Speededness in Achievement Testing: Relevance, Consequences, and Control

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

am Fachbereich für Psychologische Methodenlehre

Lebenswissenschaftliche Fakultät

der Humboldt Universität zu Berlin

vorgelegt von

Benjamin Becker, M.Sc.

Präsidentin der Humboldt-Universität zu Berlin

Prof. Dr. Julia von Blumenthal

Dekan der Lebenswissenschaftlichen Fakultät

Prof. Dr. Dr. Christian Ulrichs

Gutachter:innen:

1. Prof. Dr. Manuel Völkle
2. Prof. Dr. Frank Goldhammer
3. Prof. Dr. Hans Anand Pant

Tag der mündlichen Prüfung: 30.10.2023

Danksagung

Für die Unterstützung bei meiner Promotion und Begutachtung dieser Arbeit möchte ich mich bei Prof. Dr. Manuel Völkle, Prof. Dr. Frank Goldhammer sowie Prof. Dr. Hans Anand Pant herzlichst bedanken.

Besonderer Dank gebührt auch Prof. Dr. Petra Stanat, Dr. Sebastian Weirich und Dr. Malte Jansen für das Ermöglichen dieses Promotionsvorhabens am IQB. Insbesondere bin ich dankbar für die Unterstützung und die Freiheiten, die mir das IQB bezüglich dieser Promotion ermöglicht hat: bei der Themenwahl dieser Arbeit, externen Kooperationen, bei der Teilnahme an Konferenzen, meinem Forschungsaufenthalt sowie den zeitlichen Ressourcen während der Fertigstellung dieser Arbeit.

Dr. Sebastian Weirich möchte ich darüber hinaus für die stets engagierte Unterstützung und Geduld in allen Lebenslagen danken: Sei es bei der Betreuung dieser Arbeit, als (ehemaliger) Vorgesetzter im Methoden-Projekt oder bei Fahrrad-Reparaturen in der Tiefgarage.

Bei Prof. Dr. Steffi Pohl möchte ich mich an dieser Stelle für das Wecken meiner Begeisterung für die Psychometrie und das Heranführen an das wissenschaftliche Arbeiten danken. Ich habe in meiner Zeit an der FU Berlin, neben vielem Fachlichen, viel über Zielstrebigkeit und wissenschaftliches Netzwerken gelernt.

Bei allen Ko-Autor:innen der Einzelbeiträge möchte ich mich ebenfalls bedanken, da ohne sie die Erstellung dieser Arbeit nicht dasselbe gewesen wäre: Dr. Dries Debeer, Dr. Sebastian Weirich, Prof. Dr. Frank Goldhammer, Dr. Peter van Rijn, Dr. Dylan Molenaar, Dr. Christoph König, Dr. Esther Ulitzsch sowie Dr. Karoline Sachse.

Besonderer Dank gilt dabei Dr. Dries Debeer, für die Begeisterung für mein Promotionsthema, die engagierte Unterstützung aller meiner Forschungsvorhaben, die zahllosen psychometrischen Diskussionen und die viele gemeinsame Arbeit auch in Form von Workshops. Ohne ihn wäre diese Arbeit nicht möglich gewesen.

Auch Dr. Peter van Rijn möchte ich einen besonderen Dank für die Ermöglichung meines Forschungsaufenthalts bei ETS Global aussprechen, der für mich äußerst inspirierend und motivierend war.

Darüber hinaus möchte ich mich bei Sandra Becker, Dr. Karoline Sachse sowie Dr. Esther Ulitzsch für das akribische und kritische Korrekturlesen dieser Arbeit bedanken.

Abschließend bedanke ich mich bei allen Kolleg:innen des IQB, die mich während dieser Promotion begleitet und unterstützt haben, mit zahlreichen Mensa- oder Kaffeemobil-Besuchen, fachlichen, sportlichen und politischen Diskussionen sowie inspirierenden Rahmungs-Playlists.

Abstract

Educational and psychological testing is an important cornerstone of modern educational systems. As examinations and assessments are often used to control access to educational programs and to assess successful participation in an educational program, their fairness and validity is of great importance. A controversially discussed aspect of standardized tests is setting time limits on tests and how this practice can result in test speededness. Indeed, different perspectives on whether tests should be speeded and whether speed should be seen as a part of measured constructs exist. Regardless of these perspectives, being able to deliberately control the speededness of tests is desirable. If a test is intended to be speeded, test designers should be able to deliberately control the degree of speededness. If a test intended to be unspeeded but a time limit is set out of practical considerations, test designers should be able to provide an unspeeded test administration for all test-takers. If multiple, parallel test forms are used interchangeably, test designers must be able to guarantee an equal degree of speededness on all test forms.

For these purposes, van der Linden (2011a, 2011b) proposed an approach to control the speededness of tests in *automated test assembly* (ATA) using mixed integer linear programming and a lognormal response time model. The approach can be used in fixed-form linear tests, computer-adaptive testing as well as multi-stage testing and is relevant for high- as well as low-stakes assessments. However, the approach by van der Linden (2011a, 2011b) has an important limitation, in that it is restricted to the two-parameter lognormal response time model which assumes equal speed sensitivities (i.e., factor loadings) across items. This thesis demonstrates that otherwise parallel test forms with differential speed sensitivities are indeed unfair for specific test-takers. Furthermore, an extension of the van der Linden approach is introduced, which incorporates speed sensitivities in ATA. Additionally, test speededness can undermine the fairness of a test if identical but differently ordered test forms are used. To prevent that the score of test-takers depends on whether easy or difficult items are located at the end of a test form, it is proposed that the same, most time intensive items should be placed at the end of all test forms. Currently, there is a lack of designated software implementations and guides on response time modeling and ATA. Therefore, the thesis provides introductions and tutorials on using the `R` package `eatATA` for ATA and using `Stan` and `rstan` for Bayesian hierarchical response time modeling. Finally, the thesis discusses alternatives, practical implications, and limitations of the proposed approaches and provides an outlook on future related research topics.

Zusammenfassung

Prüfungen und Tests sind wichtige Eckpfeiler moderner Bildungssysteme. Da Prüfungen und Tests häufig dazu dienen, den Zugang zu Bildungsprogrammen zu steuern und die Grundlage zur Abschlussvergabe am Ende von Bildungsprogrammen bilden, ist ihre Fairness und Validität von größter Bedeutung. Ein kontrovers diskutierter Aspekt standardisierter Tests ist die Verwendung von Zeitlimits und die Frage, inwiefern Zeitlimits zu Zeitdruck aufseiten von Testteilnehmenden führen können. In der Tat wird innerhalb der psychometrischen Forschung kontrovers diskutiert, ob Tests unter Zeitdruck administriert werden sollten und ob die Testbearbeitungs-Geschwindigkeit als Teil der gemessenen Konstrukte betrachtet werden sollte. Unabhängig von diesen Überlegungen sollten Testentwickler:innen in die Lage versetzt werden, den Zeitdruck einer Testadministrationen bewusst gestalten zu können. Wenn ein Test eine substanzielle Speed-Komponente haben sollte, sollten Testentwickler:innen in der Lage sein, den Grad des Zeitdrucks zu steuern. Wenn ein Test keine Speed-Komponente haben sollte, aber aus praktischen Erwägungen ein Zeitlimit gesetzt wird, sollten Testentwickler:innen sicherstellen können, dass keine Testteilnehmenden unter Zeitdruck arbeiten müssen. Wenn unterschiedliche, parallele Testhefte verwendet werden, sollten Testentwickler:innen in der Lage sein, für alle Testhefte ein gleiches Maß an Zeitdruck zu gewährleisten.

Zu diesem Zweck schlägt van der Linden (2011a, 2011b) einen Ansatz zur Kontrolle des Zeitdrucks von Tests in der automatisierten Testhefterstellung (ATA) unter Verwendung von Mixed Integer Linear Programming (MILP) und eines lognormalen Antwortzeitmodells vor. Der Ansatz kann bei konventionellen linearen Tests, bei computeradaptiven Tests sowie bei multi-stage Tests verwendet werden und ist sowohl für low-stakes als auch high-stakes Tests relevant. Der Ansatz von van der Linden (2011a, 2011b) hat jedoch eine zentrale Limitation: Er ist auf das zwei-parametrische lognormale Antwortzeitmodell beschränkt, das gleiche Geschwindigkeits-Sensitivitäten (d.h. Faktorladungen) für alle Items annimmt. Diese Arbeit zeigt, dass ansonsten parallele Testhefte mit unterschiedlichen Geschwindigkeits-Sensitivitäten für bestimmte Testteilnehmende tatsächlich unfair sind. Darüber hinaus wird eine Erweiterung des van der Linden-Ansatzes vorgestellt, die unterschiedliche Geschwindigkeits-Sensitivitäten von Items in ATA berücksichtigt. Zusätzlich kann Zeitdruck ein wichtiges Fairness-Problem darstellen, wenn Testhefte mit identischen, aber unterschiedlich angeordneten Items verwendet werden. Um zu verhindern, dass die Punktzahl der Testteilnehmenden davon abhängt, ob sich leichte oder schwierige Items am Ende eines Testhefts befinden, wird vorgeschlagen, dass die zeitintensivsten Items am Ende aller Testhefte platziert werden soll-

ten. Derzeit gibt es einen Mangel an spezifischen Software-Implementationen und Leitfäden zur Antwortzeitmodellierung und ATA. Daher bietet die Arbeit Anleitungen zur Verwendung des `R`-Pakets `eatATA` für ATA und zur Verwendung von `Stan` und `rstan` für Bayesianische hierarchische Antwortzeitmodellierung. Abschließend werden Alternativen, praktische Implikationen und Grenzen der vorgeschlagenen Ansätze diskutiert und Vorschläge für zukünftige Forschungsthemen gemacht.

Contents

List of Abbreviations

1PL model	one-parameter logistic IRT model
2PL model	two-parameter logistic IRT model
2PLN model	two-parameter lognormal model
3PLN model	three-parameter lognormal model
APIs	application programming interfaces
ATA	automated test assembly
B-GLIRT	bivariate generalized linear item response theory
CAT	computerized adaptive testing
CBA	computer-based assessment
CDM	cognitive diagnostic modeling
CFA	confirmatory factor analysis
DIF	differential item functioning
eatATA	educational assessment tools: Automated Test Assembly (R-package)
ESS	effective sample size
GMITC	generalized maximum information per time unit criterion
GRE	Graduate Record Examinations
IIF	item information function
IRT	item response theory
LSA	large-scale assessment
LSAT	Law School Admission Test
MIC	maximum information criterion
MILP	mixed integer linear programming
MITC	maximum information per time unit criterion
MST	multi-stage testing
PIAAC	Programme for the International Assessment of Adult Competencies
PIRLS	Progress in International Reading Study
PISA	Programme of International Student Assessment
PPC	posterior predictive checks
SAbT	speed-ability trade-off
SEM	structural equation modeling
SRT	signed residual time
STA	shadow-test approach

SWD	students with disabilities
TBA	technology-based assessment
TIF	test information function
TIMSS	Trends in International Mathematics and Science Study
TOEFL	Test of English as a Foreign Language
WAIC	widely applicable information criterion
WLE	weighted likelihood estimation

1 Theoretical Background

Educational and psychological achievement tests are among the cornerstones of modern educational systems. Throughout their school careers, students are tested on a regular basis to monitor their abilities and knowledge gains. This assessment practice continues in vocational or higher educational programs. In these contexts, assessments can serve a variety of purposes: First, achievement tests are used by teachers, institutions, and organizations to measure whether applicants have the required skill set or knowledge to attend an educational program or work in a specific profession. Examples of such assessments include college admission tests and assessment centers used in applicant selection. Second, achievement tests are used to assess whether someone has attended an educational program successfully, for instance licensing or university exams. Such assessments are frequently referred to as *summative assessments* (Dixson & Worrell, 2016; Dolin et al., 2018). Third, *formative assessments* are important tools for teachers and trainers to monitor and aid learning processes (Dixson & Worrell, 2016; Dolin et al., 2018). Finally, assessments are used on the institutional or policy level, for example to evaluate and compare the success of different educational systems. Due to this variety of purposes, measuring cognitive ability and achievement has always played a major role in the history of psychometrics and – by extension – in the history of psychological and educational sciences (Jones & Thissen, 2006).

According to the *Standards for Educational and Psychological Testing* by the American Educational Research Association et al. (2014), the three fundamental requirements for achievement tests are validity, reliability, and fairness. However, in practical administrations, the validity, reliability, and fairness of an assessment can be threatened for a multitude of reasons, even if the test itself is perfectly designed. For instance, test-takers may cheat on assessments (Bernardi et al., 2008), test-takers may not fully engage with the test due to lack of motivation (Wise & DeMars, 2005), or a test administration may be disturbed due to external factors such as a fire alarm. Another such threat, especially for the validity and fairness of assessments, stems from the use of time limits and speededness (Y. Lu & Sireci, 2007). Almost all formalized assessments use time limits to guarantee comparable testing conditions for test-takers and out of practical considerations. However, if a time limit is set, this means that some test-takers may not have sufficient time to work on the assessment to their fullest ability while other, faster test-takers do have ample time. The term *test speededness* refers to the phenomenon that a test-taker would have performed better on a test, would they have been given more time (Cintron, 2021).

To illustrate how speededness can threaten the validity and fairness of an assessment consider the following two examples: First, assume a foreign language teacher constructs a test to assess the language skills of a class they have just been assigned to. The teacher tries to cover various aspects of language skills, such as reading fluency, spelling, grammar, and vocabulary. As the teacher only has weekly one-hour lessons with the class, the teacher administers the test in one of these one-hour lessons. During test administration, some students easily finish within the time limit. However, some students struggle to finish on time, especially those with slow reading speed. Inadvertently and without bad intentions, the teacher has created a test which is speeded for some test-takers. Instead of measuring students' foreign language skills, the test measures whether students are able to answer foreign language questions in a rapid fashion. If a student scores low on the spelling section of the test, this could mean that either (a) they have poor spelling or (b) they were working too slowly on the test. Therefore, the assessment cannot be considered valid for its intended purpose, due to its speededness.

Second, assume a university which uses an admission test to determine which applicants are suited to attend a study program. The university seeks to measure whether potential students can solve math problems efficiently in a limited amount of time. To prevent students from collaborating or copying answers from other students, the university administers two different test forms A and B with distinct item sets but of comparable difficulty. However, the test forms differ in terms of their workload, meaning that test form A contains items with less reading material and test form B contains items with more reading material. Especially for slower reading students it is now of substantial importance whether they are assigned test form A or B, as test form B penalizes them stronger for being slow than test form A. Even though the test forms are equivalent regarding difficulty, they cannot be considered fair due to test speededness. For slow but able students, university admission depends on which test form they are randomly assigned.

These examples illustrate that controlling the speededness of a test is crucial for the construction of fair and valid tests. Indeed, there exist substantial amounts of research on statistical methods for detecting whether tests are speeded and on statistical methods for how psychometric properties of tests can be maintained when a test is speeded (Cintron, 2021). In contrast, the number of publications on how the speededness of an assessment can actually be controlled and under which circumstances speededness may be harmful to the validity and fairness of an assessment is small. One of the only approaches feasible for a wide

variety of assessment contexts is the approach presented by van der Linden (2011a, 2011b). Yet, even this approach is limited as it makes the rather strict and unconventional assumption of equal speed sensitivities (factor loadings) across items and it is therefore unclear, whether the proposed approach is appropriate for practical applications.

Furthermore, fairness issues related to test speededness can also arise in different ways. For instance, in high-stakes testing, equivalent test forms are often used to prevent cheating (Smith et al., 2004). These test forms should not yield different results for the same test-taker. However, equivalence on the test form level may not be sufficient in such scenarios as speededness is often expected to affect specific parts of the test (i.e., the later parts of the test) more so than others (Mollenkopf, 1950). Already Leary and Dorans (1985) noted that in such circumstances certain item orders may be more beneficial for test-takers than others. Yet, research on item ordering and speededness has stagnated in the last centuries and no applicable solution to the problem is available.

To address the aforementioned research gaps, this thesis builds on the work of van der Linden (2011a, 2011b) as well as Leary and Dorans (1985) and makes the following contributions: It (a) illustrates shortcomings of the current state-of-the-art approach by van der Linden (2011a, 2011b) for controlling speededness in practical applications, (b) presents a generalization of the approach suggested by van der Linden (2011a, 2011b) to overcome these shortcomings, (c) shows that differently ordered but otherwise identical test forms can be unfair if speededness is present, (d) presents easy to implement approaches for dealing with the effects of item order due to speededness, and (e) provides statistical software and guidance for implementations of all presented methods for practitioners.

To increase readability, all of the substantial research work in this thesis focuses on specific assessment contexts, mainly on fixed-form linear high-stakes assessments. In such assessment contexts, impact of speededness on fairness and validity is often most pronounced. However, controlling speededness and the effects of speededness is relevant in nearly all assessment contexts. The thesis starts with describing general testing standards, which assessments have to adhere to, and their relationship to speededness. Then, the contexts in and for which assessments are used and why speededness is relevant for almost all of them are discussed. Next, general approaches for (automatically) assembling tests from item pools are described, which will subsequently be applied to incorporating speededness controls. This is followed by a discussion of different response time modeling concepts. Response time modeling not only serves as the basis for controlling speededness when assembling tests but also as the

foundation for understanding the concept of speededness in the first place. Then, the concept of the speed-ability trade-off is discussed and a formal definition of speededness is provided. A detailed overview of the existing research on controlling speededness in assessments is given, including the current shortcomings. From this, the research questions of this thesis are derived. In the main body of this thesis, three substantial research works and two tutorial papers, which make the discussed methods accessible for practitioners and applied researchers, are presented. The thesis provides a summary and critical discussion of the research work and closes with an outlook for future research on the topic.

1.1 Achievement Testing

In this thesis, the terms *assessment* and *test* will be used interchangeably. It is assumed that assessments generally seek to measure a unidimensional latent (i.e., not directly measurable) *construct*. The term *achievement* refers to the fact that in most educational assessments this latent construct refers to a latent *ability*, such as knowledge or skills that can be acquired. This thesis will focus mainly on educational achievement testing, but all presented approaches and discussions are applicable to other psychological testing applications such as intelligence testing or aptitude testing as well. For general overviews on psychological testing and its applications see, for example, Kaplan and Saccuzzo (2017) or Murphy and Davidshofer (2005).

1.1.1 Fairness and Validity

The *Standards for Educational and Psychological Testing* by the American Educational Research Association et al. (2014), from now on simply referred to as the *testing standards*, list three concepts as the foundation of high-quality assessments: (1) validity, (2) reliability/precision and errors of measurement, and (3) fairness. This section will focus on the concepts of fairness and validity, as these are the concepts most directly affected by the speededness of an assessment[1].

The term *validity* refers to the fact that there should be empirical evidence justifying the use and interpretation of test scores, for example for pass/fail decisions or selection for educational programs (American Educational Research Association et al., 2014)[2]. McDonald (2013, p. 133) describes validity as follows: "A test score is valid to the extent that it measures the attribute of the respondents that it is employed to measure, in the popula-

[1]Some researchers argue that strict time limits have a strong impact on the reliability of achievement tests as well (Gernsbacher et al., 2020), yet this discussion is beyond the scope of this thesis.

[2]For a broader, historical overview on the topic of validity, see also Tiffin-Richards and Pant (2017).

tion(s) in which it is used." Messick (1993) argues that the two major threats to validity are *construct-irrelevant variance* and *construct under-representation*. The testing standards describe construct-irrelevant variance as "[...] processes that are extraneous to the test's intended purpose" (American Educational Research Association et al., 2014, p. 12). Examples of such construct-irrelevant variance can be (un)familiarity with specific item types, varying testing conditions due to online-administered tests, or test-takers being affected by strict time limits used for power tests for practical reasons. Construct under-representation refers to a test that "[...] fails to include important dimensions or facets of the construct" (Messick, 1993, p. 9). Examples of tests with construct under-representation could be a language assessment that measures only reading skills and no oral or writing skills, a test for mathematical literacy only focusing on geometry, or a speeded test for reading fluency whose time limit is too lenient, thereby failing to detect differences in reading speed.

The concept of *fairness* relates to the fact that all test-takers should have equal opportunity to display their ability in an assessment (American Educational Research Association et al., 2014). Kingston and Kramer (2013, p. 193) describe fairness as follows: "[Fairness] means there must be no construct irrelevant variance associated with being a member of a definable subgroup." The fairness of an assessment is threatened, if a certain subpopulation is favored in the assessment, for example due to familiarity with the assessment framework and item types, compared to other subpopulations. Speededness can be a substantial threat to the fairness of assessments as (a) certain subpopulations may have higher working speeds than other subpopulations and (b) interchangeably used test forms may contain different amounts of workload.

It is apparent that the concept of fairness is strongly connected to the concept of validity. Indeed, if an assessment is not fair it can never be considered valid. However, fairness is not a sufficient condition for validity, meaning that an assessment can be fair but still not be valid.

1.1.2 Stakes for Test-Takers

To understand the relevance of speededness, it is essential to consider the different contexts in which achievement tests are used. One of the most basic classification criteria is the stakes for test-takers associated with an assessment. The term *low-stakes assessment* refers to assessments which have little to no consequences for test-takers on the individual level (Rios, 2021). Examples include educational large-scale assessments such as the *Programme of In-*

ternational Student Assessment (PISA; OECD, 2016b), the *Programme for the International Assessment of Adult Competencies* (PIAAC; OECD, 2013), the *Progress in International Reading Study* (PIRLS; Martin et al., 2017), or the *Trends in International Mathematics and Science Study* (TIMSS; Martin et al., 2020). These assessments usually focus on results on the group level (e.g., country) and aim at monitoring and comparing educational systems (Kirsch et al., 2013). Other examples of low-stakes assessments are assessments used by teachers to inform and improve their teaching, such as the German *Vergleichsarbeiten* (comparative performance tests, VERA; Ophoff & Cramer, 2022; Pant, 2013).

The term *high-stakes assessment* refers to assessments with substantial consequences for test-takers on the individual level (Rios, 2021). Such consequences are, for instance, admission to educational programs, selection for a job, or certification after concluding an educational program. Examples of such assessments are school or university exams, language proficiency tests like the *Test of English as a Foreign Language* (TOEFL; Educational Testing Service, 2020), or college admission tests like the *Graduate Record Examinations* (GRE; Davey & Lee, 2011) or the *SAT*[3] (formerly known as the Scholastic Aptitude or Scholastic Assessment Test; College Board, 2015).

These different stakes[4] have strong implications for how test-takers approach an assessment and what potential sources of construct-irrelevant variance must be considered by test designers and administrators. Test-takers can be expected to be much more extrinsically motivated in high-stakes assessments than in low-stakes assessments and thus to perform substantially better (Cole & Osterlind, 2008; Rios, 2021; Steedle & Grochowalski, 2017; Wise & DeMars, 2005; Wolf & Smith, 1995). Low test motivation is known to be connected to low test-taker effort, showing in increased item omissions and rapid guessing (Wise & Gao, 2017) or even quitting the assessment (Pools, 2022; Ulitzsch et al., 2020). In low-stakes assessments, low test-taking effort is therefore frequently considered a substantial source of construct-irrelevant variance. In contrast, high-stakes assessments even motivate test-takers to specifically prepare for the assessment (Devine-Eller, 2012). This can, for example, entail studying for an exam or attending a preparation course for a standardized assessment like the TOEFL test. For recent overviews on test preparation for standardized high-stakes

[3]The term SAT is no longer used as an abbreviation but as the full name of the assessment.

[4]The terms high and low-stakes solely refer to the stakes for test-takers. Low-stakes assessments usually also have stakeholders who have great interest in the results of assessments. For instance, teachers who seek to improve their teaching or policy makers who seek to evaluate their policies have an interest in valid and fair measurement. Furthermore, financial benefits can depend on the results of low-stakes assessments, such as teacher salary depending on student performance (*teacher performance pay;* Podgursky & Springer, 2007) or educational institution funds depending on accountability programs (Cole & Osterlind, 2008).

assessments see the works of Kim (2021) and Powers (2017). Unfortunately, cheating is a prevalent threat in high-stakes assessments and can hence be considered a relevant source of construct-irrelevant variance in this context (Bernardi et al., 2008; Chirumamilla et al., 2020). Other sources of construct-irrelevant variance, such as familiarity with item types or speededness due to strict time limits, are however potential sources of construct-irrelevant variance in all assessment contexts.

1.1.3 Standardization

Independent of their stakes, educational achievement tests can also vary greatly in their degree of standardization. For example, college admission tests or language proficiency tests, such as TOEFL, employ professionally developed and maintained item pools as well as standardized procedures for assembling equivalent test forms, which have to conform to numerous specifications (Armstrong et al., 2005; College Board, 2015; Educational Testing Service, 2010). They aim to ensure that assessments are comparable within an assessment cycle but also between assessment cycles (e.g., across years). In contrast, school or university examinations often have a much lower degree of standardization due to limited budgets and resources (Elton, 2004; Frey et al., 2020). In such contexts, the person teaching a class or course will typically also be the person designing and administering the test. Comparability across assessment cycles is achieved based on expert judgments, not via explicit psychometric modeling, and quality control is much less strict. This means, for example, that items are used without any pretesting or external expert validation (Frey et al., 2020). An even lesser degree of standardization can be found in formative assessments whose main goal is to inform teachers and students about learning progress. For this purpose, informal assessments are commonly utilized, such as quizzes or question-and-answer sessions (Dixson & Worrell, 2016; Dolin et al., 2018). From a validity and fairness perspective, however, speededness is relevant for all of these assessments. In this thesis, the term *assessment practitioner* therefore refers to a wide range of persons, from professional test designers and administrators to school teachers or university lecturers.

1.1.4 Measurement Models

A common challenge for achievement assessments is that these assessments are designed to measure constructs which are not directly observable, so called *latent constructs*. An additional challenge is finding measurement models that are flexible enough to accustom

simple assessment contexts (e.g., a single test form used on a specific sample) or complex assessment contexts (e.g., a variety of test forms used on different subpopulations on various occasions). *Item Response Theory* (IRT) models haven been developed for this purpose. IRT models conceptualize responses as outcomes of random processes determined by item and person parameters (van der Linden, 2005). In achievement tests, responses to items are frequently dichotomous, meaning that items can be answered either correctly or incorrectly. One of the most popular IRT models for dichotomous items is the *two-parameter logistic* (2PL) model. In the 2PL model, the probability $P(y_{ik} = 1)$ of a person $i = 1, ..., n$ answering an item $k = 1, ..., j$ correctly, is given by

$$P(y_{ik} = 1|\theta_i, a_k, b'_k) = \frac{\exp(a_k(\theta_i - b'_k))}{1 + \exp(a_k(\theta_i - b'_k))}. \tag{1}$$

θ_i denotes the *ability* parameter associated with a person taking the test, indicating their unidimensional latent ability level. Parameters a_k and b'_k are item parameters, with b'_k being a *difficulty* parameter and a_k being a *discrimination* parameter. It should be noted that the 2PL model can also be written as

$$P(y_{ik} = 1|\theta_i, a_k, b_k) = \frac{\exp(a_k\theta_i - b_k)}{1 + \exp(a_k\theta_i - b_k)}. \tag{2}$$

While the two models are interchangeable, the meanings of the difficulty parameters b'_k and b_k differ between Equation 2 and 1, as $b_k = a_k b'_k$ (de Ayala, 2022; Fox, 2019). Throughout this thesis, if not stated explicitly otherwise, the term 2PL model will refer to Equation 2. A special case of the 2PL model is the *one-parameter logistic* (1PL) model, which is conceptually equivalent to the Rasch model (Rasch, 1960). In this model, all discrimination parameters a_k are fixed to 1. Thus, the model assumes that all items discriminate equally between test-takers with different ability levels. In the 1PL model, the probability $P(y_{ik} = 1)$ of a person i answering an item k correctly, is given by

$$P(y_{ik} = 1|\theta_i, b_k) = \frac{\exp(\theta_i - b_k)}{1 + \exp(\theta_i - b_k)}. \tag{3}$$

Regardless of the specific measurement model, the parameter of interest in achievement testing is usually the person ability parameter θ_i. See, for example, the work of de Ayala (2022) for an extensive and general overview of the IRT literature and research. For a more historical perspective, including comparisons of IRT to *classical test theory*, see also van der Linden (2005). It should be noted that numerous extensions of the presented models, such

as the three-parameter logistic model (for modeling a guessing probability) or the generalized partial credit model (for modeling partial credit on items) exist (de Ayala, 2022). All approaches for controlling speededness presented later in this thesis are, however, applicable independent of the specific IRT models used.

An important, universal feature of IRT is that the information provided by an item for the ability estimation of a test-taker can be quantified. The *item information function* (IIF) is derived from the more general *Fisher information* and in the 2PL framework is defined for an item k and ability level θ as (e.g., de Ayala, 2022; DeMars, 2010):

$$I_k(\theta) = a_k^2 P_k(\theta)(1 - P_k(\theta)). \tag{4}$$

$P_k(\theta)$ refers to the probability of a correct response for item k given θ under the 2PL model. Equation 4 illustrates that the IIF is determined by two aspects: (a) the discriminative power of the item, expressed by its discrimination parameter a_k, and (b) the probability to solve an item correctly $P_k(\theta)$ multiplied by its inverse probability. The latter term is largest if the probability of a correct answer is $P_k(\theta) = 0.5$, so if $b_k = \theta_i$. On an aggregate level, the *test information function* (TIF) is the sum of all IIFs of the items $k = 1, ..., j$ in a test:

$$I(\theta) = \sum_{k=1}^{j} I_k(\theta). \tag{5}$$

The TIF is a frequently used statistic of interest when designing a test. For instance, if a test is used as a licensing exam, test forms should be highly informative around the cut score used for pass/fail decisions. Alternative statistics, such as the *test characteristic curve*, exist (e.g., Debeer et al., 2017). However, all later presented approaches for controlling speededness are independent of any other design features or requirements of a test. As controlling the speededness of a test is part of the assembly of the test, the following section provides a short overview on the general topic of test assembly.

