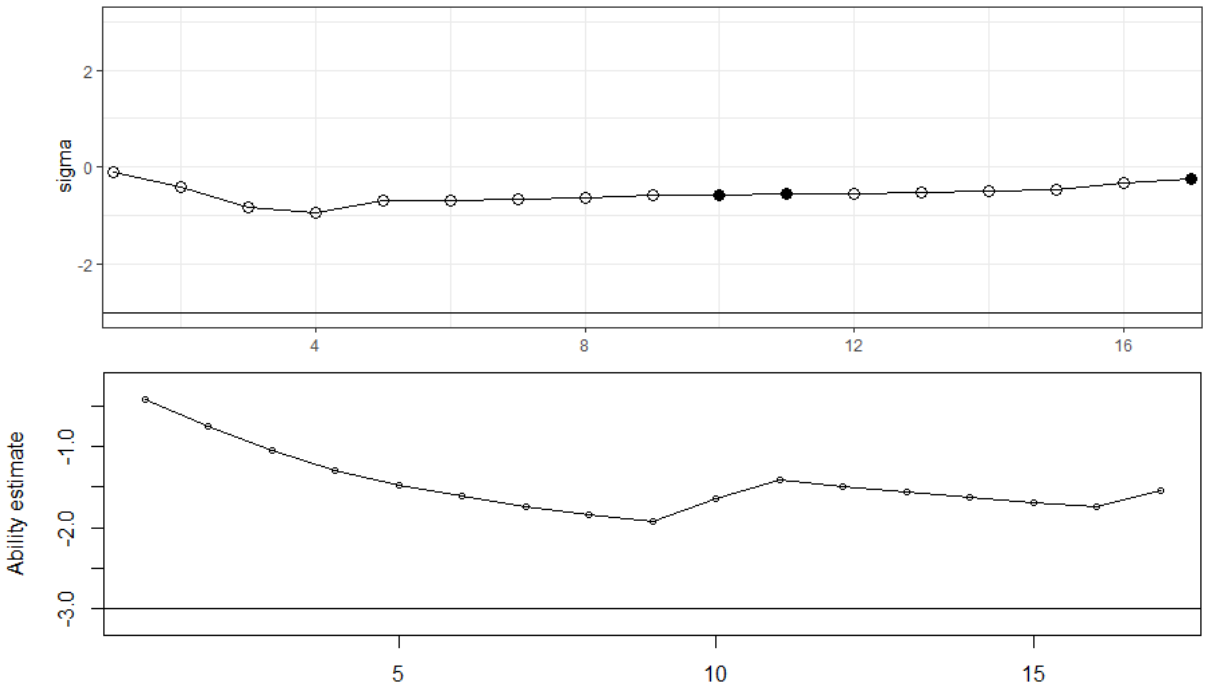


Figure D18

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = -3 for sub-test 4 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