1.2 Test Assembly

The testing standards describe the design and development of a test consisting of four phases (American Educational Research Association et al., 2014):

1. Development and evaluation of the test specifications

2. Development, tryout, and evaluation of items

3. Assembly and evaluation of new test forms

4. Development, procedures, and materials for administration and scoring.

This section focuses on step (3), and more specifically on the assembly of new test forms, as in this step the workload of a test is determined and important decisions regarding its speededness are made. The term *test assembly* describes the process of assembling items from an item pool to one or multiple test forms. Usually, test assembly is a complex process as there are specific requirements test forms have to adhere to (see step (1), the development and evaluation of the test specifications). These *test specifications* usually cover a broad range of requirements and include content, format, and psychometric specifications, such as representation of subdomains within a test form, desired test length of a test form, or average difficulty of a test form (American Educational Research Association et al., 2014). Step (2) above illustrates that items are ideally tested in form of *pre-tests* (also termed *pilot tests* or *pilot studies*) before they are used in the actual test administration, also termed the *operational test*.

1.2.1 Parallel Test Forms

The complexity of test assembly procedures varies greatly between assessment contexts. For instance, in university examinations, it is common that test designers develop just enough items for a single test form. In extreme cases, there is basically no test assembly procedure required, as the single assembled test form simply equals the complete item pool. In educational large-scale assessments, multiple test forms are typically assembled within multiple matrix booklet designs to ensure broad content coverage (Gonzalez & Rutkowski, 2010). In these cases, test forms differ in the tested (sub-)domains but still should be equivalent in many aspects, such as test length or average difficulty, so that test-takers have a comparable testing experience regardless of the specific test form they are randomly assigned to. In standardized high-stakes assessments, multiple test forms are mainly used to increase test security. If identical test forms were used for all test-takers, test-takers could copy answers from each other or collaborate during a test administration (Bernardi et al., 2008; Chirumamilla et al., 2020; Smith et al., 2004). Furthermore, such assessments are usually offered repeatedly. Different test forms make it harder for test-takers to share relevant test content with future test-takers and are therefore crucial to the integrity (and fairness) of such assessments (Kippel, 1985; van der Linden, 2022). Such test forms consisting of distinct item sets are referred to as *nonoverlapping* test forms (Belov & Chen, 2014). However, in high-stakes assessments, test

forms have to be exchangeable, as tests should be comparable on the individual level across test forms and testing occasions.

Equivalent test forms which can be used interchangeably are also referred to as *parallel*. As illustrated above, the degree of required parallelism can vary from application to application. Historically, parallel test forms were defined by equal true scores and error variances across test forms. Lord (1980) refers to such test forms as *strictly parallel*. In the context of IRT, Samejima (1977) proposed the differentiation of *strongly parallel* test forms with parallel item pairs in contrast to *weakly parallel* test forms, which are parallel regarding their TIF. Samejima (1977) argues that weakly parallel test forms suffice for practical applications. The testing standards (American Educational Research Association et al., 2014, p. 35) describe parallel test forms as "[...] designed to have the same general distribution of content and item formats, the same administrative procedures and at least approximately the same score means and standard deviations in some specified population [...]." This definition illustrates that a variety of aspects have to be considered for evaluating whether test forms are parallel. Furthermore, different assessment applications can require different degrees of parallelism. For instance, in standardized high-stakes assessments, test forms are usually assembled with great care from sufficiently large item pools (e.g., College Board, 2015) to guarantee fair and truly exchangeable test forms.

Several different practical approaches to test assembly exist. Test form assembly can be performed in one step, with test forms being directly assembled from an item pool. Alternatively, the test assembly procedure can consist of multiple steps. For instance, in a first step, *blocks* are assembled from an item pool and, in a second step, test forms are assembled from these blocks (van der Linden, 2005). Such a two-step test assembly procedure is frequently applied in large-scale assessments (e.g., Kuhn & Kiefer, 2015; OECD, 2016b). For simplicity and readability reasons, this thesis focuses on test forms being assembled from items only. However, all reasoning and implications apply to multi-step procedures as well and respective extensions are rather straightforward.

1.2.2 Degree of Adaptivity

The traditional approach in achievement testing is to use *pre-assembled, fixed-form linear tests* (Luecht & Sireci, 2011; van der Linden & Glas, 2010a). This means that test forms are assembled a priori and then administered to test-takers. During administration, every test form is fixed. This assessment mode is rooted in the era of paper-based assessment, where

any form of on-the-fly modification of a test was at least cumbersome if not impossible. The advancement of computer-based assessment has also led to an advancement of more adaptive testing modes. In *computerized adaptive testing* (CAT), a provisional ability estimation is performed after every item and the next item is selected based on the current ability estimate, thereby increasing measurement precision (van der Linden & Glas, 2010a). This procedure makes use of the fact that items are differentially informative given a specific (provisional) ability level. However, in CAT, it can be challenging to control item exposure (i.e., how often is a single item used across test-takers) and further test specifications such as content domains or item type distributions across test-takers. A popular alternative to CAT is therefore *multi-stage testing* (MST), in which *modules* consisting of sets of items, are pre-assembled and administered depending on the performance of test-takers on prior modules (Yan et al., 2016; Zenisky et al., 2010). All mentioned assessment modes (fixed-form linear tests, CAT, MST) are used in both high- and low-stakes contexts. For instance, in high-stakes testing, the SAT and TOEFL use fixed-form linear tests (College Board, 2015; Educational Testing Service, 2020), the revised GRE uses MST (Davey & Lee, 2011) but previously used CAT (Bridgeman & Cline, 2004). In low-stakes testing, PISA and PIAAC have been using MST designs in the recent past (OECD, 2013, 2019b), while TIMSS and PIRLS have been using fixed-form linear tests (Martin et al., 2017, 2020).

Regardless of the degree of adaptivity of an assessment, the final test form(s) should usually conform to test specifications. Therefore, test assembly is a relevant challenge in all three assessment modes: In MST, modules consisting of multiple items are usually pre-assembled while having to meet specific requirements (Yan et al., 2016). Often, modules of the same stage are targeted at different ability levels and therefore should have different test information functions but should otherwise be parallel (regarding, e.g., length, content coverage, and item formats). In the CAT framework, meeting complex test specifications is more challenging. An elegant approach for dealing with complex test specifications in CAT is the *shadow-test approach* (STA). In the shadow-test framework, a full test assembly method is performed after each administered item which results in a test form satisfying all relevant test constraints. From this full test form the most informative item is selected. Through this iterative procedure, it can be guaranteed that the final test form fulfills all required test specifications as well (van der Linden & Reese, 1998).

The substantial research works of this thesis will focus on fixed-form linear tests, as the extension of the respective test assembly procedures to MST or CAT using the STA framework

is rather straightforward. The relevance of controlling speededness for MST and CAT will be discussed in later parts of this thesis (see Chapter 7).

1.2.3 Conventional/Manual Test Assembly

Originally, test forms were assembled manually using trial and error approaches, as for example described in van der Linden (2005). This means that test administrators created an overview of the item pool, for example using a spreadsheet, and manually assigned items to different test forms. By trying out different solutions, test administrators ruled out options and searched for the optimal (or a practically sufficient) solution. Such an approach can be feasible for small test assembly problems (e.g., assembling a test form of 20 items from an item pool of 60 items) but will quickly reach its limitations for more complex test assembly problems (e.g., creating multiple parallel test forms from a large item pool). Although large-scale testing programs have used manual test assembly procedures in the past for more complex test assembly problems as well, as no other approaches were available or out of habit, such procedures are typically very time-consuming and are likely to yield suboptimal solutions (Armstrong et al., 2005).

1.2.4 Automated Test Assembly

Automated test assembly (ATA) refers to approaches which use computer algorithms to solve test assembly problems. For example, Theunissen (1985) was the first to suggest using *linear programming* for this purpose. Since then, automated test assembly approaches have evolved steadily. The work of van der Linden (2005) on how to use *mixed integer linear programming* (MILP) for ATA provides researchers with extended theoretical guidance on these methods.

The general idea is that test specifications can be translated into mathematical constraints and an objective function in the form of equations and inequalities. The term *constraints* refers to an implementation using explicit cut-offs or equalities. The term *objective function* refers to the fact that for this specification no strict cut-off is set, but a minimization or maximization criterion is defined. In the following, frequently used constraints and objective functions are discussed. The respective *decision variables* used in these (in-)equations are usually binary (0 or 1), indicating whether an item is included in a test form. Decision variables can be denoted as x_{kf} for item k and test form f.

$$x_{kf} = \begin{cases} 1 & \text{if item } k \text{ is selected in test form } f \\ 0 & \text{if item } k \text{ is not selected in test form } f. \end{cases} \tag{6}$$

In his work, van der Linden (2005) categorizes constraints into quantitative, categorical, and logical constraints. Examples for quantitative constraints include specifying the estimated average test time for the test, the total number of items in the test or a minimum test information function the test is supposed to have. Specifying the minimum TIF (T_θ) as the sum of the item information functions $I_k(\theta)$ at ability level θ can be formulated as

$$\sum_{k=1}^{j} I_k(\theta) x_{kf} >= T_\theta. \tag{7}$$

Examples for categorical constraints include the distribution of items types or items belonging to specific subdomains across test forms. For instance, assuring that the number of items of the subset V_c (e.g., multiple-choice items) is below the maximum n_c^{max} and above the minimum n_c^{min} in test form f can be written as:

$$\sum_{k \in V_c} x_{kf} <= n_c^{max} \tag{8}$$

$$\sum_{k \in V_c} x_{kf} >= n_c^{min} \tag{9}$$

It should be noted that in practice, categorical constraints can easily be transformed into quantitative constraints if the grouping factor is translated into a set of dummy coded variables and these 0-1 variables are then used in quantitative constraints. For instance, if d_k is a dichotomous variable indicating whether item k is a multiple-choice item ($d_k = 1$) or not ($d_k = 0$), Equations 8 and 9 can be written as:

$$\sum_{k=1}^{j} d_k x_{kf} <= n_c^{max} \tag{10}$$

$$\sum_{k=1}^{j} d_k x_{kf} >= n_c^{min} \tag{11}$$

Logical constraints refer to the conditional selection of items. Examples of these are cases of exclusion (e.g., item 1 and item 2 cannot be in the same test form f) and inclusions (e.g., if item 1 is included in a test form, item 2 has to be also included). Exclusions, also called

enemy items, can, for instance, be due to items containing the solution to other items and can be written as

$$x_{1f} + x_{2f} <= 1. \tag{12}$$

Equation 12 enforces that either none of the two items ($x_{1f} = 0$ and $x_{2f} = 0$) or maximally one of the items ($x_{1f} = 1$ or $x_{2f} = 1$) is included in the test form f, but not both of them. In contrast, inclusions, also called *friend items*, can occur due to a shared stimulus and can be written as

$$x_{1f} - x_{2f} = 0. \tag{13}$$

Equation 13 is only satisfied if both items are in the test form ($x_{1f} = 1$ and $x_{2f} = 1$) or none of the items is in the test form ($x_{1f} = 0$ and $x_{2f} = 0$). Note that logical constrains can be reformulated into quantitative constraints comparable to categorical constraints via the use of dichotomous indicator variables.

Besides these quantitative, categorical, and logical constraints, usually a single *objective function* is formulated in ATA. Examples of common objective functions include maximization of the TIF, minimization of test length or testing time, or minimization of the difference of the TIF between multiple test forms. Formulating the maximization of the TIF at ability level θ, for example, would be expressed as

$$\text{maximize} \sum_{k=1}^{j} I_k(\theta) x_{kf}. \tag{14}$$

A common challenge in practical applications is the decision, which test specification should be implemented as the objective function and which test specifications should be implemented as constraints. Almost all test specifications can be approximated both by fixed constraints or an objective function. For instance, the TIF can be either maximized or constrained directly to a high level. Unfortunately, there is no easy answer to this dilemma. Some practitioners may prefer to implement more than one test specification as an objective function, because frequently multiple test specifications are not strict (in-)equalities but "as good as possible" requirements. However, this would lead to a multi-objective optimization problem. Unfortunately, multi-objective optimization problems are notoriously difficult to implement and often require conceptual compromises (van der Linden, 2005; Veldkamp, 1999).

In general, as the computational power of computers has long surpassed the computational power of the human brain with regard to combinatorial optimization, the convenience and elegance of ATA procedures versus manual test assembly is apparent. However, there are still limitations to the accessibility and usability of ATA approaches. While the works of van der Linden (2005) and Kuhn and Kiefer (2015) provide in-depth theoretical guidance on ATA, resources on practical implementations such as software tutorials are still scarce. An overview of programs that are in principle suitable for ATA implementations can be found in Donoghue (2015). It seems that the only existing tutorial on software implementations of ATA can be found in Diao and van der Linden (2011). However, Diao and van der Linden (2011) illustrate how ATA approaches can be implemented via a lpSolver-API R package, which is not specifically designed for ATA and whose usability is arguably sub-optimal. It can be argued that the lack of designated ATA-software and the lack of tutorials for ATA implementation is a substantial burden for assessment practitioners who want to use ATA for their test assembly.

Besides ATA, the second prerequisite for controlling the speededness of tests is response time modeling. Therefore, the next section provides a short overview on response time modeling in general.

1.3 Response Time Modeling

The concept of measuring response times on cognitive tests and thereby modeling an underlying speed factor dates back to the beginnings of experimental psychology (e.g., Baxter, 1941). For detailed, historical reviews, see also Schnipke and Scrams (2002), Luce (1986), Lee and Chen (2011), van der Linden (2009a), as well as Kyllonen and Zu (2016). In the recent past, response time modeling has gained a lot of traction in the psychometric literature, as response times are more easily and unobtrusively available in computer-based assessments than they were in paper-based assessments. *Computer-based assessments* (CBA), sometimes also termed *technology-based assessments* (TBA), refer to assessments which are administered on computers, such as laptops, desktop computers, or tablets (Csapó et al., 2012; Kröhne & Goldhammer, 2018). Computer-based assessments have a variety of advantages compared to paper-based assessments, for instance the possibility to measure new domains or subdomains such as information and communication technology literacy or reading in technology-rich environments (Csapó et al., 2012). Another great advantage is that process data, such as response times, can be automatically tracked in the background via the assessment system

without additional proctoring or asking test-takers to track their own response times (Kröhne & Goldhammer, 2018).

In the literature, competing definitions of response times exist (van Rijn & Sinharay, 2023). For instance, assessments such as PISA or PIAAC define response times as time spent on an item (e.g., Goldhammer et al., 2020; OECD, 2022); others define response times as the time a test-taker spends on an item until the final response is made (e.g., Kröhne & Goldhammer, 2018; Li et al., 2017). Frequently, psychometric literature on response time modeling does not address this issue at all. For reasons of simplicity and because in practice, researchers often depend on what kind of data is made available in public use files, throughout this thesis it is assumed that response times are available from assessments and represent exactly the time a test-taker has taken to work on an item. A further discussion on this topic is provided in Chapter 7.

Through the advancement of CBA in the last decades, response times are now more easily available, fostering not only research utilizing response times but also their practical use in data analyses or assessment design. For instance, response times are used as additional information to score omitted responses (Weeks et al., 2016) or to differentiate between test-takers who ran out of time and test-takers who quit the assessment ahead of time (Pools, 2022; Ulitzsch et al., 2020). Another important application is the use of response times from pre-testing for controlling the speededness of test forms during test assembly (van der Linden, 2005, 2011a, 2011b). A recent review of the psychometric response time modeling literature can be found in De Boeck and Jeon (2019). The authors divide response time models into four categories: (a) *pure response time models* with response times being the sole end variable, (b) *joint models* with response times and responses as end variables, (c) *dependency models*, which are joint models but include dependencies beyond latent relationships, and (d) *response times as covariate models*. In this thesis, the focus will be on (a) pure response time models and (b) joint response and response time models, as these model types are utilized by van der Linden (2011a, 2011b) for controlling the speededness of test forms. Models belonging to the family of (c) dependency models will, however, be covered in Chapter 5 and their relevance for controlling speededness will be discussed in Chapter 7.

1.3.1 Lognormal Response Time Model

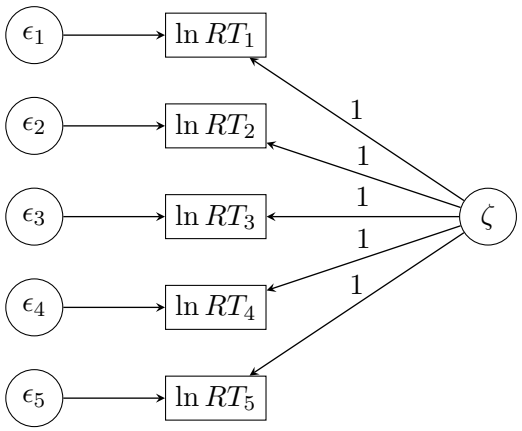
The lognormal response time model by van der Linden (2006) is one of the most popular response time models in the literature (e.g., van Rijn & Sinharay, 2023). It is a pure response

time model as categorized by De Boeck and Jeon (2019). In general, the lognormal distribution is frequently used in response time modeling as response times are truncated at 0 (i.e., negative response times are impossible). The lognormal model by van der Linden (2006) assumes that response times are lognormally distributed and that the lognormal response times $\ln RT_{ik}$ for persons denoted as $i = 1, ..., n$ and items denoted as $k = 1, ..., j$ follow the measurement model

$$\ln RT_{ik} = \lambda_k - \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{15}$$

In the model, λ_k represents the *time intensity* or workload of an item. ζ_i represents the person *speed* parameter, indicating how fast a person works on the assessment. $\sigma^2_{\epsilon_k}$ is an item-specific residual variance. Throughout this thesis, this model is termed the *two-parameter lognormal* (2PLN) model, as it contains two item-specific parameters (time intensity and residual variance). It should be noted that the 2PLN model is a purely descriptive model, which implies that it makes no assumptions regarding the processes leading to response times. In contrast, other response time models, such as *race models* or *diffusion models*, make stronger assumptions about the specific underlying processes leading to response times (De Boeck & Jeon, 2019).

Figure 1: *Illustration of the Two-Parameter Lognormal Model by van der Linden (2006).*



Readers familiar with *confirmatory factor analysis* (CFA) or *structural equation modeling* (SEM) will note that the 2PLN model equals a linear one-factor model for the log-transformed response times with freely estimated intercepts and residual variances, but omitted factor loadings (see also, Molenaar, Tuerlinckx, & van der Maas, 2015b; van Rijn & Sinharay, 2023). An exemplary illustration of the 2PLN model as a CFA model with five items can be

seen in Figure 1. By relating the original lognormal model to confirmatory factor analysis, it becomes apparent that the 2PLN model makes a rather unconventional assumption: In omitting the item-specific factor loadings, the model implicitly assumes that all items have equal factor loadings (fixed to 1). For instance, Ranger and Ortner (2012a, p. 133) write: "[...] the model of van der Linden (2006) contains a restriction of the different [factor loading] parameters to the same value. [...] such constraints are unusual in factor analysis [...]" While such an assumption is sometimes used in CFA, for instance under the term *τ-equivalence* (Brown, 2006), this assumption is expected to be empirically tested and justified, or should at least be stated explicitly. Therefore, a logical generalization of the 2PLN model is adding freely estimated factor loadings. The first researchers to propose such a generalization were Fox et al. (2007), Klein Entink, Fox, and van der Linden (2009) and Ranger and Ortner (2012a). The resulting measurement model can be written as

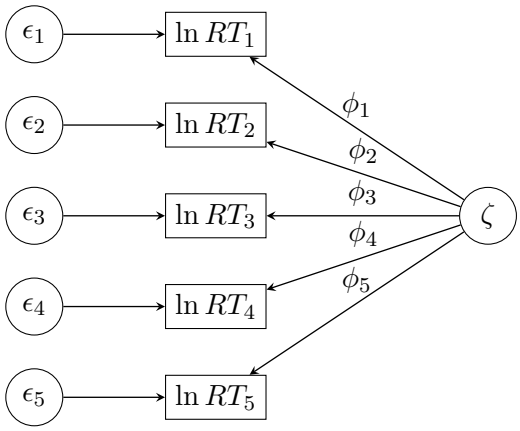
$$\ln RT_{ik} = \lambda_k - \phi_k \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{16}$$

Thereby, ϕ_k represents the *speed sensitivity* of an item, indicating how sensitive an item is to differences in speed between test-takers. Throughout this thesis, this model is termed the *three-parameter lognormal* (3PLN) model, as it contains three item parameters (time intensity, speed sensitivity, residual variance). An exemplary illustration of the 3PLN model with five items can be seen in Figure 2. In-depth explanations and comparisons of the 2PLN and the 3PLN model can be found in Chapter 2. It should be noted that even though the 2PLN model makes a rather unconventional, often untested assumption, the model is used in many contexts such as the joint modeling of responses and response times (van der Linden, 2007), the modeling of missing responses (Pohl et al., 2019; Ulitzsch et al., 2019b, 2020), the identification of aberrant response behavior such as item pre-knowledge (Kasli et al., 2022; Man & Harring, 2020; Man et al., 2018; van der Linden & Guo, 2008), or the control of speededness during test assembly of fixed-form tests (van der Linden, 2011a, 2011b) as well as the administration of CAT (van der Linden & Xiong, 2013). Yet, this assumption is frequently neither explicitly stated or questioned, nor empirically tested.

1.3.2 Hierarchical Framework

For the specific purpose of controlling speededness, it can be argued that pure response time models are sufficient (van der Linden, 2011a, 2011b). However, to foster a better conceptual understanding of the speed-ability relationship and because the interplay of speed and ability

Figure 2: *Illustration of the Three-Parameter Lognormal Model.*



is often relevant in practical applications (e.g., the time intensity of items correlates with the difficulty of items), joint models are required which model responses and response times simultaneously. For this specific purpose, and because a joint estimation of both constructs can aid model and parameter estimation (van der Linden et al., 2010), van der Linden (2007) proposed a joint, hierarchical framework for ability and speed. In this thesis, joint models will be mainly used to generate plausible simulation conditions for the evaluation of approaches for controlling speededness. Central assumptions of the hierarchical framework include multivariate normal joint item and person parameter distributions and local stochastic independence of manifest responses and response times given this latent structure. Beyond that, various measurement models for ability and speed can be used within the hierarchical framework. Common choices for the respective measurement models for responses and response times are the 2PL and the 3PLN model (e.g., Debelak et al., 2014; Goldhammer & Klein Entink, 2011; Scherer et al., 2015). Mathematically, the resulting joint person parameter distribution of the hierarchical framework with the 2PL and the 3PLN model can be written as

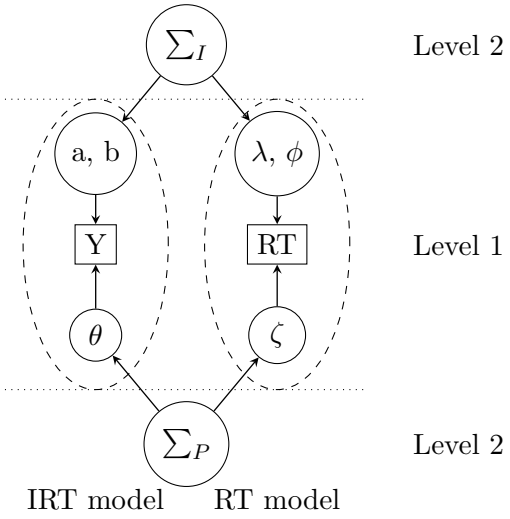
$$(\theta_i, \zeta_i) \sim \mathcal{N}(\boldsymbol{\mu_P}, \boldsymbol{\Sigma_P}). \tag{17}$$

$\boldsymbol{\Sigma_P}$ denotes the covariance structure of the person parameters. Note that the joint item parameter distribution depends on the parameterization of the respective measurement models. Again, for the 2PL and 3PLN model case it can be written as

$$(a_k, b_k, \lambda_k, \phi_k) \sim \mathcal{N}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}). \tag{18}$$

$\boldsymbol{\Sigma_I}$ denotes the covariance structure of the item parameters. The latent structure of the hierarchical framework using these two specific measurement models can be seen in Figure 3.

Figure 3: *Structure of the Hierarchical Response Time Model Using the 2PL and 3PLN Measurement Models, Extended from Klein Entink, Kuhn, et al., 2009.*



The introduced hierarchical framework is indeed very flexible. The model can be extended or modified in various directions, for example by using *cognitive diagnostic modeling*[5] (CDM) instead of IRT models (Huang, 2019), by using mixture components (e.g., Ulitzsch et al., 2022; C. Wang et al., 2018) or explicitly modeling residual relationships (e.g., Bolsinova, de Boeck, & Tijmstra, 2017). A common alternative that has been frequently used in the literature is the substitution of the 3PLN with the 2PLN model as the response time measurement model. The resulting common item parameter distribution with the 2PL and 2PLN model can be denoted as

$$(a_k, b_k, \lambda_k) \sim \mathcal{N}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}). \tag{19}$$

Multiple studies have compared the fit of hierarchical frameworks using the 2PLN model with hierarchical frameworks using the 3PLN model on empirical data, indicating superior fit of the latter (Debelak et al., 2014; Goldhammer & Klein Entink, 2011; Scherer et al., 2015). Further model comparisons are performed and reported in Chapters 2 and 3. Unfortunately, while pure response time models such as the 2PLN or 3PLN model are rather straightforward to implement in standard statistical software, the implementation of the hierarchical framework and further extensions of it are not.

[5]For an introduction to CDMs, see, for instance, de la Torre (2011).

1.4 Speed-Ability Trade-Off

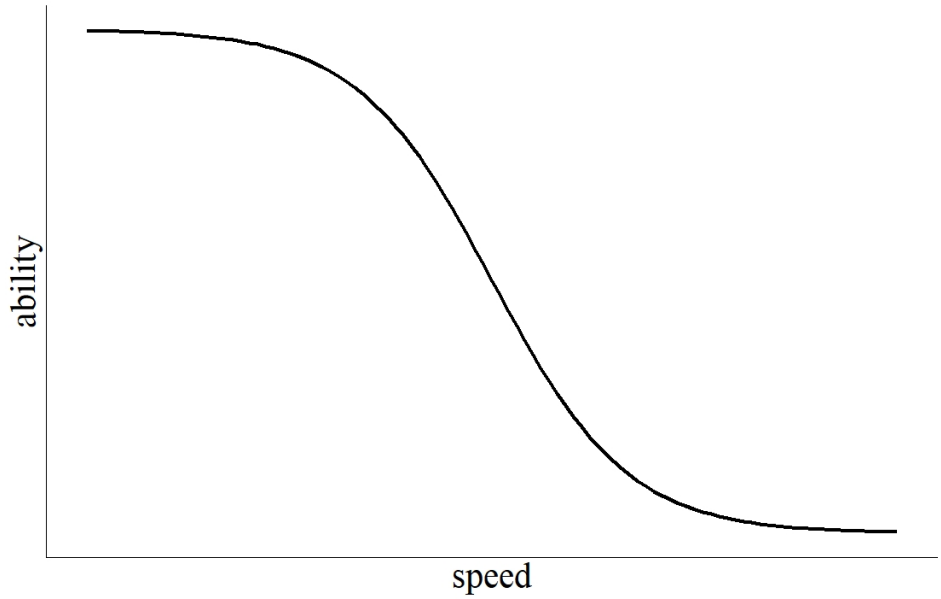
One of the main reasons to model ability and speed simultaneously is to investigate their relationship. The *speed-ability trade-off* (SAbT) is a concept dating back to the earliest days of psychological science (e.g., Henmon, 1911; Spearman, 1927). Back then, experimental researchers investigated the assumption that test-takers work with increased accuracy if they invest more time into a task. If they invest less time, their accuracy displayed on the task decreases. As already Spearman (1927, p. 250) noted: "[...] any increase in speed at a mental operation tends towards a decrease in its goodness, whilst inversely a greater goodness can always be attained by some sacrifice of speed." For an extensive historical overview, see the work of Heitz (2014). Indeed, the notion of working with greater accuracy when given more time seems almost trivial. It is not only a well-researched phenomenon among humans (e.g., Goldhammer & Kröhne, 2014; Wickelgren, 1977) but even among animals, such as bees (Chittka et al., 2003), ants (Franks et al., 2003), monkeys (Heitz & Schall, 2012), or rats (Kaneko et al., 2006).

Goldhammer (2015) provides an in-depth overview on the topic focusing on the area of psychological and educational testing. He differentiates between the speed-accuracy trade-off investigated in experimental research, which focuses on manifest accuracy, and the speed-ability trade-off, with ability being a latent construct, as is usually the case in achievement testing. Furthermore, Tijmstra and Bolsinova (2021) argue that the described conceptualizations of speed always refer to a *response* or *working speed*, not to an inherent cognitive speed. Speed-accuracy and speed-ability trade-offs are typically assumed to follow a curvilinear relationship (H. Chen et al., 2018; Goldhammer, 2015). For the purpose of this thesis, the specific shape of the speed-ability trade-off is irrelevant. However, it is assumed that the speed-ability relationship is monotonically decreasing (i.e., with an increase in speed, ability always decreases or remains constant but never increases) and from a certain point on, additional time does not lead to an increase in accuracy. For an exemplary depiction of a curvilinear speed-ability trade-off as discussed in Goldhammer (2015) see Figure 4.

The described SAbT is a within-person phenomenon (Goldhammer, 2015). This means that, for example, if a person is able to freely choose their working speed for a task or assessment, this person effectively decides where on their individual SAbT curve they want to be working on (ideally guided by the test settings, such as time limits and instructions)[6]. It can be assumed that persons may vary in different aspects of their SAbT, for instance their

[6]However, person characteristics or situational aspects, such as regulatory focus may play a role as well (Förster et al., 2003).

Figure 4: *Conceptual Illustration of the Speed-Ability Trade-Off as Described in Goldhammer (2015).*



slope (i.e., how strongly does the ability level drop off for a person if their working speed level is increased), their intercept (i.e., what is a person's maximum ability/ability ceiling if they work sufficiently slowly), or the speed level at which ability converges asymptotically against 0 (i.e., chance level) (Bolsinova & Tijmstra, 2015; Goldhammer et al., 2017; Tijmstra & Bolsinova, 2018). Regardless, test designers must be aware that different test-takers may simply choose different locations on their individual SAbT curves, independent of how they differ in their ability and speed-ability trade-off. In practice, two persons with very different SAbTs can still work with identical speed and ability, if their SAbTs cross at that specific point. On the other hand, two persons with identical SAbTs can choose very different speed levels to work within an assessment. This also explains why it can be assumed that, conceptually, the within-person SAbT is supposed to be stable (slower means more accurate), yet the between-person relationship between speed and ability empirically shows great variation between assessments, with correlations being sometimes positive, negative, or zero (Tijmstra & Bolsinova, 2021).

In the field of experimental psychology, various techniques have been used to experimentally vary the speed-accuracy trade-off within persons (Heitz, 2014; Wickelgren, 1977). These include verbal instructions (instructing study participants to work fast and/or accurate, e.g., Howell & Kreidler, 1963), payoffs (rewarding/penalizing study participants for fast and correct/incorrect responses, e.g., Swensson & Edwards, 1971), the response-signal paradigm

(study participants have to respond immediately after a signal is given, e.g., Goldhammer & Kröhne, 2014) and deadlines (setting time limits on tasks, e.g., Pachella & Pew, 1968). In the context of educational and psychological testing, the common practice of setting fixed time limits on tests roughly equals the deadlines-method from experimental psychology[7].

1.5 Speededness

Although everyone who has taken a test with a time limit has probably an intuitive understanding of the concept of speededness, a formal definition is required. Historically, tests have been categorized as pure speed or as pure power tests. Following Spearman (1927, p. 252), to measure ability, tests must be used "in which ample time is allowed, so that speed has little or no scope" and to measure speed, tests must be used "in which the time allowed is too brief for any but the fastest subjects to reach the end, so that here speed becomes of vital importance." According to Gulliksen (1950, p. 230), pure speed tests are tests "[...] composed of items so easy that the subjects never give the wrong answer to any of them. The answers are correct as far as the subject has gone in the test. However, the test contains so many items that no one finishes it in the time allowed." In contrast, "[...] in a pure power test all the items are attempted so that the score on the test depends entirely upon the number of items that are answered, and answered incorrectly."

However, in reality, almost all standardized assessments use a fixed time limit and are a mixture of speed and power tests (Goldhammer, 2015; Rindler, 1979). This was already noted by Gulliksen (1950, p. 230): "At present most tests are a composite in unknown proportions of speed and power, which makes the development of appropriate theorems in test theory more difficult than for the pure type tests." While some researchers argue that power tests should avoid time limits at all costs (Gernsbacher et al., 2020), most tests underlie practical considerations, such as limited facility availability, test administrator availability, or cost constraints (Sireci & Botha, 2020), requiring test administrators to use time limits. However, this means that the measured construct can no longer be referred to as a pure, "maximum" ability as it represents a compound measure of ability and speed (e.g., Wilhelm & Schulze, 2002). This is an issue, as research indicates that speed and ability are not the same construct (e.g., Partchev et al., 2011). For an extensive, historical overview on the matter (focusing mainly on cognitive abilities), see, for example, Carroll (1993). However, assessment and test designers differ in the extent to which they see speed as a nuisance factor or as a substantial

[7]Which does not mean that test-takers adhere to this manipulation of speed but may still work slower or faster than the time limit requires them to work.

and desirable part of the measured construct. The following section discusses scenarios in which speed is seen as a nuisance factor and in which speed is seen as a substantial part of the construct and argues that the control of speededness is vital for both scenarios.

1.5.1 Speed as a Nuisance Factor

Frequently, the ability dimension, which is to be measured, is seen as a "pure" ability dimension, free of the influence of speed. Such a view is closely related to the concept of pure power tests referred to by Spearman (1927) and Gulliksen (1950). Examples of assessments that conceptualize speed as a nuisance factor could be classroom assessments that are supposed to inform teachers whether students have understood the material from the last few lessons. Similarly, a university admission test for a psychology master program could seek to determine whether potential students have basic statistical knowledge, irrespective of whether they can reproduce it in a fast or slow manner.

If in such instances speededness is involuntarily introduced through practical time limits, and test-takers do not have sufficient time to answer all items to their fullest ability, they are faced with a decision: Either they (a) work slowly and carefully and not reach the end of the test, (b) work slowly but omit items to reach the end of the test in time, or (c) work faster but less carefully. Of course, combinations of these "pure" decisions are possible and there is research on how these decisions may be made in real tests, such as when the answer mode switches from solution to rapid guessing behavior (Schnipke & Scrams, 1997). It is apparent that all these aforementioned options reduce the number of correct answers of a test-taker and are counterproductive to a valid and fair measurement of "pure ability". In fact, the influence of speed can be seen as construct-irrelevant variance or multidimensionality (de Ayala, 2022; Y. Lu & Sireci, 2007). If a test aims to measure whether test-takers can perform a task independent of time constraints, speededness will threaten the overall validity of the test. If test-takers work at different speed levels, the test may have fairness issues, as some test-takers have enough time to finish the task and others do not (Kane, 2020). Kane (2020) refers to the difference between hypothetical unlimited-time scores and the actual, time-limited scores as *time-limit errors*.

A specific concern for test administrators is the fact that the speededness of a test can affect whole subpopulations differently. For instance, language proficiency influences reading speed, thereby leading to higher risks of speededness for non-native speakers on a test for mathematical literacy (Ercikan et al., 2020). Research on gender differences, for example, has

shown that time limits affect male and female test-takers differently (Steinmayr & Spinath, 2019; Stoevenbelt et al., 2022; Voyer, 2011). Similar evidence exists for different cultures or ethnic backgrounds (Evans & Reilly, 1972; Knapp, 1960; Lawrence, 1993; Sehmitt & Dorans, 1990). Such findings are supported by research indicating that pacing behavior may vary across countries or ethnical backgrounds (Lee & Haberman, 2016; Llabre & Froman, 1987). It is therefore evident that, if speed is seen as a nuisance factor, test designers must ensure that assessments provide all test-takers with sufficient time to be not speeded, both for fairness and validity reasons.

1.5.2 Speed as a Substantial Part of the Construct

In contrast, some achievement constructs are a mixture of both speed and power (Kane, 2020; Mollenkopf, 1960). They are neither only trivial tasks which should be completed at maximum speed (i.e., a pure speed test) nor difficult tasks which can be completed at any amount of time (i.e., a pure power test), but difficult tasks that should be completed in limited amounts of time. Kane (2020) calls these assessments *time-sensitive performance tests*. Tijmstra and Bolsinova (2018) refer to ability measured under deliberate time pressure as a *target ability* displayed at a specific *target speed* level.

Examples of such assessments are situational performance tests, such as *in-basket tests*, in which prioritization of work is one of the key skills being measured (Mollenkopf, 1960). Other examples could be tests of reading literacy, in which time efficient processing of reading material is relevant. In such scenarios it is important that the degree of speededness in the assessment is deliberately and carefully chosen, as any deviation will change the composition of the measured construct. Indeed, if a test is accidentally unspeeded while speed should be a part of the construct being measured, test designers are constructing a test with construct under-representation. Therefore, it can be argued that the control of the speededness of a test is vital regardless of whether speed is seen as a nuisance parameter or as a substantial part of the measured construct.

1.5.3 Defining and Measuring Speededness

Despite the substantial amount of research on the topic, no commonly accepted method to determine if or to which degree a test is speeded has been established. Recent and comprehensive reviews on the topic of speededness have been published by Cintron (2021) and Jurich (2020). The historically most widely accepted rule set for determining whether a test

is speeded appears to be the so called *Swineford Guidelines* (Swineford, 1956, 1974). The Swineford Guidelines state that if (1) all examinees reach at least 75% of the items and (2) at least 80% of the examinees reach all of the items, a test can be considered not speeded. The definition explicitly contains the statement that a substantial part of the population (up to 20%) may not reach the end of a test, in fact may not finish up to 25% of the test, and the test is still considered not speeded. Therefore, already Swineford (1974, p. 9) noted that "these are arbitrary criteria and should not, of course, be too strictly applied." Swineford also noted that practical considerations play a major role in using this definition: "[If enough time for all test-takers to finish the test would be given,] the test supervisors would be faced with the problem of a restless group, eager to get away from the examination room." Other common definitions of test speededness include: "Speededness refers to the situation where the time limits on a standardized test do not allow substantial numbers of examinees to fully consider all test items" (Y. Lu & Sireci, 2007, p. 29); "A test is speeded when some portion of the test-taking population does not have sufficient time to attempt every item in the test within the allocated time" (Bejar, 1985, p. 1); "The larger the proportion of items examinees lack time to attempt, the greater is the speededness of the particular test" (Rindler, 1979, p. 261).

However, these historical speededness definitions have a common pitfall: If a test is speeded, one would expect smart and experienced test-takers to work with the appropriate level of speed, instead of omitting items, rapid guessing, or not reaching the end of the test. Tijmstra and Bolsinova (2018) refer to such an "ideal" speed level as a *target speed level*. Unfortunately, under most historical speededness definitions, a test administration would not be considered speeded if a test-taker adjusts their level of speed appropriately, as these speededness definitions (and speededness measurement models based on them) rely on the aforementioned symptoms of speededness. More precisely, these definitions and models rely on test-takers who do not deal with test speededness in an ideal way.

Finally, a central weakness of past speededness definitions is that speededness is seen as a property of the test. However, asking whether a test is speeded is in some sense like asking whether a test has a high average score or is reliable. With IRT, psychometricians have moved past the notion that expected average scores or reliability are intrinsic to a test but instead a function of ability (Samejima, 1977). Similarly, van der Linden (2011a, 2011b) was the first to propose a definition which moved passed this notion regarding speededness: He defines speededness as an interaction of the time limit of a test, the speed of the test-taker

and the workload of a test. This means that speededness is not a fixed property of the test but must be seen in relation to the properties of the test-taker. Therefore, this speededness definition of van der Linden (2011a, 2011b) is adopted throughout this thesis.

Unfortunately, this definition still does not provide researchers with an appropriate tool for measuring speededness. Yet, even if the speededness of an assessment for a specific test-taker could be measured after the test administration, it remains unclear how negative impacts of speededness could be remedied by statistical means (Tijmstra & Bolsinova, 2018). As Mollenkopf (1960, pp.228-229) already noted, explicitly referring to speededness: "And let me anticipate myself here a bit by saying that no statistical adjustment of the scores after the fact can control the problem, because this is like trying to catch the bees after the hive is upset. The proper thing to do, it seems to me, is to take steps to prevent luck from making any significant difference in the scores." When discussing research designs, Light et al. (1990, p. viii) once noted: "You can't fix by analysis what you bungled by design." It can be argued that this statement is equally true for designing assessments. Therefore, this thesis focuses on how speededness can be controlled during the design and assembly of tests.

1.6 Controlling Speededness in Test Assembly

In almost all testing contexts and regardless of whether speed is seen as a nuisance factor or a substantial part of the measured construct, being able to control the speededness of a test is of crucial importance. Controlling speededness is also crucial if multiple, parallel test forms are assembled. In such cases, differential speededness can lead to unfair results based on assignment of specific test forms, as the more speeded test form is then more difficult than a less speeded test form.

Controlling speededness in test assembly in general and ATA in particular has received only limited attention in the literature so far. The testing standards state that "For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure" (American Educational Research Association et al., 2014, p. 90). However, it is not only unclear (a) how such a speed component should be identified but also (b) how the appropriate speed component can be deliberately chosen.

An intuitive approach to controlling speededness in ATA is constraining the expected test times directly, for example based on mean item response times from pilot studies. Such an approach is, for instance, discussed by van der Linden (2005). Similar approaches using

mean, median, or standard deviation of response times are also discussed by Scrams and Smith (2010) as referenced in Lee and Chen (2011). However, these approaches are rooted in classical speededness definitions and ignore that test-takers usually work with very different working speeds. Beyond these basic approaches, van der Linden (2011a, 2011b) was the first to suggest a formal approach based on response time modeling. As was already mentioned above, this approach is also unique in that speededness is seen relative to a test-taker's speed.

1.6.1 van der Linden Approach

The approach by van der Linden (2011a, 2011b) is based on the 2PLN model for response times (van der Linden, 2006). The lognormal model is assumed for the item response time distributions conditional on a specific speed level. In addition, the approach takes advantage of the fact that the total test time distribution can be described by its cumulants, even though there is no known functional form of this distribution. More specifically, because the cumulants of the test time distribution are the sums of the cumulants of the item time distributions in the test, MILP constraints can be formulated for the cumulants of the test time distribution. As such, this approach can be used to set time limits on tests (van der Linden, 2011a) as well as to control the speededness of tests in ATA (van der Linden, 2011b).

Based on Equation 15, van der Linden (2011b) suggests that the total test time can be sufficiently approximated as a lognormal distribution. In addition, when the first two cumulants of the test time distribution are set, a corresponding lognormal distribution can be found. The first two cumulants (i.e., mean and variance) of the lognormal distribution for item k assuming the 2PLN model are defined as:

$$\mathrm{E}(RT_k|\zeta) = \exp\left(\lambda_k - \zeta + \frac{\sigma^2_{\epsilon_k}}{2}\right) \tag{20}$$

$$\mathrm{Var}(RT_k|\zeta) = \exp\left(2\lambda_k - 2\zeta + \sigma^2_{\epsilon_k}\right)\left(\exp(\sigma^2_{\epsilon_k}) - 1\right) \tag{21}$$

In his approach van der Linden (2011a, 2011b) also makes use of the following reparameterization, factoring out speed level ζ[8]:

$$q_k = \exp(\lambda_k + \frac{\sigma^2_{\epsilon_k}}{2}) \tag{22}$$

[8]For a more intuitive implementation and further explanations on the approach of van der Linden (2011a, 2011b), see Chapter 4.

$$r_k = \exp(2\lambda_k + \sigma^2_{\epsilon_k})(\exp(\sigma^2_{\epsilon_k}) - 1) \tag{23}$$

Parameters q_k and r_k can then be used to formulate quantitative constraints in the ATA framework. Using target value T_q (e.g., from an existing test form) and allowed deviation δ_q gives:

$$\sum_{k=1}^{j} q_{kf} x_{kf} <= T_q + \delta_q \tag{24}$$

$$\sum_{k=1}^{j} q_{kf} x_{kf} >= T_q + \delta_q \tag{25}$$

The same applies for the constraints of r_k:

$$\sum_{k=1}^{j} r_{kf} x_{kf} <= T_r + \delta_r \tag{26}$$

$$\sum_{k=1}^{j} r_{kf} x_{kf} >= T_r + \delta_r \tag{27}$$

Subsequently, van der Linden and Xiong (2013) extended his proposed approach for fixed-linear form testing to the shadow-test approach in the CAT framework. A potential application to MST seems straightforward. However, only a few studies have built on the ideas of or critically investigated the approach by van der Linden (2011b) so far. Finkelman et al. (2020) use the approach by van der Linden (2011b) to apply it to the CDM framework. Huang (2019) applies the approach to CDM-CAT. Veldkamp et al. (2017) focus on using the 3PLN model to constrain mean expected response time via ATA in the mixture modeling framework. Mixture modeling is frequently applied in response time modeling, as test-takers often quantitatively differ in their working speed, both between (some test-takers give their full effort, some only rapid guess) and within test-takers (some test-takers may adjust their working speed at the end of the test due to strict time-limits; Fox & Marianti, 2016).

As stated above, controlling the speededness of test forms during test assembly is one of the applications in which the 2PLN model has frequently been used without questioning the rather unconventional assumption of the 2PLN model of equal factor loadings (i.e., equal speed sensitivities) of items. It is both unclear (a) if using the restricted 2PLN model can be justified even if it does not hold in empirical applications and (b) how the 3PLN model could be used instead of the 2PLN model in test assembly. Furthermore, van der Linden (2011a,

2011b) focuses in his work on controlling the speededness on the level of test forms. Unfortunately, there are scenarios where such an approach is not sufficient, such as assessments that use parallel test forms with different item orderings under speededness conditions.

1.7 Parallel Test Forms and Differential Effects of Speededness

In high-stakes tests, it is common practice to use multiple, parallel test forms to increase test security (Bernardi et al., 2008; Chirumamilla et al., 2020; Smith et al., 2004). Based on the arguments presented so far, it is obvious that such test forms should be parallel regarding speededness to prevent test forms from being differentially speeded and therefore unfair. However, an argument can be made that it is not sufficient to control speededness on the test form level to guarantee fair test forms.

It is a common assumption when statistically modeling test speededness that speededness does not affect all parts of a test equally. Instead, it is assumed that later parts of the test are disproportionately stronger affected than early parts of the test. Practically, this means that test-takers are expected to start working with a slower than optimal working speed so they have to speed up at a later point in the test or not reach the end of the test (Mollenkopf, 1950). This assumption is frequently made and tested in mixture-modeling approaches which, for instance, model a switch in response behavior from solution to rapid guessing behavior (e.g., Goegebeur et al., 2008; Schnipke & Scrams, 1997; Yamamoto, 1995). Furthermore, in the literature, speededness is frequently regarded as a potential source of conditional dependencies. In this line of research, speededness is assumed to lead to local stochastic dependencies between items at the end of the test due to test-takers running out of time (e.g., Douglas et al., 1998; Yen, 1993).

Therefore, it can be assumed that item order and the speededness of a test administration can interact. If a test-taker runs out of time on five easy items at the end of the test, this will have a different effect compared to the test-taker running out of time on three difficult (and more time intensive) items instead. Already in the 1960s, researchers voiced their concerns that the effects of speededness may depend on the characteristics of items placed at the end of a test. Therefore, Sax and Cromack (1966) argue that test designers should order items in ascending order of difficulty. Similar concerns are raised by Leary and Dorans (1985) and Lawrence (1993). Oshima (1994) investigated the effect of running out of time at the end of a test on item and person parameter estimation. For this, he also included various item orders (random, easy-to-hard) and found substantial effects of speededness on

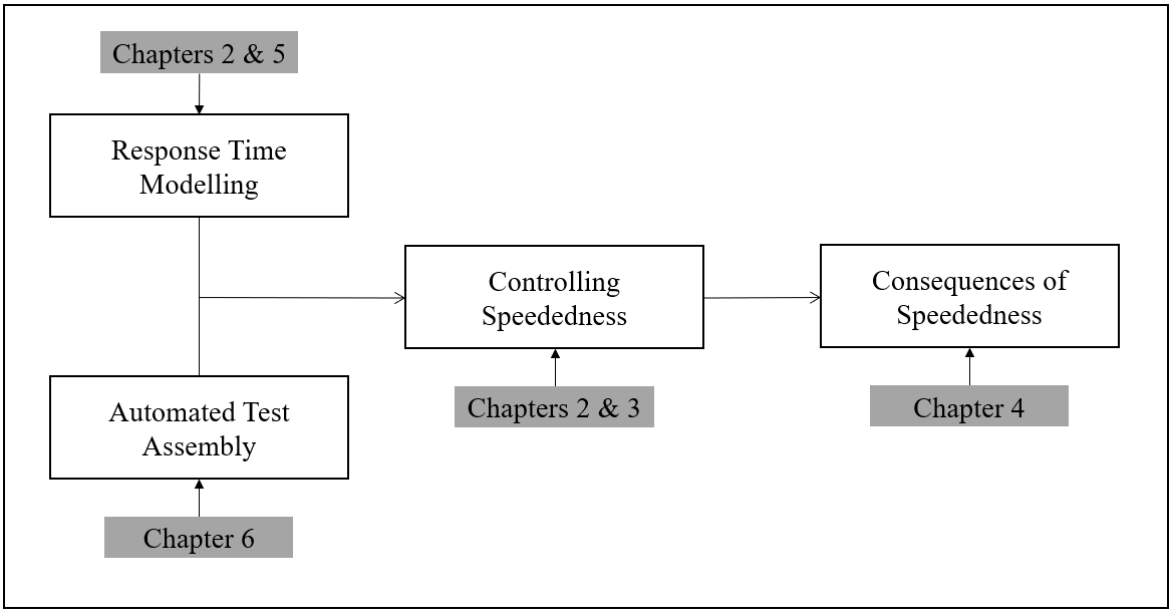
accuracy of parameter estimation. However, Oshima focused primarily on item parameter estimation and not so much on bias in person parameter estimation. This thesis argues that the impact on person parameter estimation could be much more impactful and substantial. The issue of differential effects of speededness due to different item orderings should therefore be investigated to assess whether it is indeed a major concern for the fairness of high-stakes tests.

1.8 Aims and Scope of the Present Work

It is apparent that controlling speededness when assembling test forms is crucial for the design of valid and fair assessments. To this day, a substantial number of studies on how to determine the degree of speededness of a test has been published (Cintron, 2021). However, there is little research on the topic of preemptively controlling the degree of speededness of an assessment. While van der Linden (2011a, 2011b) presents promising approaches for doing so, this work has important limitations. Van der Linden's approach focuses on using the 2PLN model, which makes a restrictive assumption regarding the speed sensitivity of items, as described above. Furthermore, even if test forms are perfectly controlled regarding speededness, test speededness can still interact with item order when multiple test forms are used. The goal of this thesis is to pinpoint weaknesses in the existing approaches for controlling speededness and to expand them beyond their limitations. Finally, even if approaches exist, this does not mean that they are applicable for assessment practitioners with limited technical experience and time resources. Therefore, this thesis explicitly aims at providing software as well as software tutorials for assessment practitioners enabling the implementation of the suggested approaches and procedures. An overview of the content of the present thesis can be seen in Figure 5. Chapters 2, 5, and 6, focus on laying the foundation for controlling speededness, namely response time modeling (Chapters 2 & 5) and automated test assembly (Chapter 6). Chapters 2 and 3 deal with controlling speededness and Chapter 4 deals with potential consequences of speededness on the fairness of assessments.

More precisely, Chapter 2 focuses on response time modeling and the differences between the 2PLN and the 3PLN model. In the literature, there is some confusion around the parameterization of the lognormal response time model. One goal of the Chapter is to provide detailed explanations on the conceptual meaning of the different parameters of the model, focusing on speed sensitivity ϕ and residual variance σ_ϵ^2. Another goal is to investigate how the 2PLN and 3PLN model perform in empirical situations. Finally, the main aim of the Chapter

Figure 5: *Overview of the Conducted Research.*



is to investigate if using the 2PLN model even though the 3PLN model were appropriate can impair the fairness of multiple test forms in the context of high-stakes testing. Chapter 3 builds on the ideas of Chapter 2 and investigates how the ATA framework of van der Linden (2011b), which uses the 2PLN model, can be generalized to the 3PLN model. Furthermore, the goal is to provide hands-on software solutions for practitioners who want to control the speededness of assessments in practice. Chapter 4 shifts the focus to the consequences of speededness, focusing on high-stakes tests in higher-education, such as university exams, in which multiple test forms with identical items but varied item orders are used. The Chapter illustrates that, if a test is speeded, varying item orders can have a substantial impact on test scores, even if the different test forms are otherwise completely identical. Furthermore, this Chapter seeks to give hands-on advice on how such effects can be mitigated in practical assessment situations.

Implementation of joint response time models in standard statistical software is currently sparse and challenging but can be essential for understanding the speed-ability relationship. Therefore, Chapter 5 provides an in-depth guide on how various response-time modeling approaches within the flexible, hierarchical framework of van der Linden (2007) can be implemented in the general purpose Bayesian estimation software Stan (Carpenter et al., 2017). Chapter 6, in contrast, focuses on the implementation of ATA methods in the statistical programming environment R. While extensive theoretical explanations on ATA (e.g., van der

Linden, 2005) and a small tutorial on ATA using lpSolve (Diao & van der Linden, 2011) exist, it is still very cumbersome for researchers to implement ATA in practice. The R package eatATA (Becker, Debeer, Sachse, & Weirich, 2021) aims at closing this gap, by providing an accessible user interface for general ATA methods. In Chapter 6, the focus is on introducing the eatATA package and providing guidance on how to use it for general ATA purposes. Chapter 7 provides a summary of the presented research findings. Important limitations of the findings are discussed and an outlook on future research topics is given.

2 On the Speed Sensitivity Parameter in the Lognormal Model for Response Times and Implications for High-Stakes Measurement Practice

Published as: Becker, B., Debeer, D., Weirich, S., & Goldhammer, F. (2021). On the speed sensitivity parameter in the lognormal model for response times and implications for high-stakes measurement practice. *Applied Psychological Measurement, 45*(6), 407-422, https://doi.org/10.1177/01466216211008530

Abstract: In high-stakes testing, often multiple test forms are used and a common time limit is enforced. Test fairness requires that ability estimates must not depend on the administration of a specific test form. Such a requirement may be violated if speededness differs between test forms. We investigated the impact of not taking speed sensitivity into account on the comparability of test forms regarding speededness and ability estimation. We compared the lognormal measurement model for response times by van der Linden (2006) with its extension by Klein Entink et al. (2009), which includes a speed sensitivity parameter. An empirical data example was used to show that the extended model can fit the data better than the model without speed sensitivity parameters. A simulation was conducted, which showed that test forms with different average speed sensitivity yielded substantial different ability estimates for slow test-takers, especially for test-takers with high ability. We therefore recommend the use of the extended lognormal model for response times for the calibration of item pools in high-stakes testing situations. Limitations to the proposed approach and further research questions are discussed.

2.1 Theoretical Background

In high-stakes assessments like college administration tests (e.g. SAT; College Board (2015)) or language proficiency tests (e.g. TOEFL; Educational Testing Service (2020)), important consequences result from test scores, such as admission to university or other educational programs. The high-stakes connected to the test outcome have important implications for the design and analysis of the respective tests. First, in order to increase test security, often multiple parallel test forms are used. This prevents cheating during testing sessions with multiple test-takers and sharing knowledge about the test by former test-takers (Luecht & Sireci, 2011). Second, for reasons of fairness, testing conditions are standardized across test-takers and test occasions. For instance, the time limit for the test is equal regardless of the test form. Third, due to the high-stakes, test-takers are often assumed to be highly motivated. Therefore, missing responses are commonly considered informative, that is, they are scored as incorrect responses. This scoring rule is communicated to test-takers, to prevent test-takers from strategically not responding to items they feel unable to provide a correct response to. Ignoring missing values as a scoring rule could incentivize test-takers to omit these items and thereby lead to biased and unfair ability estimates.

When multiple tests forms are used, they are often required to be parallel, which in the strict sense means that for every test taker, the test forms have the same true score and the same error variance (Lord & Novick, 1968, p. 48). Within an IRT framework where maximum likelihood is used to estimate ability, the expected ability estimate $E(\hat{\theta})$ as well as the expected standard error $E(\widehat{SE_\theta})$ for all test-takers should be independent of the administered test form z, which corresponds to so-called weak parallelism (Samejima, 1977). When missing responses are scored as incorrect, differences in the speededness of the test forms can violate this requirement[9]. Imagine one test taker, working at a specific speed, and a test with two test forms A and B that only differ in their expected testing time for the specific test taker. The time limit for the test administration is 60 minutes and the expected total response time of the test taker on test form A is 60 minutes but 70 minutes on test form B. When confronted with test form B, the test taker has to choose from three strategies:

a) Work with the identical speed as on test form A and not reach the end of the test,

b) work with the identical speed as on test form A and omit items, or

c) work with increased speed and respond to all items in time.

[9]given that ability and speed are distinct constructs, which appears to be a reasonable assumption (e.g., van der Linden, 2009a).

Missing responses resulting from (a) and (b) are scored as incorrect. Working with increased speed (c) usually leads to decreased accuracy (cf. the within-person speed-accuracy trade-off; Goldhammer, 2015). Hence, all strategies will result in a lower expected ability estimate on test form B compared to test form A. Combinations of the three strategies are also plausible but will have similar consequences on the ability estimate.

The example illustrates that the speededness of a test is an interaction of the time intensity of its items, the time limit set on the test and the exerted working speed of the test taker (van der Linden, 2011b). As the speed level usually varies between persons, the degree of speededness of a test can also be expected to vary between persons. A fast and proficient test taker will score higher on a test with a time limit than an equally proficient but slower test taker that has to engage in one of the above described strategies to deal with the insufficient time available. Consequently, however, the measured latent construct is no longer a pure ability measure, but a composite measure of speed and ability. Whether this is seen as a conceptual property of the test or a byproduct of the testing conditions differs. In this paper we make no assumptions on the nature of speed differences between persons and to which degree they should affect ability measurement in high-stakes testing[10]. Instead, we focus on how to hold the level of speededness constant across all test forms within each individual test taker.

In the following section, we briefly outline the typical test assembly process and analysis that is commonly performed to obtain individual ability estimates in high-stakes assessments. Based on this, we describe the state-of-the-art approach to prevent differentially speeded test forms, which uses latent response time modeling. We explain an important shortcoming of this model and discuss a common model extension that mitigates this shortcoming.

2.1.1 Assessment Framework

Test Assembly. The common process of creating multiple parallel test forms contains of the following steps (College Board, 2015; van der Linden, 2005): (1) developing items, (2) using items on a piloting sample (Piloting) (3) item parameter estimation (Calibration), (4) assembly of items from an item pool to parallel test forms (Test Assembly). Criteria for the assembly of tests, besides test speededness, include the test information function, comparability of content, and similar distribution of item types (van der Linden, 2005). Due to the emergence of computer administered testing, balancing speededness has become sub-

[10]For a discussion of this issue see, for example, the work of Tijmstra and Bolsinova (2018).

stantially easier. In this paper we assume that response times are available from a computer administered piloting study.

Ability Estimation. For the estimation of latent abilities an often-used choice is the 2PL model. As already described, we assume that missing responses are scored as incorrect. Throughout this paper, we adopt the notation of Fox (2010), denoting items as $k = 1, ..., j$ and persons as $i = 1, ..., n$, with correct responses denoted as $y_{ik} = 1$. In the 2PL model, the probability to solve an item k correctly can be denoted as

$$P(y_{ik} = 1|\theta_i, a_k, b_k) = \frac{exp(a_k\theta_i - b_k)}{1 + exp(a_k\theta_i - b_k)}. \tag{28}$$

2.1.2 Balancing Speededness

Several strategies have been proposed to balance speededness across the test forms of a test administration, for example using observed response times from a piloting study (e.g., van der Linden, 2005). In the following section, we discuss the current state-of-the-art approach, which uses a latent measurement model for response times.