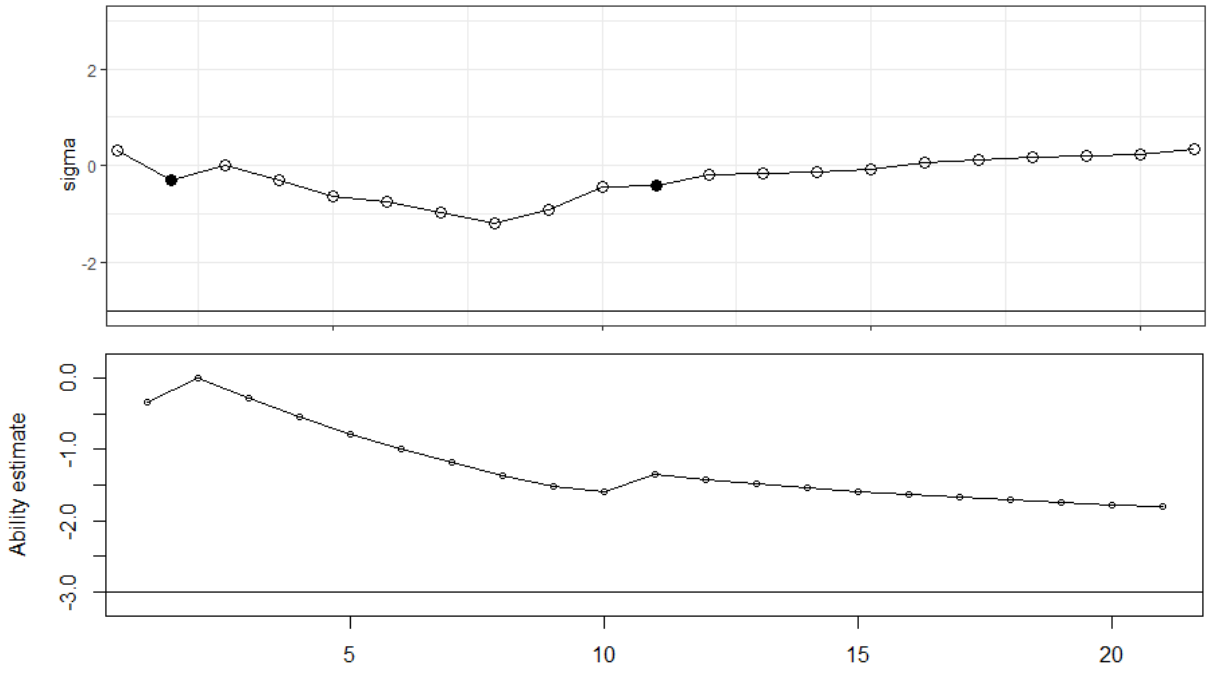


Figure D19

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = 3 for subtest 2 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

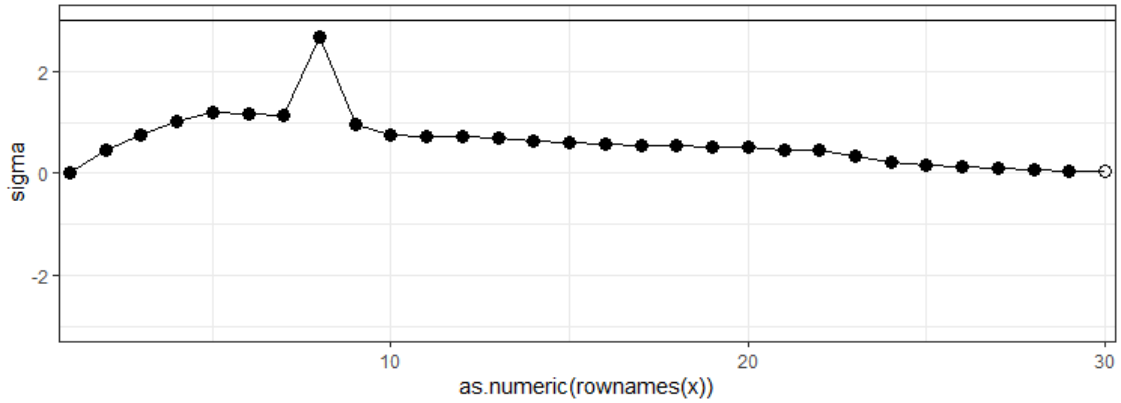


Figure D20

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = 3 for subtest 3 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