Lognormal Measurement Model. Recently, van der Linden (2011b) proposed the use of a lognormal latent measurement model for response times (van der Linden, 2006) for balancing speededness across test forms. The model assumes responses times to be lognormally distributed and parameterizes these lognormal response times $\ln RT_{ik}$ as

$$\ln RT_{ik} = \lambda_k - \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{29}$$

λ_k represents the *time intensity* of a specific item, while ζ_i represents the average speed with which a person works (*person speed parameter*). In addition, an item-specific residual variance $\sigma^2_{\epsilon_k}$ is estimated. As the model is parameterized with two item specific parameters, we refer to it as the *two-parameter lognormal* (2PLN) model. In his paper, van der Linden (2011a) proposed controlling the expected testing time, conditionally on the speed parameter, according to the 2PLN model. He showed that this approach performed better than using observed response times to balance speededness across test forms (van der Linden, 2011b), as it for example also controls for differing variances in response times between items.

Speed Discrimination. According to the 2PLN model, items can have different intercept parameters λ_k and different residual variances $\sigma^2_{\epsilon_k}$. Furthermore, van der Linden (2006)

introduced the inverse of the residual variance as the discrimination parameter α_k:

$$\alpha_k = \frac{1}{\sigma^2_{\epsilon_k}}. \tag{30}$$

α_k thereby represents the precision of the response time distribution (Molenaar, Tuerlinckx, & van der Maas, 2015a). In this manuscript, however, to avoid confusion, we will only refer to the residual variance, and not to its inverse. Compared to models from confirmatory factor analysis, the 2PLN model resembles a tau-equivalent measurement model (Brown, 2006, pp. 236–252) for log response times. This means the model lacks a slope parameter and therefore, speaking in terms of more generalized models, assumes that the slope parameter is equal across all items or indicators. This equals the assumption that items with equal residual variances correlate all equally strong with the measured latent construct. Conceptually speaking for response time modeling, this means that the 2PLN model assumes that items do not differ in their sensitivity to speed differences across persons.

In this paper, however, we will argue that items can differ in the extent to which they are sensitive to speed differences, and that this variability across items needs to be taken into account when assembling test forms that should have equal speededness for each test taker. In the next section, we will discuss an extension of the lognormal measurement model for response times which allows differences in speed sensitivity across items.

Extension of the Lognormal Measurement Model. Klein Entink, Fox, and van der Linden (2009) proposed an extension of the 2PLN model which we call the *three-parameter lognormal* (3PLN) model. It introduces a slope parameter ϕ_k. This measurement model resembles a congeneric measurement model for log-transformed response times in confirmatory factor analysis (Brown, 2006, pp. 236–252):

$$\ln RT_{ik} = \lambda_k - \phi_k \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{31}$$

Conceptually, the parameter ϕ_k allows for individual items being more sensitive to speed differences between test-takers than other items. To avoid confusion with the α_k parameter that van der Linden (2006) labels as a discrimination parameter in the 2PLN model, we will use the term *speed sensitivity* to refer to ϕ_k throughout this paper.

Difference between the 2PLN and the 3PLN model. There has been some confusion around the 2PLN and the 3PLN model and the meaning of their respective item

parameters in the literature[11]. It is important to note that the 2PLN and the 3PLN models are not equivalent formulations of the same model. This can be illustrated by comparing the model implicit correlations between the response times of two items k and l of the 2PLN and the 3PLN model. For the 2PLN model this correlation is defined as

$$\rho_{RT_k,RT_l} = \frac{\left[\exp\left(\sigma_\zeta^2\right) - 1\right]}{\sqrt{\left(\exp(\sigma_{\epsilon_k}^2 + \sigma_\zeta^2) - 1\right)\left(\exp(\sigma_{\epsilon_l}^2 + \sigma_\zeta^2) - 1\right)}}. \tag{32}$$

In contrast, for the 3PLN model this correlation is defined as

$$\rho_{RT_k,RT_l} = \frac{\left[\exp\left(\phi_k\phi_l\sigma_\zeta^2\right) - 1\right]}{\sqrt{\left(\exp(\sigma_{\epsilon_k}^2 + \phi_k^2\sigma_\zeta^2) - 1\right)\left(\exp(\sigma_{\epsilon_l}^2 + \phi_l^2\sigma_\zeta^2) - 1\right)}}. \tag{33}$$

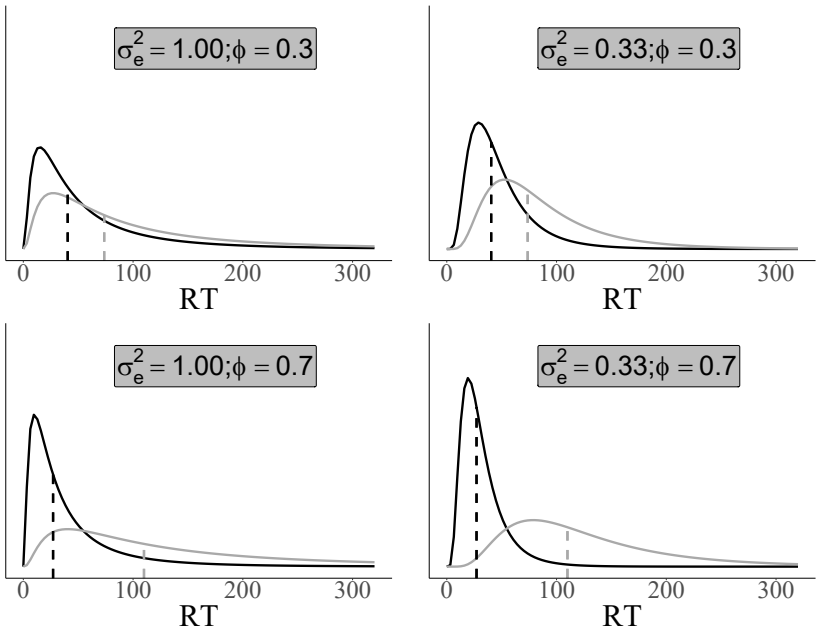
For the derivation of both formulas see Appendix A.1. For a similar remark on the model implicit covariances of the response times of two items, see Fox and Marianti (2016).

To illustrate the difference between the residual variance and the speed sensitivity parameter, Figure 6 shows response time distributions conditional on two different speed levels ($\zeta_1 = 1$, $\zeta_2 = -1$) for four different items. The left side of the figure shows the distributions for items with a high residual variance $\sigma_{\epsilon_k}^2 = 1.00$, the right side for items with a low residual variance, $\sigma_{\epsilon_k}^2 = 0.33$. Furthermore, the upper half of the figure depicts the distributions for items with low speed sensitivity, $\phi_k = 0.3$, the lower half items with high speed sensitivity, $\phi_k = 1$. The graphs illustrate how the residual variance controls the broadness of the distributions (and is strongly connected to the concept of reliability), while ϕ_k controls how far the medians of the response time distributions differ between persons with differing speed levels. An identical figure for the log transformed response times can be seen in Appendix A.2.

As an illustration of the conceptual meaning of the speed sensitivity of items, consider the following two hypothetical math items with equal time intensity (e.g., $\lambda_1 = \lambda_2 = 4$). The first item embeds a simple task in a long text; the second item has no text to read, but requires a lengthy calculation. It seems plausible to assume that the second item is more sensitive to working speed specific to math items (e.g., $\phi_1 = 0.7$), because the calculation is longer. In contrast, the first item could be less sensitive to mathematical working speed, because the response time mostly depends on the reading speed ($\phi_2 = 0.3$) . As reading and mathematical literacy are assumed to be distinct constructs, this is plausibly also the case for

[11]This may be caused by the labeling of the inverse of the residual variance as a discrimination parameter, which is usually a term used for slope parameters. For example, Bertling and Weeks (2018) cite van der Linden (2006) but introduce the model with α_k as a slope parameter instead of the inverse of the residual variance.

Figure 6: *Conditional Response Time Distributions for a Fast Speed Level with $\zeta_1 = 1$ (Black Line) and a Slow Speed Level with $\zeta_2 = -1$ (Grey Line) on Four Different Items, all with $\lambda_k = 4$. Dashed Lines Indicate the Medians of the Corresponding Distributions.*



reading and mathematical speed. The consequences for the Response Time Characteristic Curve, as also described in Fox (2010), can be seen in Appendix A.3. These two items would not lead to differences in response times for medium speed levels ($\zeta_k = 0$) but to substantial differences for slow (ζ_k = -1) and fast test-takers ($\zeta_k = 1$), with differences increasing with increasing deviation from $\zeta_k = 0$. For a test-taker with $\zeta_i = -1$ the expected response times of the two example items are 73.70 and 109.95 seconds. As time pressure usually only occurs for slow participants, generally only differences in response times for slow but not for fast participants will be relevant for the estimation of ability in educational assessments.

Hierarchical Framework. For model estimation in the context of test assembly, van der Linden (2011a) proposed embedding the lognormal latent measurement model for response times in a hierarchical framework (van der Linden, 2007). The resulting model assumes two latent dimensions, ability and speed, with common item and person parameter distributions. Conditional on these joint distributions, the model assumes independently distributed responses and response times. The framework benefits the estimation of the two dimensions, especially if the two dimensions are correlated (van der Linden et al., 2010). The joint person parameter distribution with either the 2PLN or the 3PLN model is a multivariate normal

distribution with

$$(\theta_i, \zeta_i) \sim \mathcal{N}(\boldsymbol{\mu_P}, \boldsymbol{\Sigma_P}). \tag{34}$$

The joint item parameter distribution with the 2PLN model together with a 2PL model for ability is also a multivariate normal distribution[12] with

$$(a_k, b_k, \lambda_k) \sim \mathcal{N}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}). \tag{35}$$

The joint item parameter distribution with the 3PLN model together with a 2PL model for ability also includes ϕ_k:

$$(a_k, b_k, \lambda_k, \phi_k) \sim \mathcal{N}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}). \tag{36}$$

2.1.3 Research Questions

The questions arise, whether

- the hierarchical framework with the 3PLN model as a measurement model for response times fits empirical response time data better than the hierarchical framework with the 2PLN model and, if this is the case,
- what the consequences would be for ability estimation in high-stakes assessments.

To our knowledge, hierarchical frameworks with the 2PLN model and the 3PLN model have not yet been compared using data from educational competence tests. Moreover, there have only been a few comparisons using empirical data at all, so far focusing on intelligence tests (Goldhammer & Klein Entink, 2011), complex problem solving tasks (Scherer et al., 2015), and mental rotation tasks (Debelak et al., 2014). In all three studies the framework with the 3PLN model showed better fit than the framework with the 2PLN model according to the DIC (Spiegelhalter et al., 2002). In addition, the hierarchical framework with the 3PLN model has been applied to non-educational vocational credentialing high-stakes data (Fox & Marianti, 2017) and low-stakes data of chess tasks (Fox & Marianti, 2016). In both cases substantial variance in the speed sensitivity parameter was found across the items. The aforementioned studies provide general evidence for the relevance of the proposed model extension. However,

[12]Note that van der Linden (2006) also includes the inverse of the residual variance $\sigma^2_{\epsilon_k}$ in the joint item parameter distribution. For better comparability with the 3PLN model in the hierarchical framework we slightly modify the joint item parameter distribution by assuming a univariate distributed item specific residual variance, independent from the distribution of the other item parameters (see, for example, also Pohl et al., 2019).

they do not focus on educational assessment data. Therefore, we conducted an empirical data analysis, in which we applied and compared the hierarchical framework with the 2PLN model and the 3PLN model to data from an educational assessment, to investigate whether items differ in their speed sensitivity. This analysis is discussed in the "Empirical Data Analysis" section below.

If the appropriateness of the model extension indeed holds in educational competence testing and items vary in their speed sensitivity, those differences may also accumulate over test forms of educational high-stakes assessments. This could result in test forms that, despite having equal time intensities and similar average observed response times, differ in their sensitivity to speed differences and therefore in their conditional distributions of expected testing times. Especially the substantial differences in expected response times for slow test-takers would be important, as they could lead to differences in ability estimates across test forms. In the section "Simulation Study", we investigate and describe the possible consequences of unbalanced test forms on ability estimation using simulated data from test forms with item properties as found in the empirical example.

2.2 Empirical Data Analysis

2.2.1 Data Description

For the empirical data analysis, we used data from the 2015 Programme of International Student Assessment (PISA, OECD, 2016b), for which responses and response times on item level are publicly available. The competences measured by PISA resemble competences that are often assessed in high-stakes educational assessments. Note that it is not uncommon to calibrate items for a high-stakes context based on data from low-stakes conditions, when piloting in high-stakes conditions is cumbersome or impossible (e.g. College Board, 2015; Educational Testing Service, 2020). In those situations, it is implicitly assumed that items function similarly in low- and high-stakes conditions. In that sense, the results of this empirical low-stakes data analysis also have implications for high-stakes assessments. The Canadian subsample was chosen because it is the largest among the 72 countries participating in PISA.

To avoid substantial numbers of missing responses by design, we analyzed test booklets separately and included only the test-takers who had worked on the respective booklet. In PISA 2015, every test form consisted of four booklets and booklets were assembled to a whole of 66 different test forms in the computer administered version. Returning to items within a booklet was only possible within the items sharing a common stimulus and otherwise

prohibited. Response times were accumulated across multiple visits of the same item (OECD, 2016b, pp. 45–47). We analyzed all math booklets used in the assessment (named "M01" - "M05" and "M06ab"), which appeared each in overall eight different test forms, at every position twice. For simplicity, we dichotomized all polytomous items, scoring fully correct responses as correct and partially incorrect responses as incorrect. This resulted in data sets of 10 to 12 dichotomous items and 1863 to 1929 persons.

2.2.2 Methods

The software `JAGS` (Plummer, 2017) together with the R package `rjags` (Plummer, 2016) was used for model estimation. We used the hierarchical framework with both the 2PLN and the 3PLN model to analyze the data set. In the actual analysis of the PISA data set, omitted responses are scored incorrect and number of not reached responses is used as a manifest variable in the background model for the plausible value generation (OECD, 2016b). Because the aim of this empirical example is the unbiased estimation of item parameters (as in an actual pilot study for a high-stakes assessment), all missing responses were treated as if the items were not administered to the corresponding persons, which is the recommended practice for estimating item parameters (Finch, 2008).

Model estimation. Priors were uninformative and chosen in correspondence to Fox (2010) and Pohl et al. (2019). An inverse Wishart distribution was used as a hyperprior for the distribution of the three (b_k, a_k, λ_k) or respectively four item parameters (b_k, a_k, λ_k, ϕ_k). Further information on the prior distributions can be seen in Appendix A.4. The DIC was calculated and compared between the two models to assess model fit (Spiegelhalter et al., 2002). The posterior distributions of the speed sensitivity parameters and their mean and standard deviation were investigated.

2.2.3 Results

Inspections of the MCMC chains were conducted using the R packages `coda` (Plummer et al., 2006) and `rjags`. Trace plots indicate good convergence for all parameters in both models in all data sets. The point estimates of the univariate potential scale reduction factors (Gelman & Rubin, 1992) for all parameters in all booklets were below 1.03 (95% upper confidence interval limits at or below 1.10) and below 1.05 (95% upper confidence interval limits at or below 1.19), for the framework with respectively the 2PLN and the 3PLN model. This indicates satisfactory convergence (Gelman & Shirley, 2011). The correlation of the person

ability and person speed parameter ranged between $r_{\theta_i\zeta_i} = -.62$ in booklet "M01" and $r_{\theta_i\zeta_i} = -.49$ in booklet "M02", indicating a medium negative relationship between ability and speed. Similar results have been reported and are often explained by the fact that test-takers need more time if they actually solve an item (Debelak et al., 2014; Goldhammer & Klein Entink, 2011; Scherer et al., 2015). If test-takers are not able to solve an item, they may guess and move on to the next item.

Regarding model fit, DIC indicated better fit with the 3PLN model as a measurement model for all booklets (Appendix A.5). Table 1 shows the statistics for the resulting speed sensitivities for all booklets. The mean of speed sensitivities M_{ϕ_k} within booklets ranged from to 0.37 to 0.47, while SD_{ϕ_k} ranged from 0.32 to 0.36. The 95% Highest Posterior Density (HPD) interval for the standard deviation SD(ϕ_k) excluded 0 for all booklets. These findings provide evidence that there was substantial variation in the speed sensitivity across items in the empirical data.

Table 1: *Descriptive Statistics of Item Speed Sensitivity within all Math Booklets.*

Booklet	$M(\phi)$	$SD(\phi)$	95 % HPD	$Min(\phi)$	$Max(\phi)$	$r_{\phi,b}$	$r_{\phi,a}$	$r_{\phi,\lambda}$
M01	0.40	0.34	[0.20, 0.47]	0.15	0.75	0.28	-0.09	0.21
M02	0.37	0.36	[0.21, 0.52]	0.14	0.57	0.19	0.12	0.16
M03	0.39	0.33	[0.20, 0.46]	0.29	0.53	0.13	0.02	0.08
M04	0.42	0.34	[0.20, 0.46]	0.18	0.67	0.27	0.28	0.16
M05	0.44	0.32	[0.19, 0.44]	0.25	0.66	0.21	-0.01	0.18
M06ab	0.47	0.35	[0.21, 0.48]	0.19	0.69	0.30	0.26	0.19

Note: Descriptive statistics for speed sensitivity, including its mean M(ϕ_k), standard deviation SD(ϕ_k), the HPD interval for the standard deviation SD(ϕ_k), Minimum (Min(ϕ_k) and Maximum (Max(ϕ_k), and correlations of speed sensitivity with the other item parameters.

We also investigated the correlations of the speed sensitivities with other item parameters. Table 1 displays the means of the posterior distributions of these correlations. Speed sensitivity correlated low but consistently over all booklets with the time intensity parameter λ_k and difficulty parameter b_k. There was more variation across booklets in the correlation with the discrimination parameter a_k, but correlations were still small or close to zero. The small correlations imply that the speed sensitivity parameter is largely independent from the other item parameters and would not be indirectly balanced if the other item parameters were balanced between test forms.

2.3 Simulation Study

2.3.1 Design

The performed empirical data analyses illustrate that it is plausible to assume differences between items regarding their speed sensitivity. Therefore, the question arises, how the fairness of test forms is affected if this speed sensitivity is not controlled for between test forms. Based on the findings and parameters distributions in the empirical analyses, a simulation study was conducted to investigate how differences in speed sensitivity across test forms affect ability estimates. The simulation study reflects the operational phase of a high-stakes assessments in which item properties are known from prior piloting and the sole interest lies in person parameter estimation. We created three test forms, each with 30 items. The item parameters for the first test form were drawn from a multivariate normal distribution. Means, variances and covariances of the item parameters were set to be in accordance with the results obtained from the empirical data analysis (see Appendix A.6). ϕ_k and a_k were truncated at 0. If an item parameter draw included any ϕ_k and a_k smaller than 0, all item parameters were drawn again for this replication. The residual variance of the log response times was drawn from a univariate normal distribution with $\sigma^2_{\epsilon_k} \sim \mathcal{N}(\mu = 0.2, \sigma^2 = 0.1)$, also truncated at 0. The first test form, with $\mu(\phi_k) = 0.3$ is referred to as the *low speed sensitivity test form*. A second and third test form were created with identical item parameters but shifts in their average speed sensitivity, resulting in a *medium speed sensitivity test form* with $\mu(\phi_k) = 0.4$ and a *high speed sensitivity test form* with $\mu(\phi_k) = 0.7$. The difference in speed sensitivity between the first and second test form reflects a common difference between booklets, which can also be found in the empirical example. Therefore, the comparison between these booklets can be used to determine expected bias even if only a few test forms are assembled. The difference between the first and third test form reflects a more extreme but not implausible case[13]. This condition was chosen to illustrate the theoretically possible impact of differing speed sensitivities and potential bias if a large number of test forms is assembled. Person parameters were chosen to enable conclusions about the effect of the two differing test forms on all possible combinations of speed and ability. Therefore, we sampled 500 ability parameters from $\theta_i \sim \mathcal{N}(0, 1)$ and combined these with four different levels of speed, $\zeta_i = [-1; -0.5; 0.5; 1]$. This resulted in a complete sample of $n = 2000$ test-takers across the four speed subgroups. We simulated responses and response times of the complete

[13] In the empirical example, SD(ϕ_k) within booklets was around 0.35 and the range across booklets for ϕ_k was 0.6.

sample working on both test forms according to the hierarchical framework with the 3PLN and the 2PL model. We set the time limit to 65 minutes (3900 seconds) to introduce a reasonable amount of not reached items into the simulation. Overall, 500 replications were conducted.

2.3.2 Methods

Person abilities were estimated according to the 2PL model, with known item parameters using the weighted likelihood estimator (WLE) (Warm, 1989) via the R package `TAM` (Robitzsch et al., 2017). Not reached items were scored as incorrect. This approach reflects a high-stakes assessment, in which item parameters are obtained from a previously conducted calibration study and ability estimation is the focus (without specifically considering speed in the estimation). We compared numbers of not reached items and estimated ability for the four different speed groups between the three test forms.

2.3.3 Results

As can be predicted from the response time measurement model in Equation 31 and the response time characteristic curves described in the introduction, differences in cumulative response times between the three test forms were most severe for the fastest and slowest participants (Table 2)[14]. The fastest subgroup was much faster than the time limit of 3900 seconds, with means of 1310.08 seconds and 1953.53 seconds for the high and the low speed sensitivity test forms. In contrast, the slowest subgroup working on the high speed sensitivity test form was, on average, substantially slower than the time limit, with a mean of 5419.48 seconds. In the faster subgroups, the differences in testing time did not result in different numbers of not reached items, because for all test forms the testing times were well below the time limit. For the slowest participants, however, the medium and high speed sensitivity test form led to substantially more not reached items than the low speed sensitivity test form. Detailed numbers for items not reached on average can be seen in Table 3 and are depicted in Appendix A.7 for a single replication.

These differences in number of not reached items also resulted in differences in ability estimates, mainly for the slowest subgroup. For them, the average difference in ability estimation between the test forms with low and medium speed sensitivity was 0.09 and 0.51 between the test forms with low and high speed sensitivity. Higher average speed sensitivity

[14]Table 2 contains mean statistics across all replications, while standard deviation for the identical statistics across replications can be found in Appendix A.8

Table 2: *Test Statistics per Test Form and per Speed Group, Averaged Across All Replications.*

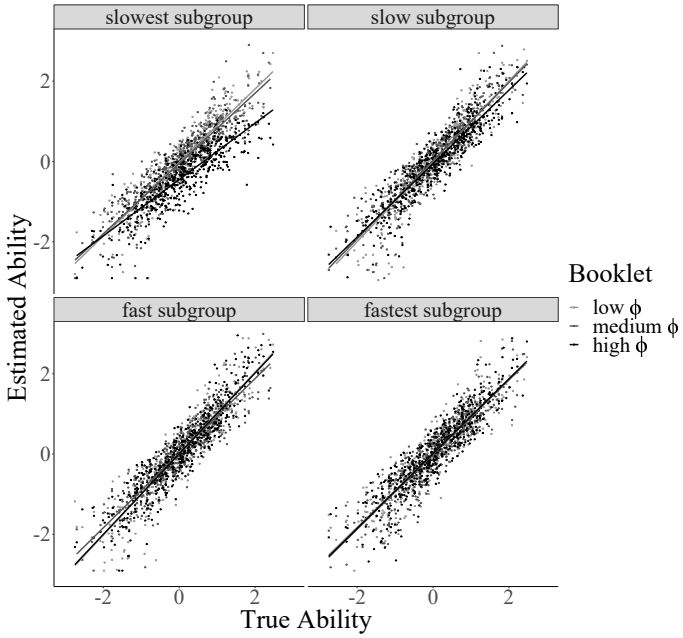
Test Form	ζ_i	$M(RT)$	$SD(RT)$	$M(mis)$	$SD(mis)$	$cor(\hat{\theta}, \theta)$	RMSE	$M(\Delta_\theta)$
low ϕ	slowest	3636.44	371.73	0.02	0.04	0.90	0.47	-0.04
low ϕ	slow	3102.47	313.97	0.00	0.01	0.91	0.45	-0.00
low ϕ	fast	2275.08	228.36	0.00	0.00	0.91	0.45	-0.00
low ϕ	fastest	1953.53	195.04	0.00	0.00	0.91	0.45	-0.00
medium ϕ	slowest	4019.63	410.14	0.06	0.07	0.89	0.51	-0.12
medium ϕ	slow	3261.12	329.52	0.00	0.02	0.91	0.46	-0.01
medium ϕ	fast	2163.99	217.06	0.00	0.00	0.91	0.45	-0.00
medium ϕ	fastest	1767.68	176.80	0.00	0.00	0.91	0.45	-0.00
high ϕ	slowest	5424.72	553.29	0.29	0.08	0.82	0.83	-0.55
high ϕ	slow	3788.69	382.56	0.03	0.06	0.90	0.48	-0.06
high ϕ	fast	1862.25	186.12	0.00	0.00	0.91	0.45	-0.00
high ϕ	fastest	1310.08	130.71	0.00	0.00	0.91	0.45	-0.00

Note: Descriptive statistics are depicted for mean cumulative response times M(RT) and the corresponding standard deviation SD(RT), mean proportion of missings M(mis), the corresponding standard deviation SD(mis), correlation between true and estimated ability $cor(\theta, \theta)$, root mean square error (RMSE) and average difference between true and estimated ability $M(\Delta_\theta)$

resulted in substantially lower ability estimates. A difference of 0.51 in the ability logit for a test taker with a true ability $\theta_i = 0$ is equal to a drop from the 50th ability percentile to the 32th ability percentile. However, differences in ability estimation were not homogeneous within the slowest subgroup. Especially slow participants with high ability had substantially different ability estimates depending on the test form (see the upper left graph in Figure 7). Average differences in ability estimation were also calculated for the quantile including only the 25% most able test-takers, resulting in differences in ability estimation of 0.15 between the low and medium and 0.78 between the low and high speed sensitive test forms. This was to be expected, because for slow but high ability test-takers there are many not-reached items (scored as incorrect) that they could have answered correctly under sufficient time conditions. This is not the case for slow and low ability test-takers, for which only minor differences in estimated abilities across the test forms occurred. Furthermore, differences in speed sensitivities between test forms resulted in higher root mean square errors (RMSE) and lower correlations between estimated and true ability parameters (see Table 2) for more speed sensitive test forms.

To conclude, the simulation shows that differences in speed sensitivity between test forms can lead to substantial differences in ability estimates especially for slow and able test-takers. This finding is independent from whether speed is seen as a nuisance parameter or part of the construct to be measured. Furthermore, if speed is seen as a nuisance parameter, the high

Figure 7: *True and Estimated Ability for the Low and High Speed Sensitivity Test Form, Across the Four Subgroups. Results Shown for a Randomly Selected Single Replication.*



speed sensitivity test forms lead to a more biased and less precise ability measurement. If speed is seen as a substantial part of the construct to be measured, differences between true and estimated ability are in fact desirable for slow test-takers, however should be identical across test forms.

2.4 Discussion

High-stakes assessments often require multiple test forms with equal speededness at the level of the test taker. So far, the use of average response times and the use of the lognormal measurement model for response time model by van der Linden (2006) have been proposed as strategies to control speededness across test forms (van der Linden, 2011b). We compared the 2PLN model to the extension of the 3PLN model by Klein Entink, Fox, and van der Linden (2009), which introduces a speed sensitivity parameter into the measurement model. We investigated which measurement model, embedded in the hierarchical framework by van der Linden (2007) fits empirical competence data better. Indeed, the 3PLN model showed better model fit and the estimated speed sensitivity parameters varied substantially across items. This implies that balancing test forms using either observed response times or the item parameters from the 2PLN model can lead to unbalanced speed sensitivity across test forms. Moreover, our simulation study shows that when missing responses are treated as incorrect

(a standard practice in high-stakes assessments) differences in speed sensitivity between test forms can lead to severe differences in ability estimation. Especially slow test-takers with a high ability were affected, because they had increased numbers of not reached items in the test forms that had higher speed sensitivities.

The issue of differential speed sensitivity can also be illustrated from an alternative perspective: As stated before, we assume that high-stakes tests usually are speeded power tests and therefore that the ability measured in the test is a composite measure of ability and speed. However, this composition changes between test forms if the test forms differ in their speed sensitivity. If a test form has a high speed sensitivity and a time limit induces time pressure for a certain speed level, the proportion of speed in the composite measure can be considered quite high. If in the same scenario a test form has low speed sensitivity, however, the proportion of speed in the composite measure for this test form will be rather low. We argue that the influence of speed on the ability estimation has to be the same across test forms within each speed level.

2.4.1 Practical Implications

We draw the following conclusions regarding the practice of assembling test forms for educational high-stakes assessments: Right now, the use of the hierarchical framework with the 2PLN model is the state-of-the-art approach when balancing test forms. However, our findings suggest that only when

- the hierarchical framework with the 2PLN model proves to better fit the data than the framework with the 3PLN model (e.g., using DIC) or
- the 3PLN model shows low variation in the speed sensitivity parameter across items,

this approach should be considered sufficient. In cases where the framework with the 3PLN model shows better model fit and items differ in their speed sensitivity, using only the hierarchical framework with the 2PLN model could lead to unfair testing situations. To be more precise, the ability estimates and the rank order of test-takers could heavily depend on the administered test form, especially for slower test-takers. Instead, the hierarchical framework with the 3PLN model should be used when calibrating the items and not only the average testing time, but also the sensitivity to speed differences should be balanced across test forms.

Another common alternative for the assembly of fair test forms is the approach of assembling unspeeded test forms. Because in most educational assessments speed is a nuisance

parameter that is conceptually not part of the construct being measured, this strategy seems promising. However, our results and results from previous studies (e.g., van der Linden & Xiong, 2013) indicate that this approach might be unfeasible because there are generally large differences in the time that test-takers require to respond to all items in an assessment (see Table 2). Assuring that even the slowest test-takers can work without time pressure would imply a time limit that is far too generous for fast test-takers and problematic both from an economical as well as from a motivational perspective. Furthermore, our results have important implications for determining the speededness of a test: So far, often experimental methods using different time limits or different numbers of items in the same time limit have been used (e.g. Bridgeman, Cline, & Hessinger, 2004; Bridgeman, Trapani, & Curley, 2004; Harik et al., 2018). But while for the majority of the test-takers more generous time limits might only have a small impact on the demonstrated ability, different time limits can still substantially affect the slowest part of the population. This effect can only be disentangled by explicitly modeling speed. If differences in ability estimation for different time limits are averaged over all test-takers or calculated for different ability levels, the degree of speededness of the test for slow test-takers could be severely underestimated. Therefore, tests that have been examined using the aforementioned experimental methods could have been falsely classified as unspeeded.

2.4.2 Limitations

There are a number of limitations to our study: First, our real data analysis is based on low-stakes data while implications are mainly relevant for high-stakes assessments. However, similar analyses on (non-educational) high-stakes data have reported similar findings (Fox & Marianti, 2017). In addition, it is not uncommon that pilot studies for item pool calibrations are conducted under low-stakes conditions. Furthermore, we do not conclude that the hierarchical framework with the 3PLN model will always demonstrate better model fit than the framework with the 2PLN model for item pools of high-stakes assessments. Rather, we argue that the assumption of equal speed sensitivity across items should be tested, just like the assumption of equal factor loadings should be tested in confirmatory factor analysis or structural equation modeling (Brown, 2006).

A second limitation relates to a general limitation of the hierarchical framework, namely the assumption of stationarity (van der Linden, 2007). The model assumes that given the common distribution of the person and item parameters, residuals between responses and

response times are independent. The assumption is for example violated if participants substantially speed up or slow down during the test. This could happen in high-stakes assessments with a time limit, if test-takers speed up when they feel they are running out of time. However, for test assembly purposes only item parameters and their relations across items are of interest. If position effects are controlled for (similar to controlling for position effects of ability item parameters estimation; e.g., Gonzalez & Rutkowski, 2010)) speeding up might only affect the precision of item parameter estimation. Avoiding speeding up seems easiest, if items were piloted in low-stakes settings.

A third limitation is that our study deals with a specific violation of the assumptions of the 2PLN model. In the past, assumptions of the hierarchical framework using the 2PLN or 3PLN model for response times have been critically reviewed using empirical data analyses (Bolsinova & Tijmstra, 2018; Entink et al., 2009; Fox & Marianti, 2016; Ranger & Ortner, 2012b). Criticism includes violations of the assumption of lognormally distributed response times and the stationarity assumption mentioned above. Although the lognormal distribution has been the standard for modeling response times in educational assessments, future research could explore alternatives, possibly also embedded in the hierarchical framework.

2.4.3 Outlook

In the past, *automated test assembly* procedures (ATA) have been developed to enable the assembly of multiple test forms from large item pools under various constraints (van der Linden, 2005). These methods are already frequently used in practice (Luecht & Sireci, 2011). To enable the use of the 2PLN model in ATA, van der Linden (2011a, 2011b) reparameterized the model. Future research should investigate how the 3PLN model can be used best in automated test assembly and if a similar reparameterization approach might be feasible.

The 3PLN model could also be useful to determine the speededness of assessments under various time constraints for different test taker populations without having to experimentally investigate all possible combinations. This would especially be valuable for determining test accommodations for students with disabilities (Lovett, 2010). Furthermore, while the current paper focuses on fixed-test forms, our findings can also be applied to computerized adaptive testing or multistage testing. Studies have shown that differential speededness of test forms is an even greater challenge in these settings (van der Linden & Xiong, 2013). Investigating, whether the 3PLN model could contribute to the fairness of these assessments, seems worthwhile as well.

3 Controlling the Speededness of Assembled Test Forms: A Generalization to the Three-Parameter Lognormal Response Time Model

Published as: Becker, B., Weirich, S., Goldhammer, F., & Debeer, D. (in press). Controlling the Speededness of Assembled Test Forms: A Generalization to the Three-Parameter Lognormal Response Time Model. *Journal of Educational Measurement.* https://doi.org/10.1111/jedm.12364

Abstract: When designing or modifying a test, an important challenge is controlling the speededness of a test. In 2011, van der Linden (2011a, 2011b) proposed using a lognormal response time model, more specifically the two-parameter lognormal model, and automated test assembly (ATA) via mixed integer linear programming to achieve this. However, this approach has a severe limitation, in that the two-parameter lognormal model lacks a slope parameter. This means that the model assumes that all items are equally speed sensitive. From a conceptual perspective, this assumption seems very restrictive. Furthermore, various other empirical studies and new data analyses performed by us show that this assumption almost never holds in practice.

To overcome this shortcoming, we bring together the already frequently used three-parameter lognormal model for response time, which contains a slope parameter, and the ATA approach for controlling speededness by van der Linden. Using multiple empirically-based illustrations the proposed extension is illustrated, including complete and documented R code. Both the original van der Linden approach and our newly proposed approach are available to practitioners in the freely available R package `eatATA`.

3.1 Theoretical Background

The term *test speededness* refers to test-takers running out of time on a test. There exist a variety of definitions of the term speededness (for comprehensive overviews, see Cintron, 2021; Jurich, 2020). In this paper we follow van der Linden (2011b), who defines speededness as the interaction of the workload of a test, the working speed of a test-taker and the time limit of a test. If a test is speeded for a test-taker, this means that the test-taker cannot answer to all presented items or that the test-taker has to increase their working speed, thereby decreasing their displayed accuracy. The latter process is a well-researched phenomenon called the *speed-accuracy trade-off* (e.g., Goldhammer, 2015).

For over 90 years, speededness has now been critically discussed in the psychometric literature. Already Spearman (1927) noted that to measure ability, tests must be used "[...] in which ample time is allowed, so that speed has little or no scope." In so called *power tests*, the speed of a test-taker is seen as a nuisance parameter (Goldhammer, 2015; van der Linden, 2017). In such a context, speededness of a test is seen as a threat to its validity because test-takers do not have sufficient time to show their true ability (Y. Lu & Sireci, 2007). Furthermore, speededness can be seen as a major threat to the fairness of assessments, since research has shown that speededness can affect different subgroups, such as ethnic groups or gender groups, differently (e.g., Evans & Reilly, 1972; Steinmayr & Spinath, 2019). In other assessment contexts, the speed of the test-taker might be seen as a substantial part of the construct being measured, for example in assessments which asses prioritization skills of test-takers (Kane, 2020) or the efficiency of reading component skills (Goldhammer et al., 2021). In such instances, the speededness of a test should be deliberately chosen to prevent construct-underrepresentation or construct-overrepresentation. In his work, van der Linden (2017) refers to test-taker speed in such assessments as an *intentional parameter*.

In practice, controlling the speededness of an assessment can be relevant for various reasons (van der Linden, 2011b). For instance, practitioners may want to create a test with a predetermined level of speededness or try to determine the appropriate testing accommodations for test-takers with special educational needs. Another common application emerges through the need for multiple but parallel test forms. In high-stakes tests, such as college administration tests, multiple forms of the same test are used to prevent copying answers or sharing test content with future test-takers (College Board, 2015). In low-stakes testing, such as the *Programme of International Student Assessment* (PISA, OECD, 2019a), multiple test forms are used as part of multiple matrix sampling designs to enable the use of large

item numbers while keeping the workload for individual test-takers manageable (Gonzalez & Rutkowski, 2010). Often, an important requirement for parallel test forms, regardless of the specific context, is that all test forms should be equally speeded.

Numerous studies exist which focus on the detection of speededness (for excellent extensive overviews see, Cintron, 2021; Jurich, 2020) and on how to overcome problematic consequences of speededness, such as bias in item parameter estimates (e.g., Bolt et al., 2002; Jin & Wang, 2014; Meyer, 2010; Oshima, 1994; Wollack et al., 2003)[15]. However, to our knowledge there exists little research that is concerned with how speededness can be controlled in advance when designing a test. We argue that controlling the speededness of a test beforehand is at least as important as dealing with the consequences of speededness during the analyses of an assessment. As already Light et al. (1990) noted: "You can't fix by analysis what you bungled by design." While Light et al. (1990) refer to the design of research studies in their work, we think that this sentiment is equally true for the design and assembly of tests.

One of the few approaches for controlling speededness during test assembly has been presented by van der Linden (2011b), who argues that the speededness of a test can be deliberately set for a specific speed level via changing either the time limit or the workload of a test. Based on this, he proposes a test design approach using a lognormal response time model in combination with mixed integer linear programming for automated test assembly. However, this approach is currently limited to a restricted, two-parameter lognormal response time model without a slope parameter. The two-parameter lognormal response time model makes a strong assumption regarding the speed sensitivity of all items, constraining them to be equal. A straightforward extension of the two-parameter lognormal response time model, namely the three-parameter lognormal response time model, already exists (Fox et al., 2007; Klein Entink, Fox, & van der Linden, 2009; Ranger & Ortner, 2012a). Indeed, empirical evidence suggests that the assumption of the two-parameter lognormal model that all items are equally speed sensitive seems to be unrealistic in practice (Becker, Debeer, Weirich, & Goldhammer, 2021). Furthermore, Becker, Debeer, Weirich, and Goldhammer (2021) have shown that ignoring differing speed sensitivities in test assembly can lead to severe fairness issues. Based on their findings, Becker, Debeer, Weirich, and Goldhammer (2021, p. 420) suggest: "Future research should investigate how the 3PLN model can be used best in automated test assembly and whether a similar reparameterization approach might be feasible." Therefore, in this paper, we generalize the approach of van der Linden (2011b) for

[15]Note that there is a natural overlap between these two aspects, i.e. approaches correcting for speededness have to detect if and for whom an assessment is speeded.

controlling speededness in automated test assembly for the well-established three-parameter lognormal response time model.

In the remainder of this paper, we first give a brief overview on automated test assembly via mixed integer linear programming, followed by a short summary of the approach by van der Linden (2011b). We then discuss the limitation of the two-parameter lognormal response time model. Based on this, we present our new contribution, how the approach of van der Linden (2011b) can be generalized to incorporate the three-parameter lognormal response time model within ATA to control the speededness of a test. The usefulness of our proposed approach is illustrated using various exemplary use cases based on the illustrative examples in the work of van der Linden (2011b).

3.2 ATA via MILP

Automated test assembly (ATA) refers to the use of computer algorithms to create test forms with specific test specifications (van der Linden, 2005). Oftentimes, *mixed integer linear programming* (MILP) is used for this purpose. Detailed explanations on general concepts can be found in van der Linden (2005), detailed illustrations on how to use MILP for ATA in practice can, for example, be found in Diao and van der Linden (2011) and Becker, Debeer, Sachse, and Weirich (2021). The general idea is to translate test specifications into mathematical constraints. More specifically, linear combinations of item values can be either constrained or optimized. For instance, the number of items can be constrained, while at the same time the *test information function* (TIF; the sum of item information function values) for a specific ability level is maximized.

Note that all item properties which are used in the constrains have to be known at the moment of test assembly and are assumed to be stable within the operational administration of the test. This means that all approaches using response times to control speededness in ATA assume that response times are available from a pilot study conducted under similar conditions and with a comparable sample as the operational test.

3.3 van der Linden Approach

Before the work of van der Linden (2011b), in order to constrain the speededness of a test, typically constraints were formulated with respect to the observed average test time (a linear combination of the observed average item response times; Cintron, 2021; van der Linden, 2005). However, when constraining the average test time, implicitly speededness is viewed

as a test property, neglecting that test-takers usually differ in their working speeds when responding to a test (van der Linden, 2011b). In addition, tests with equal average test times can still have large differences in the distribution of the test times. Therefore, van der Linden (2011b) proposed an MILP based approach that goes beyond mere average response times, consisting of the following steps: (1) assuming and estimating a lognormal response time model for the item response times (van der Linden, 2006), (2) computing the cumulants of the lognormal response time distributions (i.e., mean and variance) on item level, (3) approximating the total test time distribution based on two cumulants of the total test time distribution, because the sum of the item-wise cumulants are equal to the cumulants of the total test time distribution, and (4) constraining the sum of these cumulants in the MILP model. Using this approach, the total test time distribution can be constrained, and as such, the degree of speededness of a test can explicitly be controlled for different speed levels. In the following, we will illustrate and discuss the different aspects of this method in detail.

3.3.1 Lognormal Response Time Modeling

In response time modeling, using lognormal distributions and more specifically, using the lognormal response time model (e.g., Fox et al., 2007; Klein Entink, Fox, & van der Linden, 2009; Ranger & Ortner, 2012a; van der Linden, 2006) is a popular choice (De Boeck & Jeon, 2019). A commonly used measurement model for lognormal response times $\ln RT_{ik}$, denoting items as $k = 1, ..., j$ and persons as $i = 1, ..., n$, can be written as

$$\ln RT_{ik} = \lambda_k - \phi_k \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{37}$$

In Equation 37, λ_k is typically referred to as the *item time intensity* parameter, ϕ_k is referred to as a *speed sensitivity parameter*, $\sigma^2_{\epsilon_k}$ can be seen as the *item specific residual variance*, and the person parameter ζ_i is interpreted as the test taker's *speed*. Because the model contains three item parameters, we refer to this model as the *three-parameter lognormal* (3PLN) model. The 3PLN model corresponds to a one-factor model from confirmatory factor analysis with freely estimated intercepts (item time intensities), factor loadings (speed sensitivities) and item specific residual variances[16].

A common simplification of the 3PLN model is achieved by fixing all speed sensitivities ϕ to one (van der Linden, 2006). The resulting measurement model can then be written as

[16]For analogies and translations between lognormal response time modeling and linear factor modeling see also the work of Molenaar, Tuerlinckx, and van der Maas (2015b).

$$\ln RT_{ik} = \lambda_k - \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{38}$$

Because the resulting model contains two item parameters, we refer to this model as the *two-parameter lognormal* (2PLN) model. The 2PLN model corresponds to a one-factor model from confirmatory factor analysis with freely estimated intercept parameters (item time intensities) and item specific residual variances but factor loadings (speed sensitivities) fixed to one. For detailed descriptions of the properties and conceptual meanings of the item parameters we refer readers to the work of Becker, Debeer, Weirich, and Goldhammer (2021). It should be noted that, as the lognormal response time model is basically a one-factor model, both the 3PLN and 2PLN model can easily be estimated by any CFA software such as `lavaan` or `Mplus` (e.g., Rosseel, 2012).

3.3.2 Cumulants of the 2PLN Model

In his work on controlling speededness during test assembly, van der Linden (2011a, 2011b) made use of the restricted lognormal response time model, the 2PLN model. As van der Linden (2011a) showed, according to Equation 38 the first two cumulants of the response time distribution for a specific speed level ζ and item k are:

$$\mathrm{E}(RT_k|\zeta) = \exp\left(\lambda_k - \zeta + \frac{\sigma^2_{\epsilon_k}}{2}\right) \tag{39}$$

$$\mathrm{Var}(RT_k|\zeta) = \exp\left(2\lambda_k - 2\zeta + \sigma^2_{\epsilon_k}\right)\left(\exp(\sigma^2_{\epsilon_k}) - 1\right) \tag{40}$$

Equation 39 gives the first cumulant of the response time distribution, which is also the mean, and first moment about the origin. Equation 40 gives the second cumulant, which is also the variance and the second central moment.

Like the assumed lognormal response time distributions, the cumulants in Equations 39 and 40 depend on speed parameter ζ. However, because the ζ terms are constant in both equations, they could be factored out for all items, regardless of the respective item parameters (van der Linden, 2011a). This also implies that when two items have equal expected response times for a specific speed parameter ζ (i.e., equal first cumulants), they have equal expected response times for all possible person speed levels. This also holds for the second cumulant (i.e., the variance).

3.3.3 Cumulants of the Test Time Distribution

A central assumption of the lognormal response time model is local stochastic independence of response times given the latent measurement model. This means that for each test taker the residual response times ϵ_k are independent of each other. It is well known that the combination (i.e., the convolution) of independent normal distributions is also a normally distributed variable. The convolution of independent lognormal distributions is, however, intractable (van der Linden, 2011b). Therefore, it is typically impossible to analytically formulate the total test time distribution for a specific test taker, even when the test is a combination of items following the lognormal model with known item parameters. However, for convolutions of independent random variables, it has been established that the cumulants of the convolution are the simple sum of the cumulants of the independent random variables (Kotz et al., 2005). Consequently, the sum of the first cumulants (Equation 39) of the response time distributions of items within a test is equal to the first cumulant of the test time distribution, the sum of the second cumulants (Equation 40) of the item response times is equal to the second cumulant of the test time distribution, and so on. Formally, for the total number of j items this gives:

$$\mathrm{E}(RT_{tot}|\zeta) = \sum_{k=1}^{j} \mathrm{E}(RT_k|\zeta) \tag{41}$$

$$\mathrm{Var}(RT_{tot}|\zeta) = \sum_{k=1}^{j} \mathrm{Var}(RT_k|\zeta) \tag{42}$$

As $\mathrm{E}(RT_k|\zeta)$ and $\mathrm{Var}(RT_k|\zeta)$ are conditional on a specific speed level ζ, so are $\mathrm{E}(RT_{tot}|\zeta)$ and $\mathrm{Var}(RT_{tot}|\zeta)$. As the cumulants of the total test time distribution are linear combinations of the cumulants of the item response time distributions, the cumulants of the total test time distribution can be constrained in MILP models to perform ATA. For instance, upper and lower bounds for the cumulants can be formulated. For the sum of the first two cumulants of the total test time distribution, this can be formulated as:

$$\sum_{k=1}^{j} \mathrm{E}(RT_k|\zeta) <= T_{mean} + \delta_{mean} \tag{43}$$

$$\sum_{k=1}^{j} \mathrm{E}(RT_k|\zeta) >= T_{mean} - \delta_{mean} \tag{44}$$

$$\sum_{k=1}^{j} \text{Var}(RT_k|\zeta) <= T_{var} + \delta_{var} \tag{45}$$

$$\sum_{k=1}^{j} \text{Var}(RT_k|\zeta) >= T_{var} - \delta_{var} \tag{46}$$

T_{mean} and T_{var} are the respective target values for the mean and variance of the total test response time distribution. δ_{mean} and δ_{var} are the respective tolerance values, denoting how far a test form can deviate from the given target value. Equations 43 and 44 denote the upper and lower bound for the mean of the total test time distribution and Equations 45 and 46 denote the upper and lower bound for the variance of the total test time distribution. If these constraints are implemented within the MILP framework, only solutions will be considered that satisfy the test specifications regarding speededness.

Following the work of Fenton (1960) and Kotz et al. (2005), van der Linden (2011a) delineated that the unknown shape of the test time distribution can be sufficiently approximated using a lognormal distribution via its respective cumulants. Furthermore, he illustrated in an empirical example that constraining only the first two cumulants from Equation 39 and 40 is often sufficient for controlling the speededness of a test in practice. In his work, van der Linden (2011b) demonstrated that the presented approach is indeed very useful for controlling the speededness of a test. He illustrated that the approach can be used to create equally speeded test forms, deliberately change the level of speededness of a test, or to change other test properties while keeping speededness constant.

It should be noted that the procedure presented above slightly deviates from the work of van der Linden (2011b), who defines new parameters (called q_k and r_k) to factor out speed level ζ and uses these parameters for formulating constraints in ATA. This approach is sensible, for example, if test forms are assembled parallel to an existing test form. However, frequently test specifications are specific to a certain speed level, as, for example, test administrators want slow test-takers to have a specific probability for finishing the test in time. In such instances, constraining the cumulants of the total test time distribution directly is a more direct and intuitive approach for test administrators for controlling the speededness of a test. From a practical stand point, both procedures are interchangeable and lead to identical results.

3.3.4 Alternative Approaches for Controlling Speededness

To this date, there exist very few studies that build up on or extend these ideas of van der Linden (2011b). A recent review by Jurich (2020) on test speededness does not include any publications published after the work of van der Linden (2011b). Most of the existing studies focus on computerized adaptive testing instead of fixed form linear tests (e.g., Cheng et al., 2017; Fan et al., 2012; Finkelman et al., 2014; van der Linden & Xiong, 2013; Veldkamp, 2016). Finkelman et al. (2020) extended the approach to linear tests which use cognitive diagnostic modeling (CDMs). To our knowledge, only a single study makes use of the more general 3PLN model, namely the study by Veldkamp et al. (2017). However, Veldkamp et al. (2017) focus on mixture models and only use the first cumulant of the total test time distribution, instead of incorporating the variance of the response time distribution as well. In a recent review on measuring speededness, Cintron (2021) also explicitly notes the lack of approaches for measuring and controlling speededness, which stands in stark contrast to a great variety of response time models developed in the recent years.

3.4 Limitation of the van der Linden Approach

The approach proposed by van der Linden (2011b) has great advantages over approaches using observed average response times. However, the approach still has a major shortcoming: It is currently restricted to the 2PLN model and has not been generalized to the 3PLN model. In contrast to the 3PLN model, the 2PLN model assumes that items do not differ in the extent their response time distribution is sensitive to speed differences across test-takers on top of time intensity effects. From a practical perspective, this is an important limitation, as this assumption of the 2PLN model does seem to rarely hold in practice. In fact, in each of the empirical applications (which we are aware of), in which the 2PLN model has been compared to the 3PLN model, the 3PLN model has shown superior model fit (Becker, Debeer, Weirich, & Goldhammer, 2021; Debelak et al., 2014; Goldhammer & Klein Entink, 2011; Scherer et al., 2015) based on the DIC (Spiegelhalter et al., 2002).

Therefore, in this paper, we generalize the approach by van der Linden (2011b) to the 3PLN model. In the following, we reformulate Equations 39 and 40 to include the additional item parameter and elaborate on practical consequences of the generalization. Then, we illustrate how the new approach can be used in ATA practice.

3.5 Generalization to the 3PLN model

3.5.1 Cumulants of the 3PLN model

If items are no longer assumed to be equally speed sensitive, the derivation of the cumulants of the response time distribution has to be adapted. As a generalization of the work of van der Linden (2011a, 2011b), we delineate the following equations for the cumulants of the 3PLN response time model:

$$\mathrm{E}(RT_k|\zeta) = \exp\left(\lambda_k - \phi_k\zeta + \frac{\sigma^2_{\epsilon_k}}{2}\right) \tag{47}$$

$$\mathrm{Var}(RT_k|\zeta) = \exp\left(2\lambda_k - 2\phi_k\zeta + \sigma^2_{\epsilon_k}\right)\left(\exp(\sigma^2_{\epsilon_k}) - 1\right) \tag{48}$$

Note that, in contrast to Equations 39 and 40, where ζ can be factored out, this is not the case in Equations 47 and 48, where ζ is weighted by the speed sensitivity parameter ϕ_k. This leads to a procedural advantage of the 2PLN model over the 3PLN model: As mentioned above, if we assume that all items are equally speed sensitive and if speededness is controlled for one speed level ζ, speededness is equally controlled for all other potential speed levels. This is not the case if items have different speed sensitivities. Under such circumstances, items may lead to similar response time distributions for one speed level, but differing response time distributions for other speed levels (Becker, Debeer, Weirich, & Goldhammer, 2021). This implies that separate constraints have to be introduced if a test is supposed to be parallel for multiple speed levels.

It should be noted, as the 2PLN model is a special case of the 3PLN model, the cumulants calculation for the 2PLN model is also a special case of the cumulant calculation for the 3PLN model. If the speed sensitivity is fixed to $\phi_k = 1$, Equations 47 and 48 simplify to Equations 39 and 40. A special case occurs if speed level $\zeta = 0$ is used, in which case Equations 47 and 48 yield identical results as Equations 39 and 40.

3.5.2 Cumulants of the Test Time Distribution

As in the 2PLN model, responses are assumed to be independent given the latent factor structure in the 3PLN model. Therefore, the sum across the first cumulants of the item response time distributions in a test still equals the first cumulant of the test response time (this is true for all cumulants) (e.g., Fenton, 1960; Kotz et al., 2005). Hence, Equations 41 and 42 hold for the 3PLN model as well. Furthermore, the constraints suitable for implementing

test time specifications according to the 3PLN model are identical to Equations 43 - 46.

However, if speededness should be controlled for multiple speed levels, constraints must be formulated for multiple speed levels. This requirement is plausible, for example, if an assessment makes use of multiple, parallel test forms, which are randomly assigned to test-takers. This means that Equation 47 and 48 must be applied for multiple ζ values. In practice, this means that test developers are calculating the expected response time mean and variance for different speed levels. Furthermore, constraints according to Equations 43 - 46 must be formulated for all of these speed levels.

3.5.3 Computational Implementation

In the past, implementing ATA via MILP procedures in practice was not trivial. For example, Diao and van der Linden (2011) suggest using a lpSolve API directly via R for ATA. To lower the learning curve for test designers and practitioners we have developed the `eatATA` R package, which provides a more intuitive user interface. Furthermore, extensive resources for how to implement ATA problems in `eatATA` exist (Becker, Debeer, Sachse, & Weirich, 2021). `eatATA` currently provides access to various free and commercial solvers, namely GLPK, lpSolve, Symphony, and Gurobi. We have added both our newly developed approach, as well as the original approach by van der Linden (2011b) to the `eatATA` package for calculating the mean and variance of the expected response time distribution. Equation 39 and 40 are implemented in the `getMean2PLN()` and `getVar2PLN()` functions, and Equations 47 and 48 are implemented in the `getMean3PLN()` and `getVar3PLN()` functions. It should be noted again that the approach by van der Linden (2011b) is a special case of our newly developed approach with all $\phi_k = 1$. As `eatATA` is a general tool for ATA, the specific constraints as suggested in Equation 43 - 46 can be implemented like any other quantitative constraints, for instance using the `itemValuesDeviationConstraint()` function. Figure 8 illustrates how our newly proposed approach would be implemented to control speededness for multiple test forms (`n_forms`) for two separate speed levels, a slow ($\zeta = -1$) and a fast one ($\zeta = 1$). In the example code, test forms are constrained to arbitrary target values and the deviation from the target value is set relative to the respective target values, resulting in $\delta_{mean} = 0.1T_{mean}$ and $\delta_{var} = 0.1T_{var}$. Constraints are formulated for an item pool contained within a `data.frame` called `items`.

To illustrate the feasibility and effectiveness of the proposed approach for controlling speededness we present three use cases, which correspond to the use cases presented in van

Figure 8: *Example Code for Implementing our Proposed Approach for Controlling Speededness via the 3PLN Model.*

```
# Computing the cumulants
means_3PLN <- getMean3PLN(lambda = items$lambda, phi = items$phi,
                          zeta = c(1, -1), sdEpsi = items$sdEpsi)
vars_3PLN <- getVar3PLN(lambda = items$lambda, phi = items$phi,
                        zeta = c(1, -1), sdEpsi = items$sdEpsi)

# constraints for zeta = -1
constr_mean_slow <- itemValuesDeviationConstraint(nForms = n_forms,
                          itemValues = means_3PLN[, "zeta=-1"],
                          targetValue = 300, allowedDeviation = 0.1,
                          itemIDs = items$ID, relative = TRUE)
constr_var_slow <- itemValuesDeviationConstraint(nForms = n_forms,
                          itemValues = vars_3PLN[, "zeta=-1"],
                          targetValue = 2000, allowedDeviation = 0.1,
                          itemIDs = items$ID, relative = TRUE)

# constraints for zeta = 1
constr_mean_fast <- itemValuesDeviationConstraint(nForms = n_forms,
                          itemValues = means_3PLN[, "zeta=1"],
                          targetValue = 30, allowedDeviation = 0.1,
                          itemIDs = items$ID, relative = TRUE)
constr_var_fast <- itemValuesDeviationConstraint(nForms = n_forms,
                          itemValues = vars_3PLN[, "zeta=1"],
                          targetValue = 200, allowedDeviation = 0.1,
                          itemIDs = items$ID, relative = TRUE)
```

der Linden (2011b): (1) Creating parallel test forms, (2) modifying a test form while keeping its speededness constant, and (3) changing the level of speededness of a test form, all based on the same simulated item pool.

3.6 Illustrative Examples

3.6.1 Item Pool

In his study, van der Linden (2011b) utilized an empirical item pool of the *Law School Admission Test* (LSAT) with simulated response time model parameters, with a total size of 756 items. An operational test form containing 78 items was initially assembled from the item pool. As the LSAT item pool is not publicly available (nor is any other comparable high-stakes assessment item pool), we simulate an item pool of identical size based on data from the Canadian Programme of International Student Assessment (PISA) 2018 math data. The PISA study is an educational large-scale assessment and provides one of the few publicly available data sets including achievement data on item level with raw responses and response times (OECD, 2019a)[17]. The data set is openly available at https://www.oecd.org/pisa/data/2018database/.

To obtain realistic hyperparameters for the illustrative examples and to illustrate that the

[17]While the LSAT and PISA assessments obviously differ in many regards, the PISA data set is only used to generate plausible hyperparameters for the draw of item parameters. Note that both van der Linden (2011b) and we base our illustrative examples on an item pool with simulated response time model parameters.

Table 3: *Model Comparisons between Hierarchical Frameworks using the 2PLN and 3PLN Model as Response Time Measurement Model.*

Booklet	WAIC 2PLN	WAIC 3PLN	$\widehat{\text{elpd}}_{loo}$ 2PLN	$\widehat{\text{elpd}}_{loo}$ 3PLN	$\Delta(\widehat{\text{elpd}}_{loo})$
1	72315.55	71104.90	-36208.81	-35600.97	-607.84 (52.46)
2	73004.52	71157.95	-36579.43	-35647.11	-932.31 (61.16)
3	66316.32	65659.20	-33210.17	-32873.13	-337.04 (37.34)
4	84661.87	83607.63	-42498.40	-41918.61	-579.79 (53.25)
5	77113.19	76306.04	-38708.59	-38281.20	-427.39 (42.25)
6	63972.26	62911.83	-32056.92	-31518.43	-538.49 (51.57)

Note: Model comparisons are conducted using the Widely Applicable Information Criterion (WAIC) and Leave One Out (LOO) cross validation. For LOO, the expected log pointwise predictive density is used ($\widehat{\text{elpd}}_{loo}$) and differences as well as standard errors for the differences are reported.

assumption of the 2PLN model of equal speed sensitivities does usually not hold in practice, a joint hierarchical response and response time model (van der Linden, 2007) is estimated via the general purpose Bayesian estimation software `Stan` (Carpenter et al., 2017) and its R interface `rstan` (Stan Development Team, 2021). For model estimation, we use the default options set by `rstan`, with 4 chains, 2000 iterations each and a burn-in of 1000 iterations. Model specifications adopt the recommendations by König et al. (2023) with hierarchical prior specifications and informative or weakly informative priors aiding estimation stability. For instance, for covariance matrices, a separation strategy is utilized, with Cauchy distributions as a prior for standard deviations and an LKJ prior distribution for the Cholesky factor of the correlation matrix. To investigate whether data simulation should be performed according to the 2PLN or 3PLN model, model comparisons between hierarchical frameworks using the 2PLN and the 3PLN model are conducted. Model comparisons are performed for all Canadian math booklets using the *Widely Applicable Information Criterion* (WAIC) and the *expected log pointwise predictive density of the leave one out cross validation* ($\widehat{\text{elpd}}_{loo}$; Vehtari et al., 2017) provided via the R package `loo` (Vehtari et al., 2019). An overview of the model comparisons for all booklets can be found in Table 3. Smaller values of WAIC and larger values of $\widehat{\text{elpd}}_{loo}$ indicate better model fit. Both WAIC and $\widehat{\text{elpd}}_{loo}$ as well as standard errors for the differences in $\widehat{\text{elpd}}_{loo}$ indicate consistent and substantial better fit for the 3PLN model.