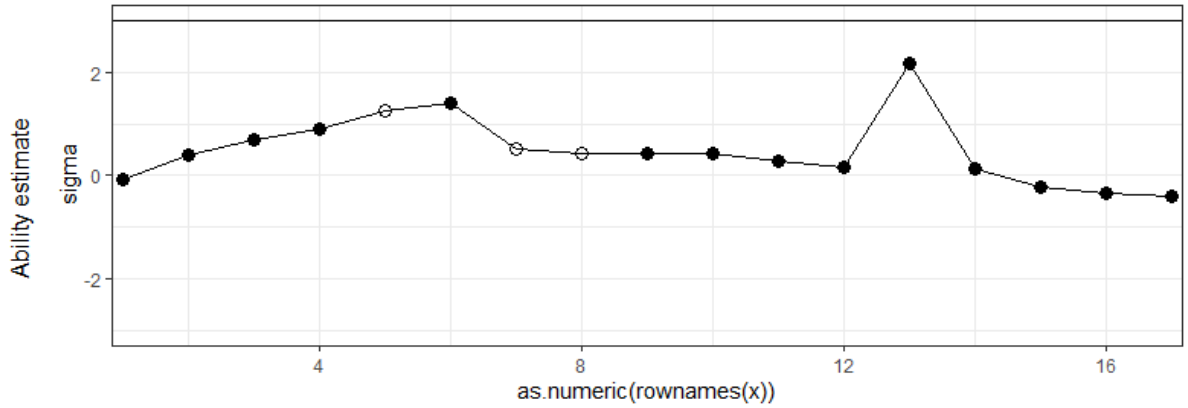


Figure D21

Item Difficulties (sigma) and theta estimates while a test run (in items) of theta = 3 for subtest 4 with black dots in the upper graph being correct answers and white dots in the upper graph being wrong answers.

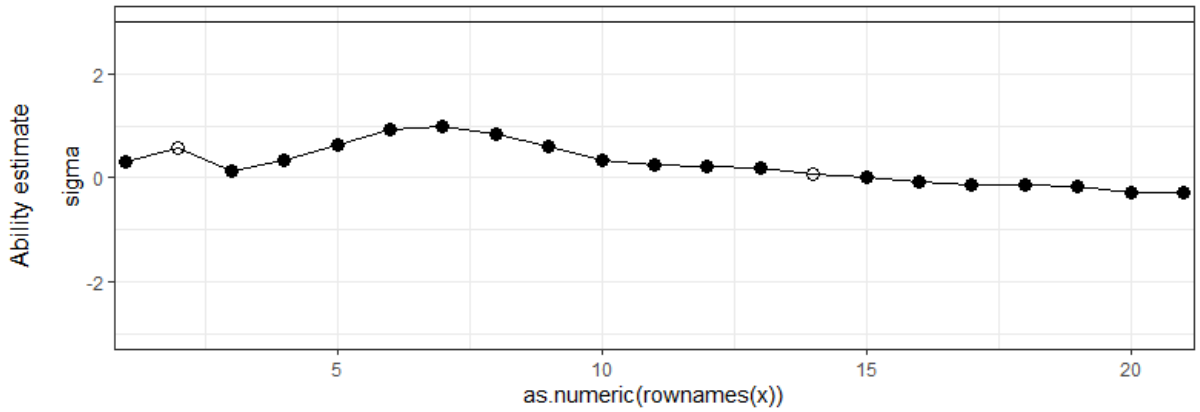


Figure D22

Scatterplot of true vs. estimated θ of the generated item pool and theta of subtest 1.