Based on these findings, we draw item parameters and simulate response times according to the 3PLN model for all illustrative examples. The hyperparameter distribution for the item pool is based on the analyses of a randomly selected Canadian PISA 2018 math booklet (booklet 3) and can be seen in Table 4. Based on this hyperparameter distribution, 756 items with item parameters conforming to a 2PL IRT model and 3PLN model are drawn. These

item parameters are treated as known without uncertainty from a pilot study.

From the simulated item pool containing 756 items, a test form is assembled consisting of 78 items, which maximizes the *test information function* (TIF) at ability level $\theta = 0$. The resulting TIF is $TIF_{existing} = 122.49$. We refer to this test form as the *existing test form*. The complete Stan and R code for the estimation of the joint hierarchical model as well as the R code for all illustrative examples can be seen in the Online Supplement available at https://osf.io/ktgrf/. The complete code for the empirical data estimation can be found in the subfolder `empirical_data_estimation`. The simulation of the item pool and assembly of the existing test form can be found in the syntax file `0_item_pool_generation.R`.

Table 4: *Means, Standard Deviations and Correlation Matrix of the Item Parameters in the Hierarchical Estimated Model Using a Canadian PISA 2018 Math Booklet.*

Parameter	M	SD	a	b	ϕ
a	1.23	0.42			
b	-0.28	1.32	0.71		
ϕ	0.35	0.06	0.00	0.44	
λ	4.39	0.23	0.00	0.10	0.61

Note: Item Discrimination (a), Item Difficulty (b), Item Speed Sensitivity (ϕ), and Item Time Intensity (λ).

3.6.2 Illustration 1a: Additional Test Form

A common test assembly use case is to have multiple, parallel test forms. In accordance with van der Linden's first example, we also assemble an additional, parallel test form from the remaining item pool, for instance if the goal was to offer test-takers a second test administration. The newly assembled test form should conform to the following test specifications: (a) it should consist of 78 items, (b) its test information function should be as similar as possible to the existing test form at ability level $\theta = 0$, (c) there should be no item overlap between the existing and new test form, and (d) it should be comparable regarding its speededness, as well. To implement the test specifications, the *item information function* values (IIFs) at $\theta = 0$ are calculated. For obtaining comparable TIFs, a *minimax* objective function is utilized. The number of items in the new test form is constrained to be exactly 78.

To control the speededness of the newly assembled test form, we utilize our newly developed approach: First, the mean and variance of the item response time distributions are calculated for speed level $\zeta_i = -1$ using Equations 47 and 48. The below average speed level is chosen, as speededness of a test is mostly relevant for slow test-takers, as faster test-takers often have enough time to finish test-forms with differing levels of speededness (Becker, De-

beer, Weirich, & Goldhammer, 2021). The speededness constraints are then implemented using the mean and variance of the total test time distribution of the existing test form as target values. A relative tolerance is implemented, meaning that $\delta_{mean} = 0.0001 T_{mean}$ and $\delta_{var} = 0.0001 T_{var}$[18].

To illustrate that the approach proposed by van der Linden (2011b) is no longer suitable if items follow the 3PLN model and have differing speed sensitivities, we also assemble a parallel test form using van der Linden's approach using the 2PLN model. We therefore calculate the first two cumulants of the item response time distribution using Equations 39 and 40. Note that this means that the respective target values based on the existing test form also vary between the 2PLN and the 3PLN approach.

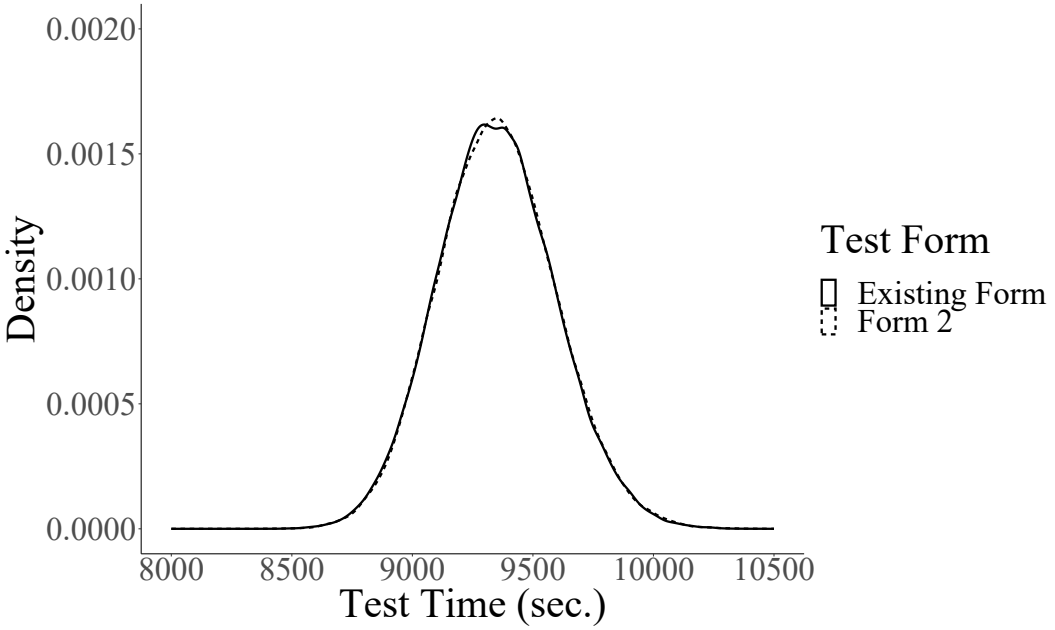
All other aspects of the implementation of speededness constraints remain the same. After assembling the test forms, item response times are simulated according to the 3PLN model for speed level $\zeta = -1$ with $n = 100,000$ and accumulated on test level. The full R code for this illustration can be seen in the syntax named `1_single_parallel_testforms.R`. Implementation of the ATA procedure is done via the `eatATA` R-package using the GLPK solver.

Results. As controlling speededness is the main focus of the illustration, we focus on results regarding test time in this section. Figure 9 depicts the resulting total test time distributions for speed level $\zeta = -1$ for both the existing test form and the newly assembled form using our proposed approach based on the 3PLN model. The almost identical distributions illustrate that the proposed approach is indeed suitable for creating a parallel test form regarding speededness and that the first two cumulants of the response time distribution are indeed sufficient for its approximation.

In contrast, Figure 10 depicts the resulting total test time distributions for both the existing test form and the newly assembled form using the original van der Linden approach based on the 2PLN model (i.e., all speed sensitivities were set to $\phi_k = 1$). The distributions are clearly distinct, which illustrates that this approach is no longer suitable, if the 3PLN model was used as the response time model for estimation. This is also reflected in the means of the total test time distributions, which differ 209.03 seconds ($\mathrm{E}(RT_k)_{existing} = 9351.59$ vs. $\mathrm{E}(RT_k)_{2PLN} = 9560.62$). Note that this is expected, as the cumulants of the item response time distributions and therefore also the target values are calculated incorrectly according to the 2PLN model, even though the data (and our response time simulation) follow the

[18] In contrast, van der Linden (2011b) uses an absolute tolerance of 0.01.

Figure 9: *Total Test Time Distributions for Speed Level $\zeta = -1$ for Two Test Forms Assembled According to Our Proposed Approach.*



3PLN model. One could therefore argue that this difference between test forms reflects the under-parameterization of the response time modeling according to the 2PLN model.

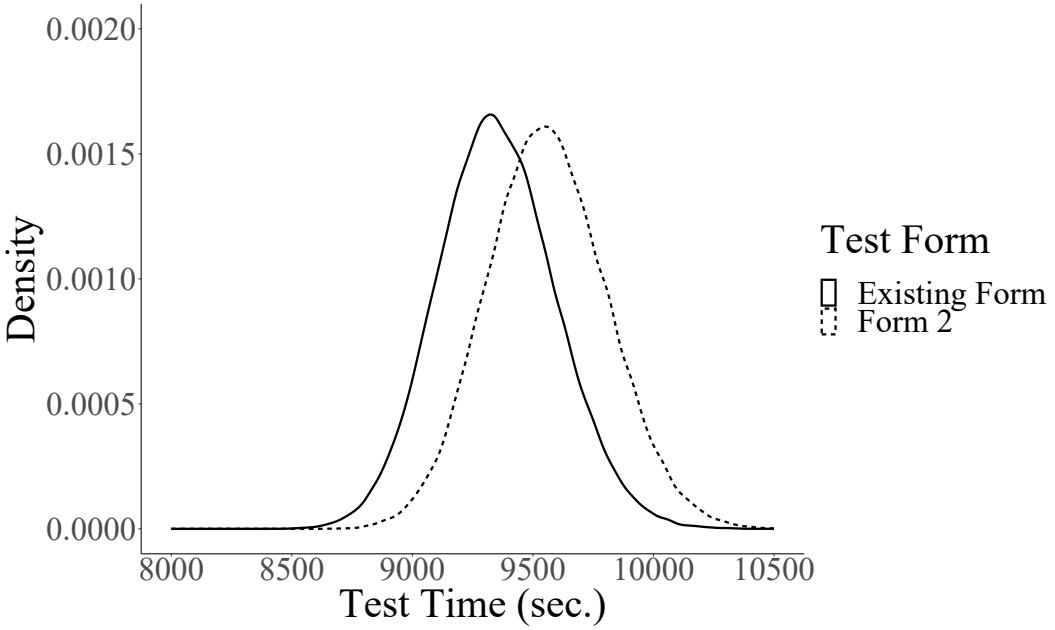
Regarding optimization of the TIF it should be noted that by the initial maximization for the existing test form depletes the resulting item pool of the most informative items. Therefore, any consecutively assembled test forms do not have access to the most informative items in the original item pool, with the maximum available TIF for a 78-item test form being $TIF = 81.10$. For both approaches, minimizing the difference of the TIF in comparison to the existing test form was comparable. The TIFs for the newly generated test forms were $TIF_{2PLN} = 80.78$ and $TIF_{3PLN} = 80.37$ respectively.

3.6.3 Illustration 1b: Additional Test Form with Multiple Speed Levels

A convenient property of the approach presented by van der Linden (2011b) is that controlling speededness is independent of the respective speed levels. As there are no speed sensitivities (i.e., all speed sensitivity parameters are fixed to one), ζ can simply be factored out of Equations 39 and 40. For the less restrictive 3PLN model including freely estimated speed sensitivities, test forms might be parallel for a specific speed level, but might not be parallel for a different speed level. If equal speededness for multiple speed levels is desired, this can be implemented via additional sets of constraints.

To illustrate this, we use essentially the same assembly problem as in Illustration 1a but

Figure 10: *Total Test Time Distributions for Speed Level $\zeta = -1$ for Two Test Forms Assembled According to van der Linden's Approach.*



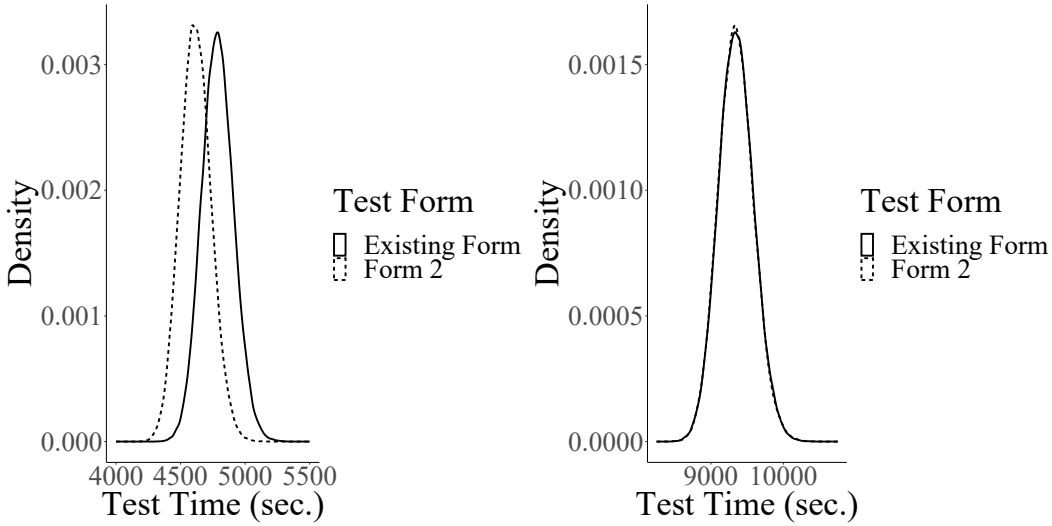
focus on multiple speed levels for controlling speededness. The resulting test specifications are: (a) the test form should contain exactly 78 items, (b) TIF should be as similar as possible to the existing test form, (c) there should be no item overlap between the existing and new test form, and (d) speededness should be identical to the existing test form both for fast and slow test-takers. In a naive approach, we utilize the same constraints as formulated in Illustration 1a, thereby ignoring all speed levels other than $\zeta = -1$. Additionally, we implement an assembly approach which controls for speededness at speed levels $\zeta = [1, -1]$. Constraints for speed level $\zeta = 1$ are formulated as they are formulated for speed level $\zeta = -1$, comparable to the illustration in Figure 8.

Consequently, we now inspect the consequences for two different speed levels $\zeta = [1, -1]$ as well, so for a slow and a fast speed level. We leave the relative tolerance at 0.0001 for the mean and increase it to 0.01 for the variance of the total test time distribution due to initial infeasibility issues given the item pool. After assembling the test forms, item response times are simulated according to the 3PLN model for the two speed levels $\zeta = [1, -1]$ with $n = 100,000$ and accumulated on test level. The full R code for this illustration can be seen in the syntax `1b_single_parallel_testforms_multi_zetas.R` in the online supplement.

Results. Figure 11 shows the resulting test time distributions if only speededness for one speed level is controlled. On the right side we see a similar picture as in Figure 9. We

controlled speededness for speed level $\zeta = -1$, therefore the response times distributions for this speed level overlap almost perfectly. However, the left graph depicting the response time distribution for speed level $\zeta = 1$ illustrates what we mentioned before: If test forms are parallel regarding speededness for one speed level, this is not necessarily the case for all other speed levels. Parallelism of the test time distributions for other speed levels will depend on whether test time distributions for similar speed levels have been controlled for. Otherwise, parallelism for speed levels will be rather random. Indeed, in our examples the two distributions are clearly shifted.

Figure 11: *Total Test Time Distributions for Two Speed Levels for Two Test Forms Assembled While Constraining Time Required for One Speed Level ($\zeta = -1$).*



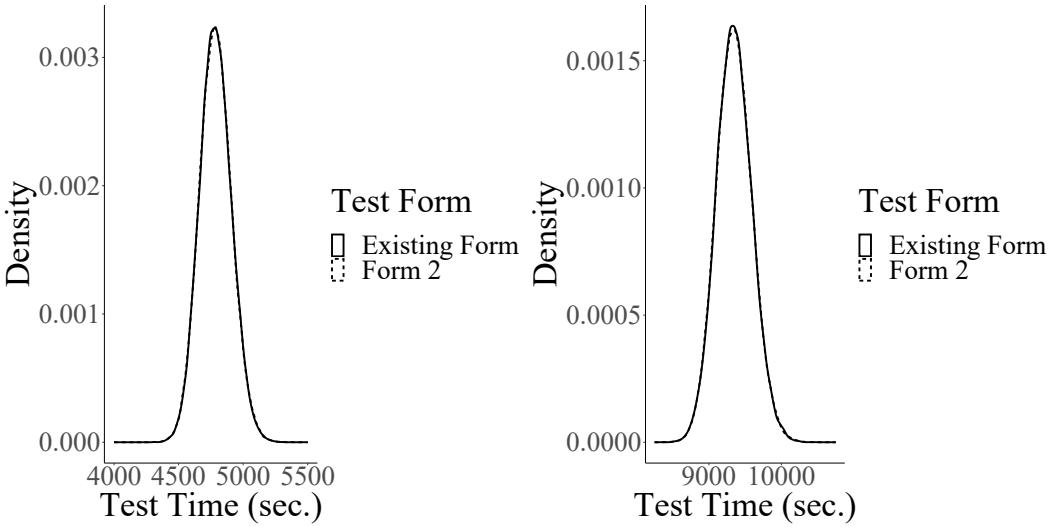
Note. The left plot contains the distributions for the fast speed level $\zeta = 1$, the right plot contains the distributions for the slow speed level $\zeta = -1$, for which the speededness was constrained.

In contrast, Figure 12 shows the resulting test time distributions if time required for the two speed levels $\zeta = [1, -1]$ is controlled. Both response time distributions overlap almost perfectly. For both approaches, minimizing the difference of the TIF in comparison to the existing test form was satisfactory. The TIFs for the newly generated test forms were $TIF = 77.61$ for the test form with controlled speededness for two speed levels and $TIF = 80.51$ for the test form with controlled speededness for one speed level.

3.6.4 Illustration 2: Changed Test Form, Identical Speededness

Another common use case for controlling the speededness of a test form is when specific properties of a test should be changed, while others should remain constant. For example, it could be desirable to have a shorter test form which is especially informative for high-ability

Figure 12: *Total Test Time Distributions for Two Speed Levels for Two Test Forms Assembled While Constraining Time Required for Two Speed Levels ($\zeta = [-1; 1]$).*



Note. The left plot contains the distributions for the fast speed level $\zeta = 1$, the right plot contains the distributions for the slow speed level $\zeta = -1$.

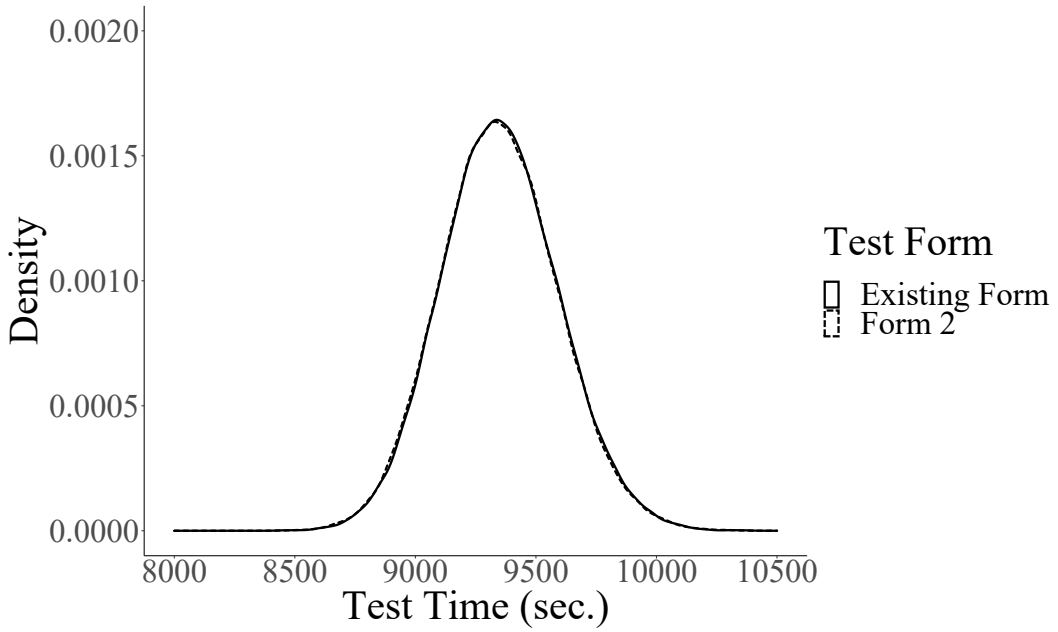
test-takers. Therefore, similar to the second example in van der Linden (2011b), we modify the existing test form by shifting the target value of θ, at which the TIF should be maximized, by 1.2. Furthermore, the test form is shortened from 78 to 69 items as in van der Linden (2011b). As this use case is not about creating an additional test form but modifying an existing test form, item overlap to the existing test form is not prohibited. The resulting test specifications are: (a) the test form should contain exactly 69 items, (b) TIF should be maximized at $\theta = 1.2$, and (c) speededness should be identical to the existing test form for slow test-takers.

For this purpose, speededness is constrained to be the same between the old and new test form via the proposed approach. This means that comparable to Illustration 1a, the target mean and variance of the response time distribution are calculated using the `getMean3PLN()` and `getVar3PLN()` functions for speed level $\zeta = -1$. The sum of these item means and variances are then constrained and the same tolerance is used as in Illustrations 1a and 1b. The full R code for this illustration can be seen in the online supplement syntax `2_changed_test_form.R`.

Results. The modified test form is indeed substantially more informative for ability level $\theta = 1.2$ (original test form: $TIF_{\theta=1.2} = 28.81$; newly assembled test form: $TIF_{\theta=1.2} = 116.73$). Meanwhile, the total test time distribution is held constant with the specified 3PLN

constraints, as Figure 13 illustrates.

Figure 13: *Total Test Time Distributions for Speed Level $\zeta = -1$ for an Initial Test Form with Maximized TIF at $\theta = 0$ and a Newly Assembled Test Form with Maximized TIF at $\theta = 1.2$ While Holding Speededness Constant for $\zeta = -1$ Using the Proposed 3PLN Approach.*



3.6.5 Illustration 3: Changed Speededness

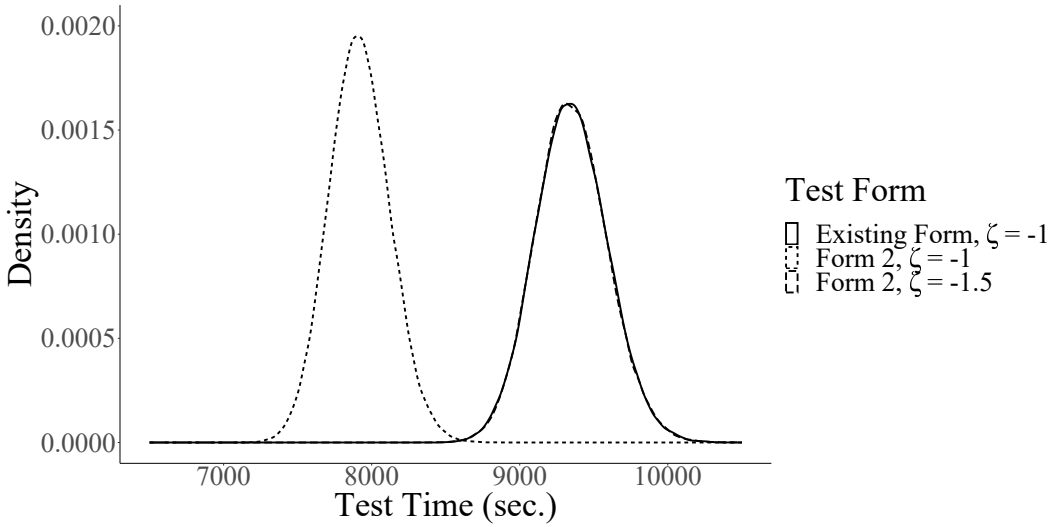
Another scenario described by van der Linden (2011b) is that properties of a test form should all remain constant (including the number of items in the test form) while its speededness is modified. For example, test-takers might have complained about excessive levels of speededness and we want to reduce the amount of time pressure of a test. This use case is in line with the third example in van der Linden (2011b). To reduce the level of speededness while leaving the time limit untouched, we shift the target speed by $\Delta = -0.5$, meaning the test time distribution of the new test form at $\zeta = -1.5$ should be the same as the test time distribution of the existing test form at $\zeta = -1$ [19]. This means that test-takers have now to cope with less workload than in the initial test form even though the number of items in the test form remains the same. Practically speaking, this means that less time intensive items have to be selected for the new test form. The resulting test specifications are: (a) the test form should contain exactly 78 items, (b) TIFs should be as similar as possible to the existing test form, and (c) speededness should be shifted as described. The full R code for this illustration can be seen in the online supplement syntax `3_changed_speededness.R`.

[19] This is a smaller shift as in van der Linden (2011b) ($\Delta = 1$), but the idea remains the same.

Results. Figure 14 shows three test time distributions: the test time distribution of the original test form at speed level $\zeta = -1$ and the test time distributions of the newly assembled test form at speed levels $\zeta = -1$ and $\zeta = -1.5$. And indeed, as Figure 14 indicates, the 3PLN approach is also suitable for modifying the level of speededness of an existing test form. The test time distributions of the initial test form at $\zeta = -1$ and of the modified test form at $\zeta = -1.5$ overlap almost perfectly. This means that very slow test-takers ($\zeta = -1.5$) now need as much time on the newly assembled test form as slow test-takers ($\zeta = -1$) initially needed on the existing test form. Moreover, the distribution on the left indicates the test time distribution of the modified test form for test-takers at $\zeta = -1$. This illustrates that slow test-takers ($\zeta = -1$) now require less time compared to the initial test form.

Test information function changes from the existing test form from $TIF_{existing} = 122.49$ to $TIF_{new} = 110.67$.

Figure 14: *Total Test Time Distributions for the Initial Test Form for Speed Level $\zeta_i = -1$ and the Newly Assembled Test Form with Shifted Speededness for Speed Levels $\zeta = -1$ and $\zeta = -1.5$.*



3.7 Discussion

Controlling the degree of speededness of a test is a frequent challenge in educational and psychological testing. In 2011, van der Linden (2011b) proposed an innovative approach for controlling the speededness of a test, emphasizing that speededness is defined by the interaction of the workload of a test, the time limit which is set and the working speed of a test-taker. However, since then, almost no further developments have been made in the area of controlling speededness for fixed form tests. To this date, the approach by van

der Linden (2011b) is limited to the 2PLN model as a response time measurement model. Meanwhile, empirical evidence has accumulated that the 2PLN model is probably too strict in most applications and instead the 3PLN model should be preferred. Therefore, we have generalized the framework of van der Linden (2011b) to the 3PLN model and illustrated its usefulness in various use cases.

It should be emphasized that our proposed approach for using the 3PLN model to control speededness is in no way more difficult or demanding for practitioners wanting to control speededness in their test(s) than the approach by van der Linden using the 2PLN model. Estimation of the 3PLN model, just as the 2PLN model, can be implemented using standard statistical software for confirmatory factor analysis models, like the R-package `lavaan` (Rosseel, 2012). Hierarchical implementation, which can aid measurement on the response and response time side (van der Linden, 2007; van der Linden et al., 2010), can be implemented using, for example, the R-package `LNIRT` (Fox et al., 2021) or general purpose Bayesian estimation software such as Stan (Carpenter et al., 2017). For an extensive tutorial on Bayesian hierarchical response time modeling using Stan, see König et al. (2023). Alternatively, the hierarchical framework can also be translated to a frequentist factor analysis framework, as illustrated by Molenaar, Tuerlinckx, and van der Maas (2015b). Furthermore, Liu et al. (2022) have proposed using machine learning approaches for model estimation in the IRT framework. The calculation of the cumulants of the expected response time distribution for both models as well as the full ATA procedure can be performed using the `eatATA` package.

If, however, the 2PLN model is used when the 3PLN model would be appropriate, issues regarding the speededness of an assembled test form can arise at two different occasions: (1) If the 2PLN model is estimated while estimation of the 3PLN model would be appropriate, estimates of model parameters will be biased. (2) If differences in speed sensitivities are ignored, target values and the cumulants of the test time distribution will be calculated incorrectly. Our first use case illustrated, that indeed incorrect calculation of the cumulants of the response time distributions and of the target values leads to issues regarding the speededness of test forms.

Furthermore, this theoretical approach of controlling the speededness of tests is not restricted to the mentioned lognormal response time models. These lognormal models have very convenient properties and are some of the most popular response time models in psychometrics (De Boeck & Jeon, 2019). Nevertheless, if there are other response time models

for which total test time distributions can be calculated or sufficiently approximated based on model parameters, these could be used instead. In addition, the proposed approach and the described implications are not only relevant for fixed-form linear tests but also for controlling speededness in multi-stage testing or computer-adaptive testing. In adaptive testing contexts speededness can be of even greater concern, as the difficulty of items is often correlated with their time intensity (Bridgeman & Cline, 2004; van der Linden, 2009b). Fortunately, our proposed approach is not only feasible for fixed-form linear tests but also for the assembly of modules in multi-stage testing as well as computer-adaptive testing via the shadow-test framework, as well (van der Linden & Xiong, 2013).

3.7.1 Practical Considerations

There are a few practical considerations for test designers and administrators when constraining the speededness of test forms. First, in the 3PLN approach, test administrators have to choose for which speed level(s) speededness should be controlled, similar to how the TIF is set for specific ability levels. Choosing the appropriate speed levels for which to control speededness for depends on the application. If, for example, speed is seen as a nuisance parameter and the test forms should not be speeded for any test-taker, it is sufficient to control speededness for a (very) slow speed level. If speed is seen as an intentional parameter, requirements might vary: If a single test-form is assembled, a single speed level might be used as a target speed level (i.e., as the optimum, desired speed level). If multiple test forms are assembled, the minimum requirement from a fairness perspective is that all test forms have to be parallel for all speed levels who are prone to running out of time.