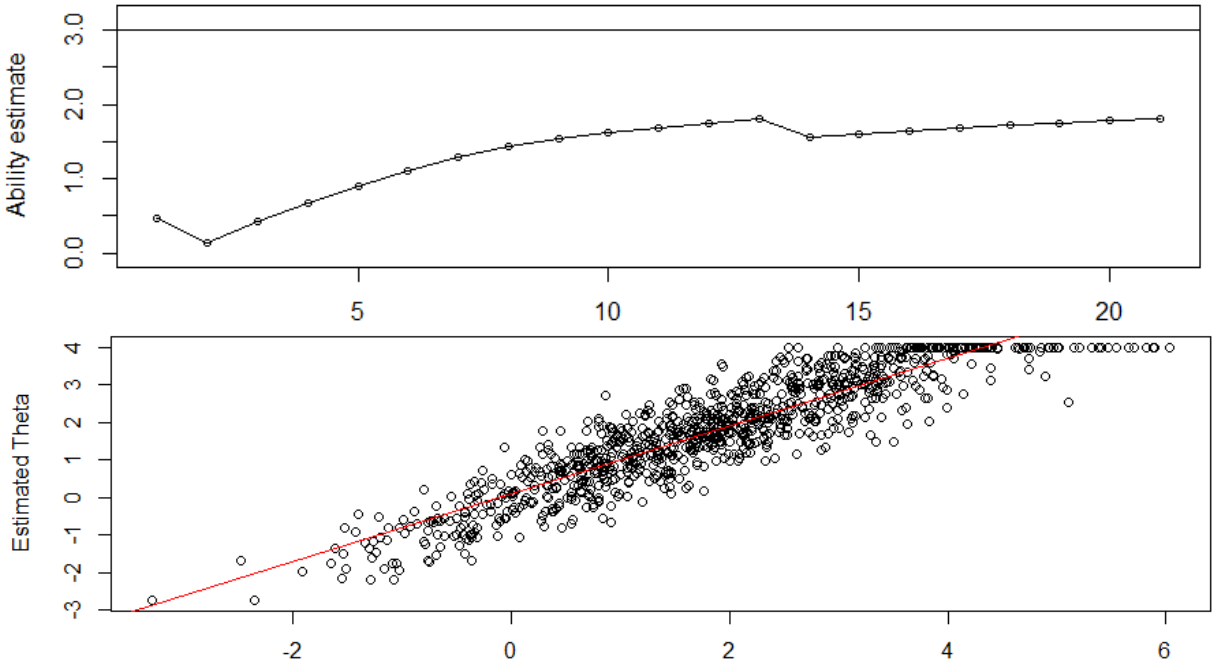


Figure D23

Scatterplot of true vs. estimated θ of the generated item pool and theta of subtest 2.

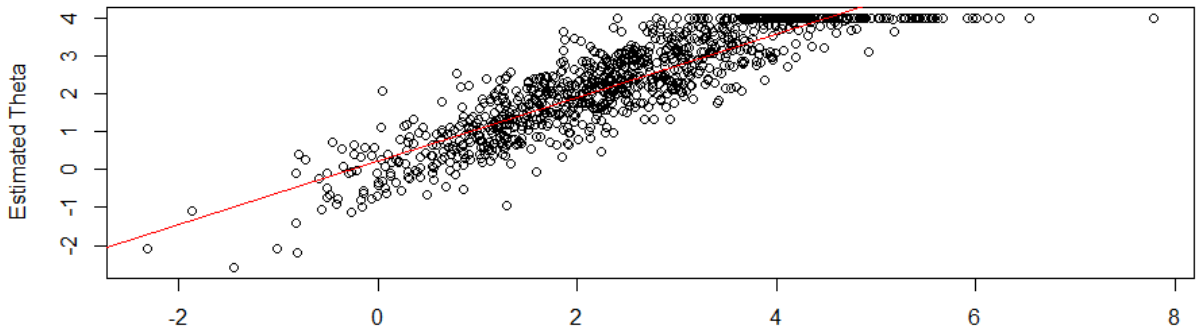


Figure D24

Scatterplot of true vs. estimated θ of the generated item pool and theta of subtest 3.