Second, in our study we have assumed that item parameters are known without uncertainty from a pilot study. However, from a practical perspective, it is not trivial how a pilot study should be designed to accurately estimate unbiased item parameters. On one hand, one could argue that testing conditions for item piloting and the operational testing phase should correspond. This could help to ensure that item properties (i.e., parameters) do not change under different conditions. If, for example, test-takers are less motivated in a low-stakes piloting setting compared to a high-stakes operational setting, item time intensities might get underestimated if test-takers work on items less thoroughly in the piloting phase.

On the other hand, however, one might argue that pilot studies should try to exterminate external influences such as speededness, decreasing effort, or fatigue from item parameter estimation. This is an issue as most analysis models assume that test-takers work with

constant ability and speed throughout a (pilot) test. Research on item position effects has shown that indeed fatigue, decreasing effort, and speededness can have a negative impact on item parameter estimation (Debeer & Janssen, 2013; Oshima, 1994; Weirich et al., 2017). The following measures could help to reduce the influence of fatigue, decreasing effort, and speededness on item parameter estimates: (a) test-takers should be given sufficient time on the pilot test, (b) the pilot test should be sufficiently short so test-takers do not experience fatigue or a decrease in effort, (c) item positions could be balanced in the pilot test design so effects on item properties are averaged across item positions, and (d) models, which take item position effects or varying speed levels into account could be used for the estimation of item parameters (e.g., Fox & Marianti, 2016).

Note, however, that these challenges regarding piloting conditions are neither specific to our proposed approach for controlling speededness nor specific to controlling speededness in general. These challenges arise whenever item properties are estimated or used based on a pilot study. For instance, if test-takers speed up during a pilot test due to time constraints and time intensity is underestimated for items positioned at the end of the pilot test, it is likely that the difficulty of items is overestimated in such cases as well.

A third practical consideration relates to parameter uncertainty. Even if ideal piloting conditions are created, item parameters are still estimated with uncertainty. This uncertainty, however, is rarely reflected in ATA approaches. To counter these issues, Veldkamp et al. (2013) proposed incorporating the uncertainty in item parameter estimation via *robust automated test assembly*. Furthermore, Veldkamp (2016) proposed using robust ATA when minimizing test time in computerized adaptive testing. Future research could look into how robust ATA might be a sensible extension to our proposed approach.

3.7.2 Conclusion

Being able to control the speededness of one or multiple test forms is important for the fairness and validity of assessments. In this manuscript, we have introduced a generalization of the approach for controlling speededness by van der Linden (2011b) which allows for incorporating differing speed sensitivities. We have implemented both our approach and the approach by van der Linden (2011b) in the freely available R package `eatATA`. Furthermore, we have illustrated in multiple practical use cases the feasibility of our approach and how the original approach by van der Linden (2011b) falls short if items indeed vary regarding their speed sensitivity. For assessment practitioners, who want to utilize the presented approaches

for controlling speededness in test assembly, our illustrative examples and online supplement should provide all required tools to do so.

4 Item Order and Speededness: Implications for Test Fairness in Higher Educational High-Stakes Testing

Published as: Becker, B., van Rijn, P., Molenaar, D., & Debeer, D. (2022). Item order and speededness: Implications for test fairness in higher educational high-stakes testing. *Assessment & Evaluation in Higher Education, 47*(7), 1030-1041. https://doi.org/10.1080%2F02602938.2021.1991273

Abstract: A common approach to increase test security in higher educational high-stakes testing is the use of different test forms with identical items but different item orders. The effects of such varied item orders are relatively well studied, but findings have generally been mixed. When multiple test forms with different item orders are used, we argue that the moderating role of speededness on item order effects cannot be neglected as missing responses are commonly scored as incorrect in high-stakes testing. If test-takers run out of time while not giving answers to easy items at the end of the test, they are penalized stronger than if instead they were unable to provide answers to difficult items. Using an illustrative real-data example of a speeded test, we show that the potential consequences of ignoring item order can be substantial with respect to test fairness. Our proposed solution consists of using a fixed item order across forms from the point at which the test may become speeded for some students. In this approach, the most time-intensive items are placed at the end of the test. A simulation based on real data of two university exams from psychology students illustrates the usefulness of this approach.

In higher educational high-stakes tests like college exams, important challenges are test fairness and test security. To assure test fairness, it is popular practice to set a common time limit for all test-takers and to score missing responses as incorrect, to prevent test-takers from choosing a specific set of items to respond to. With regard to test security, a major concern is cheating and more specifically, test-takers copying answers from other test-takers, the most popular cheating practice in crowded class room situations (Chirumamilla et al., 2020). A common approach to prevent this behavior is to create multiple test forms with rearranged item orders and to provide neighboring test-takers with differentially ordered forms (e.g., Monk & Stallings, 1970). This strategy is assumed to limit the probability that neighboring test-takers are simultaneously working on the same items, thereby making answer-copying difficult (Davis, 2017; Vander Schee, 2013). Other methods that prevent answer-copying, like test forms with distinct item sets or computer adaptive testing, exist (van der Linden, 2005) but typically require larger item pools, pretesting items, and/or computer-based test administrations. However, these requirements are often impossible to meet in conventional higher educational testing.

If multiple test forms with different item orderings are used, the resulting test scores should not depend on the ordering of the items. Or, as Lord (1980, p. 195) writes, "[...] it must be a matter of indifference to applicants at every given ability level [...] whether they are to take test x or test y". A test cannot be considered a fair test, if the test score of an individual would be different given an alternative test form.

In this paper, we investigate how different item orderings can affect test performance for test-takers and therefore violate principles of test fairness. First, we give a brief overview of the research on item order effects. Then, we introduce a modeling framework which allows us to jointly model ability and speed. Based on this framework, we introduce the concept of *speededness* and discuss why speededness may have been (partly) overlooked when explaining item order effects. Furthermore, test-takers can act in different manners when facing speededness constraints on a test. Using the concept of test-wiseness, we discuss these differences and explain how test-wiseness influences the relationship between speededness and test fairness. An empirical example of a speeded test is used to demonstrate how different item orderings can lead to unfair test forms. Finally, we propose a simple, heuristic approach to prevent unfair effects of item ordering and illustrate its effectiveness based on a short simulation study.

4.1 Theoretical Background

The question whether item order can be rearranged without affecting the fairness of a test has been extensively discussed in the literature. Leary and Dorans (1985) provide an exhaustive overview of the research before 1985, whereas L. Wang (2019) provides a more recent, but smaller overview. Most studies have focused on whether overall test difficulty varies if items are sorted (a) in random order, (b) Easy-Hard, (c) Hard-Easy, or (d) ordered according to content (*topical ordering*). Note that all specific orderings (such as b-d) can also result from random ordering. Therefore, even if test administrators plan to use random item orderings, they have to ensure that any possible differential impact on test scores due to item order is avoided.

Leary and Dorans (1985) state that sorting by difficulty usually has an effect on test scores if the test is administered under a time limit, with Hard-Easy leading to the lowest scores. They offer the explanation that in the Hard-Easy conditions "[...] when an examination is administered under strict time constraints, some examinees could be at a disadvantage as a result of spending time on hard items early in the test that they could more profitably have spent on easy items near the end." For a similar explanation see also Sax and Cromack (1966), who conclude that "[...] test constructors have a responsibility of arranging items in ascending order of difficulty if tests are lengthy or time limits restricted."[20]

Overall, however, the literature is inconclusive regarding the relation between item orderings and test difficulty: Leary and Dorans (1985) report contradicting findings; a meta-analysis by Aamodt and McShane (1992) reports small but significant effects; some more recent studies find no effects (Chidomere, 1989; Davis, 2017; Neely et al., 1994; Perlini et al., 1998; Vander Schee, 2013) while other recent studies do find difficulty differences across different item orderings (H. Chen, 2012; Pettit et al., 1986; Russell et al., 2003; Togo, 2002).

Although not aimed to explain these mixed results, studies have explored different aspects that can play a role in the relation between item order and test performance. For instance, the role of test anxiety, either as a moderator (if test anxiety is viewed as a trait, H. Chen, 2012) or a mediator (if test anxiety is viewed as a state, McKeachie et al., 1955) has been investigated. Further, the impact of topically ordered items on ease of memory retrieval has been studied (Pettit et al., 1986; Togo, 2002). In this paper, however, we focus on the role of test speededness. More specifically, we address the hypothesis stated by Leary and Dorans

[20]Note that this would not be the case if missing responses were not scored as incorrect. For example, in large-scale low-stakes assessments there is a vivid discussion revolving around alternative scoring or modeling techniques (e.g., Rose et al., 2017). However, due to reasons of test fairness these approaches are hardly applicable to high-stakes testing.

(1985) and Sax and Cromack (1966) above: If a test is administered under time constraints, different item orderings can substantially and differentially affect the test scores of individuals as it leads to test-takers distributing their time on items differently. Furthermore, we believe that this mechanism could (partly) explain the mixed findings regarding item order effects in the literature. In the following section, we illustrate how speededness can be defined and how likely it is to occur. For this, we first introduce a modeling framework that allows us to quantify ability and speed as latent constructs.

4.1.1 Modeling Framework

To investigate the effects of item ordering and speededness on test performance, a joint model for speed and ability is required. Note that we model responses and response times at the level of the item, and not at the level of the complete test. A convenient choice for the item response model is the Rasch model (Rasch, 1960). The Rasch model assumes that the probability of giving a correct response depends on a person parameter θ_i $i = 1, ..., n$, representing the person's ability (*ability parameter*) and an item parameter b_k $k = 1, ..., j$, representing the difficulty of the item (*difficulty parameter*):

$$P(y_{ik} = 1|\theta_i, b_k) = \frac{\exp(\theta_i - b_k)}{1 + \exp(\theta_i - b_k)}. \tag{49}$$

A useful property of the model is the fact that sum scores per person or item are sufficient statistics for the ability and difficulty parameters, respectively. This means that the number of correctly answered items by a person can function as a proxy for ability (*number-correct scoring*, a scoring approach that is very common for university exams). We use the terms test score and ability estimate interchangeably in this paper.

The most common model for modeling response times in cognitive testing situations is the lognormal model by van der Linden (2006), which assumes that response times are lognormally distributed. The model can be written as

$$\ln RT_{ik} = \lambda_k - \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{50}$$

The *item time intensity* in the model is represented by λ_k, the *person speed parameter* is represented by ζ_i. Note that both parameters often have substantial correlations with their ability counterparts (item difficulty and person ability) but are indeed separate parameters. ϵ_{ik} represents an item and person specific residual which is normally distributed with mean

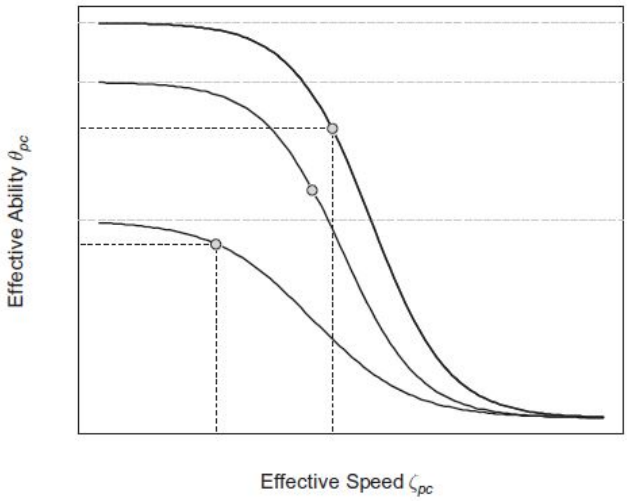
0 and an *item specific variance* $\sigma^2_{\epsilon_k}$. The joint hierarchical framework by van der Linden (2007) assumes joint multivariate normal item and person parameter distributions for the two dimensions ability and speed and allows the simultaneous estimation of both models.

4.1.2 Speededness

In his work, van der Linden (2011b) formally defines test speededness as an interaction of the time limit of a test, the amount of work a test requires and the working speed of the test-taker. This means a test is speeded for a test-taker if, given his/her optimal working speed, the person would run out of time before answering to all items. A useful concept for understanding test speededness is the so-called within-person speed-accuracy trade-off (Goldhammer, 2015). The trade-off refers to the fact that the *accuracy* or *effective ability* of a person (meaning the ability a person is able to show given a certain speed level) increases with increased amounts of time spent by the person on an item (see Figure 15). This increase has an upper bound: From a certain point on, additional time will not lead to more accurate answers. However, research in the area of response time modeling has shown that the speed distributions in test-taker samples are usually rather broad (van der Linden & Xiong, 2013), meaning that the working speed levels demonstrated by test-takers differ substantially. Meanwhile, practical constraints (e.g., limited space at universities) almost always require test administrators to use a fixed time limit in higher educational testing. Therefore, constructing unspeeded tests (so-called "pure power tests") in the context of higher educational high-stakes testing is practically impossible (Goldhammer, 2015). Instead, most tests can be considered a mixture, where at least for a small proportion of the testing population a certain level of speededness occurs on the test.

When test-takers experience test speededness (i.e., they run out of time while working on a test), they are confronted with the following three options: (a) omit items, (b) increase working speed and decrease accuracy, and (c) not reach the end of the test. An extreme form of (b) would, for example, be rapid guessing. In the context of number-right scoring, (a) is seen to be less favorable than (b) or (c), because, for example, guessing is expected to be not very time consuming while it still substantially increases the probability of a higher score (Millman et al., 1965). In high-stakes assessments, indeed omission rates are rather low and decreasing with increasing test experience of test-takers (Gafni & Melamed, 1994). In practice, for test-takers with an initial slow working speed, often a mixture of (b) from a certain point (Bolt et al., 2002; Goegebeur et al., 2008) and (c) would be expected. As

Figure 15: *Speed Accuracy Trade-Off as Illustrated by Goldhammer (2015).*



missing responses are usually scored incorrect in high-stakes assessments, all options (a) to (c) are reflected in lower test scores for test-takers that work under time pressure. Decisions of test-takers on which behavior to choose relate to the concept of test-wiseness.

4.1.3 Test-Wiseness

Millman et al. (1965) define test-wiseness as "[...] a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score." They emphasize that the construct is usually logically independent of the actual measured construct. Therefore, it is commonly seen as a source of construct-irrelevant variance in the measured scores (Rogers & Yang, 1996). Furthermore, research has shown that test-wiseness is often unevenly distributed across subgroups, for example across different ethical backgrounds (Ellis & Ryan, 2003), and depends on the cultural match of the test-taker and the test (Melikyan et al., 2019). Therefore, test administrators often seek to minimize the influence of test-wiseness or specific test preparations on test scores, for example by giving clear instructions on the test or choosing item types less connected to test-wiseness and test preparation (Powers, 1985; Powers & Rock, 1999).

A focal part of test-wiseness are time-using strategies (Millman et al., 1965). If a test is speeded for a test-taker, the test-taker has to allocate the available time in a way to maximize the expected score. This means that test-takers should identify and work on items that they are likely to answer correctly and for example guess on difficult items. Researchers

have hypothesized that time-strategies might be culture-dependent, meaning the concept of speeded tests may be more prevalent in certain cultures than in others (Melikyan et al., 2021). It is apparent that the ordering of items determines the requirement for time-using strategies: For example, if items are sorted hard to easy, test-takers have to actively decide to spend less time on the initial items of the test. If items are sorted easy to hard, this decision is not required.

4.1.4 Consequences of Different Item Orders under Speededness

The introduced frameworks can be used to illustrate theoretical implications of different item orderings if a test is speeded: If a test-taker does not respond to all items at the end of a test or works with decreased accuracy, this negatively affects the person's test scores. How much the scores are affected, however, depends on the properties of the items that are not-reached or on which a higher speed was used, as already noted by Leary and Dorans (1985). Consider an example where a test-taker works linearly with a constant and insufficient working speed on a test with a fixed time limit (i.e., only option (c) occurs). In Table 5, such an example is illustrated with not-reached items crossed out. The penalty for such a test taking behavior is much more severe on test form A, where three easy items are not-reached, than on test form B, where one hard item is not-reached. Note that such an effect is independent of specific item formats.

Table 5: *Two Reversely Ordered Test Forms with Item Difficulties b_k and Expected Response Times for a Specific Speed Level ζ_i.*

	Test form A		Test form B	
	b	RT	b	RT
Item 1	-0.1	10	1.2	60
Item 2	0.5	10	-0.6	30
Item 3	-0.2	20	~~-0.2~~	~~20~~
Item 4	-0.6	30	~~0.5~~	~~10~~
Item 5	~~1.2~~	~~60~~	~~-0.1~~	~~10~~

While it seems plausible to assume that most test-takers would speed up at the end of test form B in a realistic scenario, differences in the test scores would still occur between the two test forms. Obviously, it seems wisest for test-takers to distribute their time to items in an adaptive fashion and to use (informed) guessing on difficult items (Dodeen, 2008; Millman et al., 1965). However, the test forms in Table 5 penalize lack of speed and time-using strategies very differently, namely: Both are much more important on test form B than on test form A. Note that test-wiseness also might vary strongly between assessment contexts: For some

higher educational assessments, like TOEFL, GRE, ACT, or SAT test-takers and teachers sometimes spend considerable resources on preparation, for example trying to increase test-takers' test-wiseness (Gulek, 2003; Kulik et al., 1984), with studies showing mixed findings but in general positive effects (Kulik et al., 1984). However, in the context of university exams, this may be less common.

The impact of different item orderings on test scores depends on the following factors: (a) the time limit of the test, (b) the working speed of the test-taker, (c) the time intensity of items at the end of the test, and (d) the difficulty of items at the end of the test. Factors (a) and (b) determine the general level of speededness of the test independent of the specific item ordering. Factor (c) determines the number of items the person will not reach or work with a decreased accuracy on, depending on the specific item ordering. Finally, factor (d) determines the impact of not-reached items or decreased accuracy at the end of the test.

4.2 Illustrative Example

To illustrate potential problems of different item orderings in speeded tests, we use data from an experimental administration of a high-stakes quantitative reasoning test. The data were collected as part of a study with various experimental conditions (van Rijn et al., 2021). Participants were voluntary test-takers who wanted to prepare for the operational test and thus can be expected to be highly motivated. The overall correlation between the experimentally and later operationally measured ability was $r = .82$. The assessment contained 20 multiple-choice items. We analyze data from the conditions with a total time limit of 35 minutes. Feedback after every item was given to half of the students, but did not count towards the timing data. Item order was completely random for every test-taker. The data set consisted of 418 test-takers, of which 298 reported to be female and 119 reported to be male. The mean age in the sample was $M = 26.93$ ($SD = 5.97$). In total, 17 test-takers were excluded from the analysis due to aborted test sessions (15 cases) and technical problems (2 cases).

4.2.1 Is the Assessment Speeded?

To investigate whether the assessment is speeded, we investigated number of not-reached items and performance decline coupled with speeding up at the end of the test. Skipping unanswered items was prevented within the assessment software. Of the 401 test-takers in the data set, 5.0% did not reach the end of the test (i.e. they ran out of time before answering

to all items).

To investigate speeding up at the end of the test, we identified test-takers, who used almost all of the time available for the assessment. 127 of the 401 test-takers (31.7%) used more than 30 minutes. These test-takers are referred to as *slow test-takers*, whereas the other test-takers are referred to as *fast test-takers*. In addition, for each test-taker, we split the test in two parts according to the item order: The first 15 items and the last five items. We compared the response accuracy and the response times in the first and last part using proportion tests and median tests, respectively. Proportion correct were compared for the subset of slow test-takers including and excluding test-takers with not-reached items, as well as for fast test-takers. For both subsets of slow test-takers, on the first fifteen positions, the items are answered correctly more often (all slow test-takers: mean difference = 0.064, p = .009; slow test-takers without not reached items: mean difference = 0.065, p = .009). For fast test-takers, this difference is not meaningful (mean difference = 0.011, p = .517). The slow test-takers also take more time to answer to these items (median difference = 11.775, p = .001), while fast test-takers do not (median difference = 2.07, p = .10). Scatter plots with proportion correct and median response time on item level can be seen in Appendix B.1 Figures 38 and 39.

These findings indicate that for a substantial number of test-takers the assessment was speeded. These test-takers performed better on the items at the beginning of the test than on the items at the end of the test. This was partially due to not-reached items but also due to taking less time on the items at the end of the test which resulted in decreased accuracy.

4.2.2 What are the Potential Consequences?

To illustrate the potential consequences of different item orderings, we simulated data for the slowest test-takers in the sample. First, we estimated a joint response and response time model. Based on the estimated parameters, we simulated responses and response times and implemented different item orderings. The goal was to compare differences in sum scores within the test-takers for different item orderings.

Data Simulation. We used the R package `LNIRT` (Fox et al., 2021) to estimate a joint hierarchical framework for responses and response times with the above described models. The estimated person and item parameters were used to simulate responses and response times for the seven slowest test-takers. Note that responses and response times were also simulated for items that were originally not-reached for specific test-takers. We then applied

different orderings of items to illustrate maximum potential bias between differently ordered test forms: (a) sorting items by increasing time intensity ("Short-Long"), (b) sorting items by decreasing time intensity ("Long-Short"), (c) sorting items by increasing difficulty ("Easy-Hard"), (d) sorting items by decreasing difficulty ("Hard-Easy"). These orderings were chosen to illustrate maximally unfair ordered test forms. Response times were then accumulated. If the cumulative response times exceeded the time limit of 35 minutes, the items were scored as incorrect (in a real exam, these items would have been not-reached). We then compared the resulting sum scores for the test-takers across the differently ordered test forms. Note that this approach simulates data with a constant working speed. In real life it seems plausible to assume that some test-takers would compensate running out of time by speeding up. However, as mentioned earlier, such behavior would also result in lower test scores due to decreased accuracy.

Table 6: *Simulated Test Scores for Different Item Orderings for Seven Different Test-Takers with Different Speed (ζ) and Ability Levels (θ) of one Randomly Chosen Replication.*

ζ	θ	Short-Long	Long-Short	Easy-Hard	Hard-Easy	range(Σ)
-0.78	-0.87	5	1	5	2	4
-0.76	-0.83	3	3	3	3	0
-0.59	-0.02	7	5	8	4	4
-0.83	0.10	10	8	11	8	3
-0.64	-0.74	7	5	8	3	5
-0.59	-0.95	9	9	10	8	2
-0.58	-0.57	6	3	8	3	5

Note: Different item orderings means items were sorted in increasing or decreasing order by the respective item parameter time intensity (Short-Long or Long-Short) or difficulty (Easy-Hard or Hard-Easy). Columns contain the resulting test scores and column range(Σ) the maximum difference between these columns.

Results. Table 6 illustrates that different item orderings can indeed lead to substantially different test results. For example, one of the most extreme results occurs for the person in row five: On the test form with items sorted by increasing difficulty b the person achieves a sum score of 8, while on the test form with items sorted by decreasing difficulty b the person achieves a sum score of 3.

As the simulated responses and response times can vary substantially due to the probabilistic simulation process, we conducted 100 replications. The complete results can be seen in Appendix B.2. For each of the seven test-takers the average range in sum scores between test forms across replications was greater than 2.5. The maximum difference across replica-

tions was between 6 and 10. These are substantial differences for a test with 20 items. Note that we chose the most extreme item orderings possible in this illustrative example. However, if different versions of a test are created by ordering items randomly, these extreme orderings are also possible.

4.3 Proposed Solutions

In this paper, we are proposing two solutions to avoid item position effects in higher educational assessments. First, a certain number of items at the end can be fixed in constant ordering across test forms. This prevents differential effects of item ordering at the end of a test, as test-takers run out of time on identical items. While some may argue that this reduces test security, we would argue that at the end of a test, test-takers are less likely to work on the same item compared to the beginning of the test, because test-takers work at different speed levels. Obviously, it is not trivial to decide how many items or which portion of the test should have identical ordering at the end of the test forms. If too few items are chosen, test-takers might run out of time before the section is reached. If too many items are chosen, test security is lowered for no good reason. This can be seen as a security-fairness trade-off.

The effectiveness of the proposed approach can be enhanced by choosing to place the most time intensive items at the end of the test. By doing this, it becomes more unlikely that effects of speededness occur before the fixed set of items is reached. To investigate the effectiveness of the proposed approaches we conducted a simulation study with realistic conditions for a higher educational exam.

4.4 Simulation Study

For the simulation study, hyper-parameters were used from the analyses of two psychology exams (organizational and social psychology) at a Dutch university. Hence, the simulated data is representative for the high-stakes higher educational testing context. Both exams contained 25 multiple-choice items and one open-answer item administered under a time limit of 40 minutes. The exams were conducted on computers in an online assessment setting and taken by 527 first-year psychology students. Students were not allowed to review items and all items were presented in a random order to the students. Responses and response times to all multiple-choice items were available for analysis, while item order was not. We analyzed the data using the R package `LNIRT`. As the results were very similar for both exams, we only

report the results of the organizational psychology exam below. The hyper-parameters of the item and person parameter distributions are depicted in Appendix B.3 Table 17. The estimated correlation between item difficulty and time intensity was $r = 0.62$. The estimated correlation between speed and ability was $r = 0.24$.

4.4.1 Design

In the simulation, we used the illustrated hyper-parameters to create a realistic test containing 40 items. In each of the conditions, two test forms were created. We conducted the simulation study to answer the following questions: (a) Are the proposed approaches effective in preventing unfair effects of different item orders? (b) What are the effects if the number of items with fixed positions at the end of the test forms is too low? (c) What are the effects if time intensity is not known before the assessment and must be (imperfectly) predicted?

We varied two experimental factors: The number of items with fixed positions at the end of the test (three levels: [0; 5; 10]) and the selection and ordering of these items (three levels: [random; based on an item time intensity covariate[21]; based on true item time intensity]). Because the second factor is irrelevant when the number of items with fixed positions is equal to zero, this resulted in overall seven conditions.

To observe a variety of speed and ability levels, person parameters were created as a grid: Speed levels were $[-0.6, -0.4, -0.2, 0]$ and ability levels were $[-1, 0, 1]$. These values were chosen because effects of speededness are relevant across all ability levels, but mainly relevant for slower test-takers. The grid also represents the width of possible person parameters according to Appendix B.3 Table 17. The time limit was set at 40 minutes. Responses and response times were created according to the Rasch model and the log-normal response time model (cf. above). Test scores were calculated. In total, 1000 replications were conducted. The complete R code for the simulation can be accessed here: https://osf.io/d97b5/?view_only=804fc3db7aab466e8cb358c6f7c7fa8c.

4.4.2 Results

To analyze unfairness of the test forms we compared test scores between the two test forms for all conditions and replications. In Table 7, the average and the maximum difference between the test scores on the test forms are depicted for all seven conditions. Note that the table only contains the results for the slowest but most able test-takers. The table illustrates that

[21] Empirical mean correlation with true item time intensity of $r = .61$

there can be considerable differences between two test forms with exactly the same items but different item orderings, if no measures are taken, with $M(\Delta) = 0.90$. The table can also be used to answer the research questions stated above: (a) Indeed, the proposed measures reduce differences between test forms. In the condition with the strongest control measures (ten items fixed, sorting based on time intensity), there are almost no differences between test forms on average, with $M(\Delta) = 0.19$. In fact, the simulation indicates that even imperfect measures serve the purpose of reducing effects of different item orderings, albeit less strongly. (b) If there are only five items fixed, which are sorted based on time intensity, the resulting mean difference between test forms is $M(\Delta) = 0.43$. (c) If the sorting occurs based on a covariate of time intensity, the resulting mean difference between test forms is $M(\Delta) = 0.22$. This indicates that number of items held constant is more important than quality of the time intensity prediction[22].

Table 7: *Results of the Simulation Study: Mean (M(Δ)) and Maximum Difference (Max(Δ)) between the Test Scores for the Two Identical Test Forms with Different Item Orderings.*

Fixed Positions	Ordering Items with Fixed Positions	M(Δ)	Max(Δ)
0	Random	0.90	4.08
5	Random	0.69	2.98
10	Random	0.32	1.36
5	Based on covariate	0.58	2.46
10	Based on covariate	0.22	1.07
5	Based on time intensity	0.43	1.90
10	Based on time intensity	0.19	0.68

Note: Item ordering was either completely random (0 items constant), or random with either the last five or ten items fixed. The constant items were either picked randomly or the most time intensive items ('Based on Time Intensity') or presumably most time intensive items were fixed ('Based on Covariate').

Complete results for all person parameter combinations can be seen in the Appendix B.3 Figures 40 and 41 for mean and maximum differences across replications, respectively. Results for the other person parameter combinations are comparable, albeit decreasing with increasing speed (test-takers run out of time less early) and decreasing ability (test-takers are less punished for not answering to items as they would have had a lesser chance of answering them correctly anyway).

[22]Note that the maximum differences in sum scores between the test forms in Table 7 are less pronounced than in Appendix B.2 despite the longer test forms. This is due to the fact that for Appendix B.2 item orders have been specifically chosen to be as unfair as possible (e.g. Short-Long vs. Long-Short). Furthermore, in Table 7 results are aggregated across multiple test-takers while in Appendix B.2 results are depicted for individual test-takers.

4.5 Discussion

In the past, there have been various studies on whether different item orderings in higher educational testing are an adequate measure to increase test security or a potential source of unfairness. In a small illustration using quantitative reasoning data, we have shown that speededness plays a neglected but important role in the matter: When a test is speeded it becomes important to consider which items are placed at the end of a test, as these items are more likely to be not reached or test-takers allocate less time on them than on items at the beginning of the test. Furthermore, using the data set we illustrated how speededness can be detected by investigating missing responses and item position effects. To prevent such unfair test forms, we proposed two straightforward measures to prevent effects of item ordering in speeded higher educational tests: Fixing the last items across test forms and additionally picking the most time intensive items for these positions. In a simulation study based on data of Dutch university psychology exams, we illustrated that these approaches are indeed suitable to prevent unfair test forms regarding item ordering.

4.5.1 Practical Recommendations

From a practical point of view, the question arises how large the proportion with constant ordering at the end of a test should be and how time intensive items can be identified. In an ideal world, this should be determined be pretesting the test and determining the level of speededness. This could be done by using similar measures as in the illustrative example above or more complex modeling techniques such as change point analyses (Bolt et al., 2002; Goegebeur et al., 2008) However, a lot of higher educational exams and tests do neither have the opportunity to allow for extensive item pretesting without compromising test security nor have the required resources.