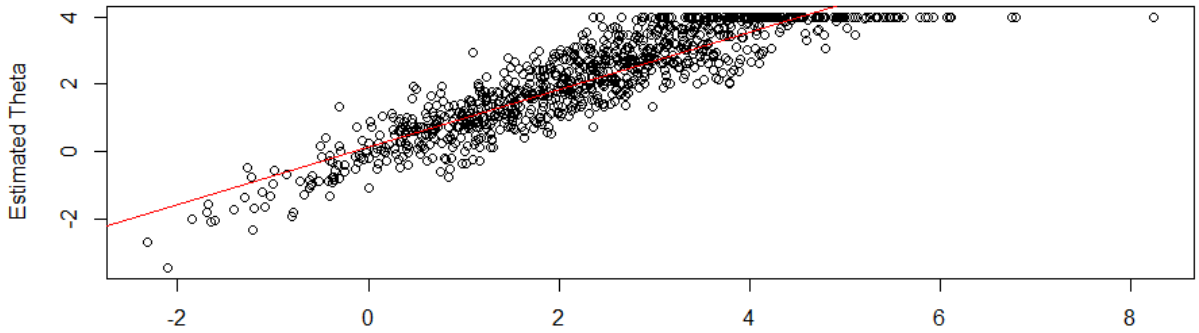


Figure D25

Scatterplot of true vs. estimated θ of the generated item pool and theta of subtest 4.

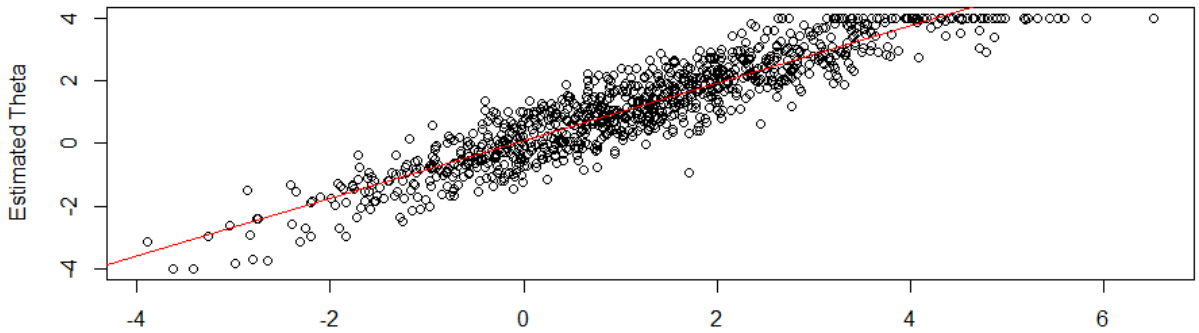


Figure D26

σ of selected items and estimated θ while a simulated test run with fixed first item of a person with true θ = 0.03 for subtest 1, true θ = 2.59 for subtest 2, true θ = 1.98 for subtest 3 and true θ = 0.23 for subtest 4.

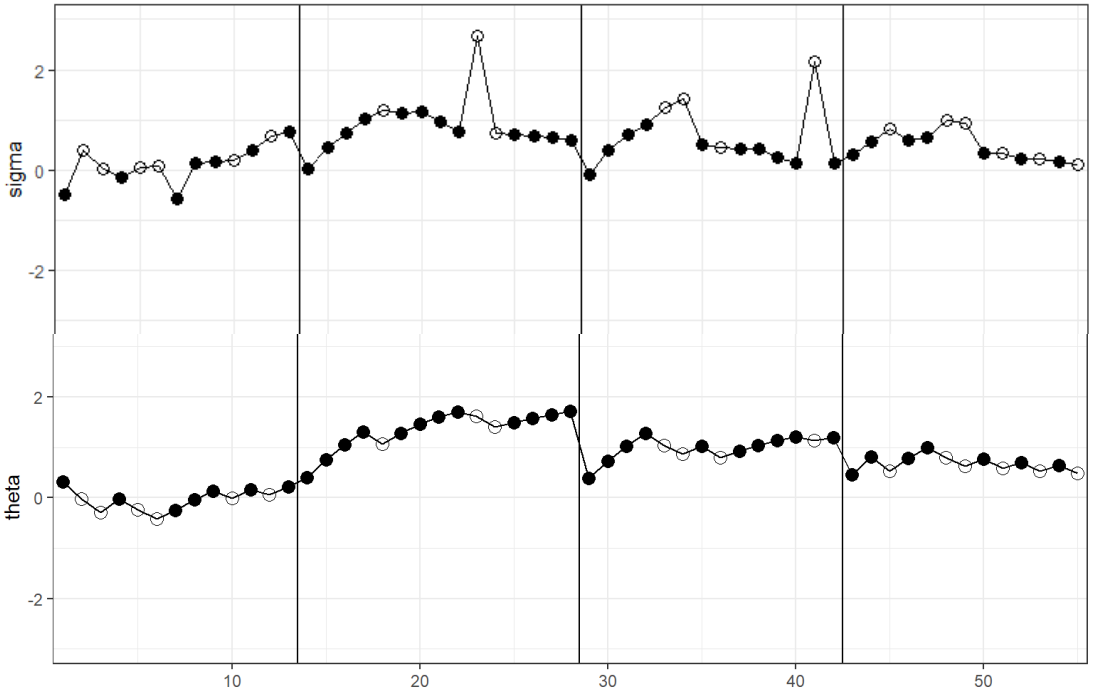
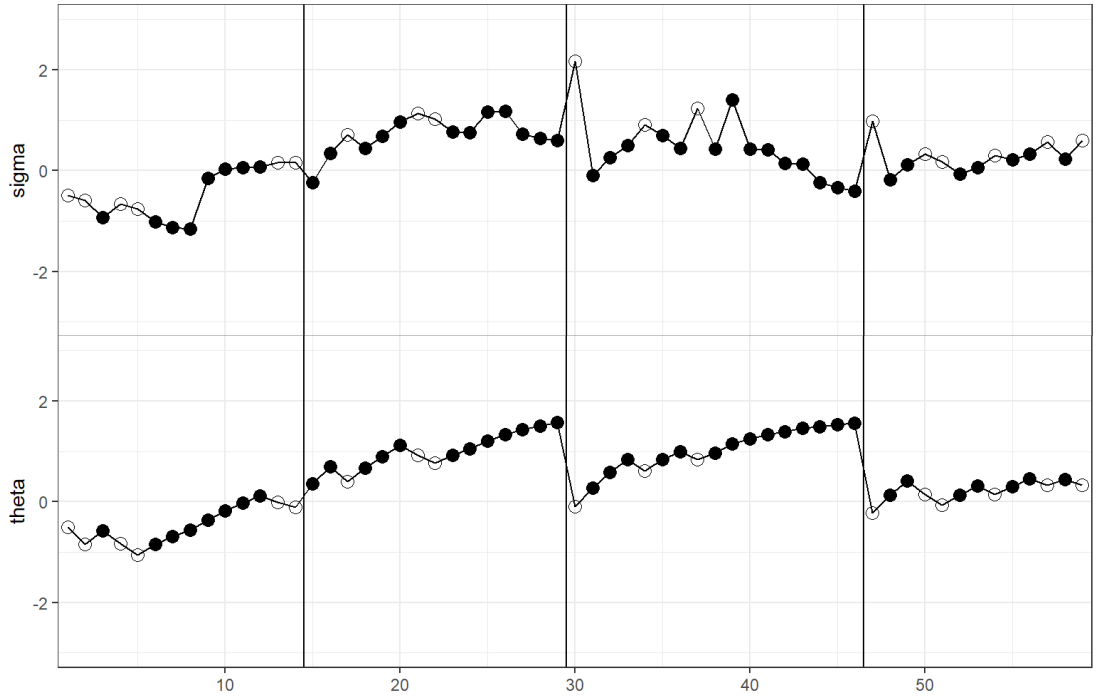


Figure D27

σ of selected items and estimated θ while a simulated test run with connected first item of a person with true θ = 0.03 for subtest 1, true θ = 2.59 for subtest 2, true θ = 1.98 for subtest 3 and true θ = 0.23 for subtest 4.



E: SubCAT Simulations of SEN Students

Table E1

Results of adaptive and linear screening simulations of real and generated item pools comparing different theta inputs of students with and without SEN and different start items.

simulation	item pool	θ	Start item	Type	*M* length	% satisfied stop	θ *cor*
1	Subtest 1	SEN	-	linear	35.00	1.00	0.94
	Subtest 2	SEN	-	linear	52.00	1.00	0.94
	Subtest 3	SEN	-	linear	30.00	1.00	0.94
	Subtest 4	SEN	-	linear	28.00	1.00	0.95
2	Subtest 1	No SEN	-	linear	35.00	1.00	0.91
	Subtest 2	No SEN	-	linear	52.00	1.00	0.90
	Subtest 3	No SEN	-	linear	30.00	1.00	0.85
	Subtest 4	No SEN	-	linear	28.00	1.00	0.92
3	Subtest 1	SEN	medium	adaptive	15.91	0.92	0.91
	Subtest 2	SEN	medium	adaptive	18.01	0.95	0.90
	Subtest 3	SEN	medium	adaptive	18.64	0.80	0.92
	Subtest 4	SEN	medium	adaptive	15.68	0.92	0.91
4	Subtest 1	No SEN	medium	adaptive	18.61	0.82	0.89
	Subtest 2	No SEN	medium	adaptive	20.56	0.92	0.85
	Subtest 3	No SEN	medium	adaptive	22.05	0.65	0.84
	Subtest 4	No SEN	medium	adaptive	17.86	0.80	0.91
5	Subtest 1	SEN	easy	adaptive	15.80	0.93	0.90
	Subtest 2	SEN	easy	adaptive	17.92	0.96	0.90
	Subtest 3	SEN	easy	adaptive	18.86	0.79	0.92
	Subtest 4	SEN	easy	adaptive	15.64	0.91	0.91
6	Subtest 1	No SEN	easy	adaptive	18.29	0.84	0.88
	Subtest 2	No SEN	easy	adaptive	20.87	0.93	0.85
	Subtest 3	No SEN	easy	adaptive	22.08	0.66	0.83
	Subtest 4	No SEN	easy	adaptive	17.88	0.80	0.90
7	Generated	SEN	-	linear	100.00	1.00	0.98
	Generated	SEN	-	linear	100.00	1.00	0.97
	Generated	SEN	-	linear	100.00	1.00	0.97
	Generated	SEN	-	linear	100.00	1.00	0.98
8	Generated	No SEN	-	linear	100.00	1.00	0.97
	Generated	No SEN	-	linear	100.00	1.00	0.95
	Generated	No SEN	-	linear	100.00	1.00	0.94
	Generated	No SEN	-	linear	100.00	1.00	0.97
9	Generated	SEN	medium	adaptive	14.00	1.00	0.91
	Generated	SEN	medium	adaptive	14.01	1.00	0.90
	Generated	SEN	medium	adaptive	15.13	1.00	0.93
	Generated	SEN	medium	adaptive	13.74	1.00	0.92
10	Generated	No SEN	medium	adaptive	14.26	1.00	0.90
	Generated	No SEN	medium	adaptive	14.70	1.00	0.87

	Generated	No SEN	medium	adaptive	16.05	1.00	0.87
	Generated	No SEN	medium	adaptive	14.62	1.00	0.92
11	Generated	SEN	easy	adaptive	13.60	1.00	0.91
	Generated	SEN	easy	adaptive	13.93	1.00	0.91
	Generated	SEN	easy	adaptive	14.24	1.00	0.93
	Generated	SEN	easy	adaptive	13.59	1.00	0.93
12	Generated	No SEN	easy	adaptive	14.11	1.00	0.90
	Generated	No SEN	easy	adaptive	14.42	1.00	0.86
	Generated	No SEN	easy	adaptive	14.80	1.00	0.86
	Generated	No SEN	easy	adaptive	14.22	1.00	0.92