Hence, in many realistic settings, test administrators will have to rely on some assumptions and heuristics: Based on our analyses, we argue that holding one fourth of the items at the end of the test constant is a reasonable measure. Thereby test security is still not severely threatened but this proportion covers the part of the test on which changes to test taking behavior might be likely to occur. Note that the requirement of items held constant depends on the discrepancy between time intensities: If a single, very time intensive item takes up one fourth of the testing time for most test-takers, it might be sufficient just to put this single item at the end of the test. Moreover, if item difficulty and item time intensity are expected to be highly correlated, items that are anticipated to be difficult can be chosen for

positions at the end of the test. Time intensity can also be expected to depend strongly on item type; open-answer items or elaborate constructed-response items can be expected to be almost always more time intensive than multiple-choice items. Finally, it should be noted that assigning time intensive items to the last item positions across test forms has the positive side effect of reducing the general influence of test-wiseness on test scores.

4.5.2 Alternative Approaches

Of course, there are also different (but more complex) approaches to prevent problems regarding item order effects: Item time limits could be set to reduce differential effects of speededness (Goldhammer, 2015), as they prevent test-takers to distribute their time unwisely on the test. Furthermore, van der Linden and Xiong (2013) proposed a useful approach to control speededness in the framework of computer adaptive testing. While these approaches seem theoretically promising, they would often pose a substantial modification to higher educational assessment practice and require computer-based testing.

Alternatively, there is a wide range of psychometric models which aim at disentangling speed and ability. Even if effects of different item orderings have occurred, these models could be used to prevent bias in ability estimation (e.g., Pohl et al., 2019; Rose et al., 2017). However, most of these models were designed for use in low-stakes assessments and might be prone to gaming (e.g., test-takers purposely not reaching the end of the test). Furthermore, they require the availability of response times for analyses, which are only available in computer-based testing.

It is noteworthy that in some contexts, test forms are created with no item overlap (e.g., different administrations of the GRE or TOEFL). In such situations, often approaches known as automated test assembly are used to create parallel test forms (van der Linden, 2005). However, when such test forms are used, having exactly the same items fixed at the end of a test is impossible, as these test forms do not share the same items. Some of the mentioned approaches above (item time limits, CAT) may be able to solve the problem of unfair test forms due to different items at the end of test forms. However, the additional administration conditions that are required for these approaches may not be feasible in all testing situations where multiple test forms are assembled. Further research could investigate how fair test forms can be assembled in the context of test time limits and differences in speed between the test-takers.

4.5.3 Conclusion

Although impact of item ordering on test fairness has been a topic of research for more than 50 year, the role of test speededness has been largely left unaddressed. In this paper, we have shown that especially when test-takers work under substantial time pressure and run out of time at the end of the test, item order plays a crucial role. Large differences in the expected test score are created between test forms that are supposed to be equivalent. To mitigate this issue, we have proposed two measures which keep the advantage of different item orders (increasing test security) while preventing unfair test forms: Keeping a certain number of, ideally time intensive, items constant at the end of a test. We believe that these measures can be easily implemented in practice and thereby help create fair test forms in the context of higher educational testing.

5 Bayesian Hierarchical Response Time Modeling – A Tutorial

Published as: König*, C., Becker*, B., & Ulitzsch*, E. (2023). Bayesian Hierarchical Response Time Modeling – A Tutorial. *British Journal of Mathematical and Statistical Psychology,* Advance online publication. https://doi.org/10.1111/bmsp.12302

*All authors contributed equally to the manuscript.

Abstract: Response time modeling is developing rapidly in the field of psychometrics, and its use is growing in psychology. In most applications, models for response times are modeled jointly with models for responses, thereby stabilizing estimation of response parameters and enabling research on a variety of novel substantive research questions. Bayesian estimation techniques facilitate estimation of response time models. Implementations of these models in standard statistical software, however, are still sparse. In this accessible tutorial, we discuss one of the most common response time models, the log normal response time model, embedded in the hierarchical framework by van der Linden (2007). We provide detailed guidance on how to specify and estimate this model in a Bayesian hierarchical context. One of the strengths of the presented model is its flexibility, which allows to adapt and extend the model according to researchers' needs and hypotheses on response behavior. We illustrate this based on three recent model extensions: (a) application to non-cognitive data incorporating the distance-difficulty hypothesis, (b) modeling conditional dependencies between response times and responses, and (c) identifying differences in response behavior via mixture modeling. This tutorial gives non-specialist and applied researchers a better understanding of the use and utility of response time models, showcases how these models can easily be

adapted and extended, and contributes to a growing use of these models to answer novel substantive research questions in both non-cognitive and cognitive contexts.

5.1 Introduction

The rise of computer-based assessments is accompanied by the opportunity to record additional information on test-taking behavior, such as response times, mouse movements, keystrokes, or clickstreams, to name just a few. Out of these, response times have received by far the most attention in psychological, psychometric, and methodological research. Response times are usually defined as time on task (how much time was spent on an item in total; e.g., OECD, 2016a). They can help obtaining more precise ability estimates (van der Linden et al., 2010) and support addressing substantive research questions related to how test-takers allocate their time (e.g., Naumann & Goldhammer, 2017). Other applications of response time analyses include the assembly of equivalent test forms while keeping speededness parallel (van der Linden, 2011b), the detection of aberrant test behavior due to fraudulent behavior or malfunctioning items (van der Linden & Guo, 2008), or detecting disengagement in low-stakes assessments (Ulitzsch et al., 2019a, 2020).

Response time analysis has a rich history in other fields such as experimental psychology, dating back multiple centuries (Craigmile et al., 2010; Luce, 1986; Vandekerckhove et al., 2011). Nonetheless, most response time models are quite recent developments (i.e. within the last 10-20 years) in the psychometric literature that are rarely used in substantive or applied research. A potential reason is that the models and methods, especially their latest extensions, are not straightforward to implement in standard statistical software (e.g. R, SPSS, or Stata). This is further aggravated by the fact that most recent advances in psychometric response time modeling utilize Bayesian hierarchical modeling.

In this accessible expert tutorial, we aim at providing guidance to non-specialists and applied researchers on how to specify and estimate one of the most popular response time models and three recent model extensions in a Bayesian hierarchical context. We first give a brief overview over the response time literature in general. We then focus on the three-parameter lognormal model for response times by Klein Entink, Fox, and van der Linden (2009) and show how this response time model can be jointly estimated with common *item response theory* (IRT) models for item responses in the hierarchical framework by van der Linden (2007). Second, we give detailed step-by-step instructions on its specification and estimation within a Bayesian hierarchical modeling approach. Third, we show how to extend the basic hierarchical framework to model (a) the distance-difficulty hypothesis in the context of non-cognitive data (Ferrando & Lorenzo-Seva, 2007), (b) conditional dependence between response times and accuracy (Bolsinova, de Boeck, & Tijmstra, 2017), and (c) qualitative dif-

ferences in response behavior based on a mixture modeling approach (Ulitzsch et al., 2019a), thereby showcasing the framework's flexibility in adjusting to researchers' needs.

5.1.1 Response Time Modeling

For historical reviews of the response time literature see Schnipke and Scrams (2002), Lee and Chen (2011), van der Linden (2009a), as well as Kyllonen and Zu (2016). The latest review is provided by De Boeck and Jeon (2019). De Boeck and Jeon (2019) classify the existing response time models into four categories: (a) response time models (with response times being the sole depended variable), (b) joint models (with an additional depended variable, most commonly response accuracy), (c) dependency models (in which joint models are extended to accommodate residual dependencies), and (d) response times as a covariate models (response times are used to predict another variable, e.g. accuracy).

This tutorial focuses on joint models and dependency models, which currently are receiving the most attention in the psychometric literature[23]. According to De Boeck and Jeon (2019), there are three different families of joint response time models: members of the *generalized linear item response theory* modeling framework (B-GLIRT, Molenaar, Tuerlinckx, & van der Maas, 2015b), diffusion models, and race models. Furthermore, there is a research line of enriching *cognitive diagnostic models* (CDM) with response time data (Zhan et al., 2018). It is important to note that CDMs, diffusion models, and race models are process models, which means that they aim at explaining the processes that lead to different responses and response times. In contrast, the B-GLIRT model family consists of purely descriptive measurement models (De Boeck & Jeon, 2019). The most popular joint model is, by far, the hierarchical model by van der Linden (2007). It can be subsumed under the B-GLIRT family (Molenaar, Tuerlinckx, & van der Maas, 2015b) and is arguably the most widely used response time modeling framework (De Boeck & Jeon, 2019).

5.1.2 The Hierarchical Framework by van der Linden (2007)

For simultaneous modeling of response times and item responses, van der Linden (2007) proposes a hierarchical framework, which basically resembles a two-dimensional latent factor model with common multivariate item and person parameter distributions. A central characteristic of the framework is its "plug and play" approach, where the component models

[23]Note that response time models are often contained within joint models; therefore, if researchers are able to implement joint models, implementing response time models without the response counterpart becomes trivial.

are open to flexible adaptation. The framework has frequently been applied to address substantive research questions (e.g., Debelak et al., 2014; Goldhammer & Klein Entink, 2011; Scherer et al., 2015) and has been subject to various extensions and modifications. Examples for these are models that take residual dependencies between responses and response times into account (e.g., Bolsinova, 2016), allow for varying speed and accuracy throughout the test (e.g., Fox & Marianti, 2016; Molenaar et al., 2016), or aim at detecting and modeling differences in response processes, for instance aberrant response behavior (van der Linden & Guo, 2008) such as rapid guessing behavior (Ulitzsch et al., 2019a; C. Wang & Xu, 2015).

In its original form, the framework consists of three components. The first two components are the measurement models specified for item responses and the associated response times. The third component consists of the joint distributions for the parameters of the measurement models. In the following, we introduce the framework in a Bayesian hierarchical context with a *two-parameter logistic* (2PL) IRT model for responses and the *three-parameter lognormal* (3PLN) model for response times as two widely employed measurement models for either type of data.

The First Component: The Two-Parameter Logistic Model for Item Responses. The 2PL IRT model is one of the most commonly used measurement models for the response model in the hierarchical framework. In the 2PL model applied in a cognitive context, the probability of person i, $i = 1, ..., I$, to solve item k, $k = 1, ..., K$, correctly can be written as:

$$P(y_{ik} = 1|\theta_i, a_k, b_k) = \frac{\exp(a_k(\theta_i - b_k))}{1 + \exp(a_k(\theta_i - b_k))}, \tag{51}$$

with θ_i denoting person i's *ability*, and a_k and b_k giving item k's *item discrimination* and *item difficulty*, respectively. For detailed information on general concepts of IRT and the 2PL model in particular, see, for example, de Ayala (2022).

The Second Component: The Three-Parameter Lognormal Model for Response Times. For response times, measurement models based on the lognormal distribution are often used, since response times are non-negative and positively skewed (Schnipke & Scrams, 1997)[24]. The 3PLN response time model by Klein Entink, Fox, and van der Linden (2009) (see also Ranger & Ortner, 2012a) is a simple generalization of the two-parameter lognormal

[24]There are alternative distributions that can be used to model response times, such as the gamma distribution, the ex-Gaussian distribution, or the Weibull distribution (for a brief overview, see De Boeck & Jeon, 2019). Nevertheless, it has been shown that lognormal models oftentimes provide the best fit, compared to models based on other distributions (Schnipke & Scrams, 2002)

response time model by van der Linden (2006). In the 3PLN model, response times RT_{ik} are assumed to be lognormally distributed and modeled as

$$\ln RT_{ik} = \lambda_k - \phi_k \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \tag{52}$$

The *time intensity parameter* λ_k represents the workload of an item, while the *speed sensitivity parameter* ϕ_k represents how strongly items differentiate between slow and fast test-takers. The *speed parameter* ζ_i represents the speed of a person i working on the test, and the parameter $\sigma^2_{\epsilon_k}$ is an *item-specific residual variance*. The model resembles a unidimensional confirmatory factor analysis model with freely estimated intercepts (time intensities), factor loadings (speed sensitivities), and residual variances. Due to the log-transformation of response times, parameter interpretation is less straightforward. For more detailed explanations see the work of van der Linden (2006) and Becker, Debeer, Weirich, and Goldhammer (2021).

The Third Component: The Hierarchical Structure of the Joint Parameter Distributions. The hierarchical framework models both item and person parameters as random effects, and assumes that the item and person parameters of both measurement models stem from common multivariate normal distributions. For the presented measurement models, this results in the following person parameter distribution:

$$(\theta_i, \zeta_i) \sim \mathcal{MVN}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}). \tag{53}$$

The respective item parameter distribution can be denoted as

$$(\ln a_k, b_k, \ln \phi_k, \lambda_k) \sim \mathcal{MVN}(\boldsymbol{\mu_K}, \boldsymbol{\Sigma_K}). \tag{54}$$

The log transformations of the item discrimination and speed sensitivities are commonly applied to enable the use of a joint multivariate normal item parameter distribution (Bolsinova, 2016; Glas & van der Linden, 2003), as these parameters are often constrained to

positive values.[25][26] The full likelihood for this and all other models discussed in this tutorial are provided in Appendix C.1 or online supplement S0. All online supplements, including model specifications and scripts necessary to estimate the models, are available on the online repository at https://osf.io/k4m3s/.

For a detailed overview over the hierarchical framework and its assumptions see the work of van der Linden (2007). Two crucial assumptions of the hierarchical framework are the constant speed assumption and the assumption of conditional independence. The constant speed assumption implies that test-takers work with the same level of speed throughout the test. This assumption mirrors the assumption of constant ability of most IRT models. The conditional independence assumption implies that responses and response times are independent conditional on the common person and item parameter distributions. Both assumptions have repeatedly been questioned (e.g., Bolsinova, de Boeck, & Tijmstra, 2017; Domingue et al., 2021) and are, for example, violated if test-takers speed up at the end of a test. Such speeding-up could occur if there is a strict speed limit on a test or if the motivation of test-takers declines throughout the test. From a practical perspective, these assumptions are reasonable for most applications and comparable to the assumptions of most "plain" IRT models. However, there are model extensions in the hierarchical framework, which allow modeling, for example, variable speed or conditional dependencies (Bolsinova, de Boeck, & Tijmstra, 2017; Fox & Marianti, 2016). One of these extensions will be presented in detail in the model extension section of the tutorial.

5.2 Doing Bayesian Hierarchical Response Time Modeling

In this section, we provide detailed guidance on how the basic hierarchical framework based on the 2PL and 3PLN models can be specified and estimated using a Bayesian hierarchical modeling approach. In this tutorial, we will use Stan (Carpenter et al., 2017) and its R package `rstan` (Stan Development Team, 2021), a general-purpose Bayesian estimation software utilizing the No-U-Turn-Sampler (NUTS) that is based on Hamiltonian Markov chain Monte Carlo (MCMC) sampling. While the basic hierarchical framework can easily be estimated

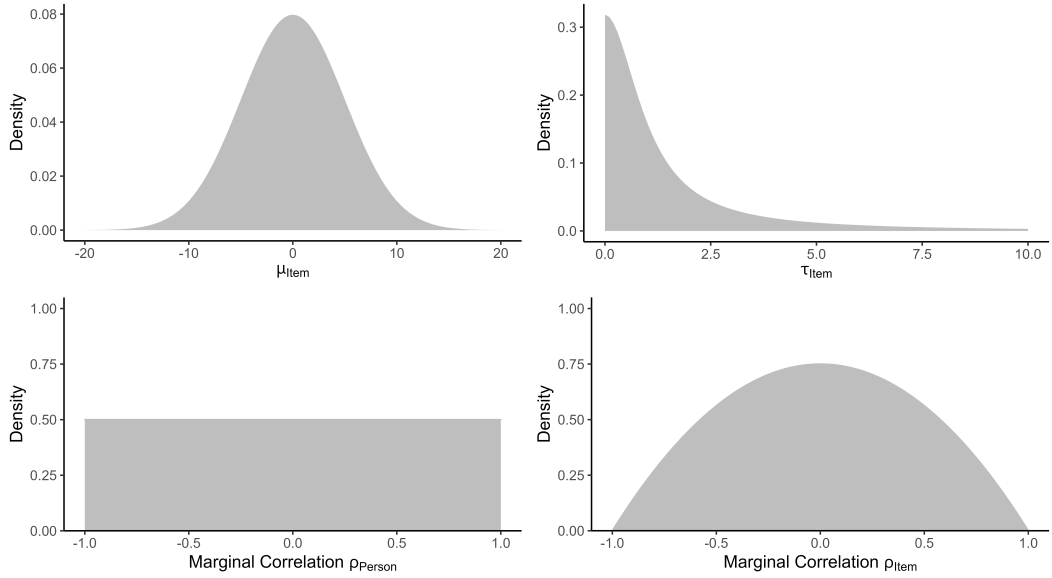
[25]Speed sensitivities and discriminations are constrained to positive values to deal with rotation indeterminacy. Note that this assumption also aligns with theoretical considerations in the context of cognitive assessments where, commonly, all items can be assumed to load in the same direction on the ability and speed factors. When researchers do not deem this assumption plausible, e.g., when analyzing items from non-cognitive assessments worded in different directions, they can, without loss of generality, model item parameters as fixed effects, and constrain the first speed sensitivity and discrimination parameter to positive values, while leaving the remaining speed sensitivity and discrimination parameters unconstrained.

[26]Note that this implementation deviates from Equation 18 presented in Chapter 1, as a and ϕ are log-transformed. However, this does not change the conceptual meaning of the respective parameters.

using Maximum Likelihood (ML) estimation implemented in standard statistical software (e.g., Mplus, lavaan), ML implementations and estimation of the more complex extensions can be challenging. Although custom-made Gibbs samplers exist for some of the extensions, such customized samplers cannot easily be adapted to researchers' needs. Furthermore, these tailored Bayesian solutions do not offer the full range of diagnostics to assess the quality of the resulting parameter estimates, in contrast to modern general-purpose Bayesian software such as Stan. As detailed general introductions to Bayesian modeling and the Stan programming language are beyond the scope of this paper, we refer readers to a collection of resources in Appendix C.2 or online supplement S8. Note that this list is by no means exhaustive but a small collection of resources we ourselves found helpful when doing Bayesian modeling using Stan.

The Bayesian hierarchical modeling approach illustrated here warrants further introductory explanations. Instead of assigning prior distributions to the individual item or person parameters, hyperprior distributions are specified for the grand means, standard deviations, and correlations of the joint multivariate item and person parameter distributions (see Equations 53 and 54). This hierarchical specification results in partial pooling, as information of all parameters assumed to stem from a joint distribution is used for estimating an individual parameter (e.g., Jackman, 2009). This is beneficial in terms of precision; the standard errors of the parameter estimates are usually smaller compared to non-hierarchical approaches. For implementing the hierarchical specification, we follow a separation strategy (Barnard et al., 2000; König et al., 2020), which allows assigning individual prior distributions for standard deviations and correlations, giving researchers more fine-grained control over these prior distributions. More details are given in the following sections describing the actual implementation in Stan. We use weakly informative prior distributions in the form of the Cauchy (or, alternatively, the exponential) distribution as hyperpriors for the standard deviations (Gelman, 2006; Polson & Scott, 2012). They introduce a small amount of information into the analysis, primarily coming from the knowledge of the typical range of the parameters. They facilitate convergence and sampling efficiency, which in turn leads to an increased accuracy of the resulting parameter estimates, especially in smaller samples (e.g., König et al., 2020). We illustrate the prior densities for the grand means and standard deviations of the item parameters, and for the marginal correlations of the person and item parameters in Figure 16.

Figure 16: *Prior Distributions for Grand Means, Standard Deviations, and Marginal Correlations.*



Data. The data has to be provided in long format, where the number of rows equals the total number of observed responses. Table 8 illustrates the first ten rows of a data set in the long format. The electronic supplement S1 contains exemplary R code for reshaping the data from wide to long format, and how to pass the data to Stan. Note that all input objects for Stan need to be supplied as a named list, with list names specifying how objects can be referred to within the Stan code. Figure 17 shows how all data input is defined in a single list object for a data set consisting of responses and response times for 12 items and 500 persons.

Table 8: *First Ten Rows of an Example Data Set in Long Format.*

ID	item	y	RT	log_RT
1	1	0	164.50	5.10
2	1	1	16.66	2.81
3	1	1	62.45	4.13
4	1	1	37.61	3.63
5	1	0	13.57	2.61
6	1	1	66.81	4.20
7	1	0	26.39	3.27
8	1	1	44.77	3.80
9	1	1	32.25	3.47
10	1	0	40.02	3.69

Note: ID: test-taker identifier; item: item number; y: response; RT: response time in seconds; log_RT: log-transformed response time.

Having prepared the data in R, we now turn our attention to the specific components of a Stan model file (with `.stan` extension). The complete `stan` code for the basic hierarchical

Figure 17: *R Code Setting up the Data Input for the pisaL Dataset Consisting of Responses and Response Times for 12 Items and 500 Persons for Stan.*

```
# data in Stan format
K <- 12
I <- 500
pisa_data <- list(K   = K,
                  I     = I,
                  Nobs  = K * I,
                  kk    = as.integer(pisaL$item),
                  ii    = as.integer(pisaL$ID),
                  y     = pisaL$y,
                  logrt = pisaL$log_RT)
}
```

model can be found in online supplement S2. The first component is the `data` block, which defines and initializes the core elements of the data (see Figure 18): the number of test-takers `I`, the number of items `K`, and the number of the observed responses and associated response times `Nobs` as integer values (see the `int`-descriptor). It is possible to specify lower and upper bounds. The dimensions of the objects are declared by the value (or other descriptors) between the square brackets. The item responses `y` (integer values of 0 and 1, denoting wrong and correct responses, respectively) and the log transformed response times `logrt` (real numbers without bounds) are defined as vectors of length `Nobs`. Finally, the vectors `ii` and `kk` are used as person and item identifiers for each observed response and the associated response time.

Figure 18: *The Data Block for the Hierarchical Framework Specification.*

```
data {
  int<lower=1> K;
  int<lower=1> I;
  int<lower=1> Nobs;
  int<lower=1> kk[Nobs];
  int<lower=1> ii[Nobs];
  int<lower=0,upper=1> y[Nobs];
  vector[Nobs] logrt;
}
```

Parameters. The `parameters` block (Figure 19) declares all model parameters to be estimated (here: individual person and item parameters), as well as the associated hyperparameters of the prior distributions (here: parameters describing the joint distribution of person and item parameters). We first declare an array `person` containing I row vectors of length 2 (for the two person parameters ability and speed), and an array `item` containing K row vectors of length 4 (for the four item parameters discrimination, difficulty, speed sensitivity, and time intensity) to store the person and item parameters, respectively. Second, we declare a vector of length `K` of residual standard deviations of the log response times `sigma_e`.

Since these residuals cannot be negative, we introduce a lower bound of zero as a parameter constraint (`<lower=0>`). Third, a vector `mu_item` of length 4 is declared that contains the grand means of the multivariate normal hyperprior distribution for the four item parameters. For identification, the means of the bivariate normal hyperprior distribution for the person parameters are hard-coded to be zero, and thus not declared as parameters. Fourth, a vector `tau_item` of length 4 is declared that contains the standard deviations of the multivariate normal hyperprior distribution for the four item parameters. Because the standard deviations of the bivariate normal hyperprior distribution for the person parameters are hard-coded to be one for model identification, they are not explicitly declared as parameters. Lastly, `L_Omega_person` and `L_Omega_item` are the Cholesky factors of the correlation matrices of the person and item parameters, respectively. Together with the vector of standard deviations `tau_item`, the Cholesky factors are central parts of the aforementioned separation strategy.

Figure 19: *The Parameters Block for the Hierarchical Framework Specification.*

```
parameters{
  row_vector[2] person[I];
  row_vector[4] item[K];
  vector<lower=0>[K] sigma_e;
  vector[4] mu_item;
  vector<lower=0>[4] tau_item;
  cholesky_factor_corr[2] L_Omega_person;
  cholesky_factor_corr[4] L_Omega_item;
}
```

Transformed Parameters. In the `transformed parameters` block (Figure 20) we declare transformations of the raw parameters. This may include transformations of auxiliary parameters into parameters of substantive interest, identification restrictions of certain parameters, or simple convenience transformations. In our case, the transformations essentially just give explicit names to the columns of `person` and `item` arrays. These are convenience transformations that support readability of the code. We declare six vectors (four of length `K`, two of length `I`) storing the item discriminations, difficulties, speed sensitivities, and time intensities (`discrimination`, `difficulty`, `sensitivity`, `intensity`), and the ability and speed parameters (`ability`, `speed`). Since item discriminations and speed sensitivities cannot be negative, we declare a lower bound of zero with the `lower=0` command. In the following, the actual transformations are specified with the `to_vector` command, which transforms the columns of the `person` and `item` arrays into column vectors. Recall that log-transformed item discriminations and time sensitivities are modeled in the joint item parameter distribution.

Thus, to obtain untransformed item discriminations and time sensitivities, the respective transformed column vectors are exponentiated. Note that all transformed parameters have to be declared before specifying the transformations, otherwise the model specification will not compile.

Figure 20: *The Transformed Parameters Block for the Hierarchical Framework Specification.*

```
transformed parameters{
  vector<lower=0>[K] discrimination;
  vector[K] difficulty;
  vector<lower=0>[K] sensitivity;
  vector[K] intensity;
  vector[I] ability;
  vector[I] speed;

  discrimination = exp(to_vector(item[,1]));
  difficulty = to_vector(item[,2]);
  sensitivity = exp(to_vector(item[,3]));
  intensity = to_vector(item[,4]);
  ability = to_vector(person[,1]);
  speed = to_vector(person[,2]);
}
```

Model. In the `model` block (Figure 21), the core of the script, we specify the prior distributions for the parameters declared in the `parameters` block, followed by the likelihood of the model. We first specify an LKJ prior distribution[27] (Lewandowski et al., 2009) for the Cholesky factor of the correlation matrix of the person parameters `L_Omega_person`, governed by the parameter $\eta \geq 1$. As η increases, the density increasingly concentrates around the identity matrix, giving more weight to smaller correlations. Our specification ($\eta = 1$) implies a uniform prior on correlations. Next, we specify a multivariate normal prior distribution for each row of `person`, with a zero-vector as grand means and covariance matrix $\boldsymbol{\Sigma_I} = \text{diag}(\boldsymbol{\tau_I}) L_{\boldsymbol{\Omega_I}}$. The `diag_pre_multiply` command returns the product of the diagonal matrix `tau` (in our case, the diagonal is given by a vector of ones) and the Cholesky factor `L_Omega_person`. Both the LKJ and the multivariate normal distributions are used in their Cholesky parameterization for efficiency purposes. This prior specification is repeated for the item parameters, where we need to assign prior distributions to the grand means and standard deviations of the multivariate normal prior distribution as well. For the LKJ prior on the correlations, setting $\eta = 1$ does not result in a uniform prior in this case, because the number of dimensions is larger than for the person parameters (see Figure 16). The density, however, concentrates its mass around zero, and makes extreme correlations less likely. Furthermore, we assign a weakly informative normal prior distribution to the grand means, and a weakly

[27] LKJ is an acronym consisting of the first letters of the family names of the authors of the original paper.

informative half-Cauchy distribution to the standard deviations. We further employ a weakly informative half-Cauchy prior distribution for the residual standard deviations `sigma_e`.

Declaring the likelihood completes the model specification. In our case, we have the 2PL model that assumes that the responses y_{ik} follow a Bernoulli distribution in logit specification, governed by Equation 51, and the 3PLN model that assumes that the log response times follow a normal distribution governed by Equation 52. The `target+=`-command is the sampling statement that, in case of the 2PL model, indicates that, for instance, a Bernoulli distribution is added to the target density (the target density being the joint posterior distribution of the model). The `_lpmf` and `_lpdf` suffixes denote log probability mass functions and log probability density functions, respectively. Alternatively, the likelihood can also be specified with simplified sampling statements: `y ~ bernoulli_logit(discrimination[kk] .* (ability[ii] - difficulty[kk])` and `logrt ~ normal(intensity[kk] - sensitivity[kk] .* speed[ii], sigma_e[kk])`. Note that the likelihood is vectorized (the `.*` operator that is the elementwise product of the `sensitivity` and `speed` vectors) and no loop over observations is necessary.

Figure 21: *The Model Block for the Hierarchical Framework Specification.*

```
model{
  target += lkj_corr_cholesky_lpdf(L_Omega_person | 1);
  target += multi_normal_cholesky_lpdf(person | [0,0],
            diag_pre_multiply([1,1], L_Omega_person);

  target += lkj_corr_cholesky_lpdf(L_Omega_item | 1);
  target += normal_lpdf(mu_item | 0,5);
  target += cauchy_lpdf(tau_item | 0,2);
  target += multi_normal_cholesky_lpdf(item | mu_item,
            diag_pre_multiply(tau_item, L_Omega_item);
  target += cauchy_lpdf(sigma_e | 0,2);

  target += bernoulli_logit_lpmf( y | discrimination[kk] .*
       (ability[ii] - difficulty[kk]));
  target += normal_lpdf( logrt | intensity[kk] - sensitivity[kk] .*
       speed[ii], sigma_e[kk]);
}
```

Generated Quantities. The `generated quantities` block allows us to calculate additional quantities that can be derived from sampled parameters and are necessary for *posterior predictive checks* (PPC), for evaluating the model fit, and for model comparisons. These calculations are done after the sampling process; thus, these parameters and additional quantities do not factor into the likelihood. In our application (Figure 22), we want to calculate the correlation ($\mathbf{\Omega_K}$ and $\mathbf{\Omega_I}$) and covariances matrices ($\mathbf{\Sigma_K}$ and $\mathbf{\Sigma_I}$) of the item and person parameters, respectively. To calculate $\mathbf{\Omega_K}$ and $\mathbf{\Omega_I}$, we simply multiply the respective Cholesky factors with their transpose, e.g. $L_{\mathbf{\Omega_I}} L_{\mathbf{\Omega_I}}^T$ (in the code block below, this operation is

accomplished by the `multiply_lower_tri_self_transpose`-command). Then, for example, $\boldsymbol{\Sigma_K}$ is simply $\text{diag}(\boldsymbol{\tau_K})\boldsymbol{\Omega_K}\text{diag}(\boldsymbol{\tau_K})$ (this operation is accomplished by the `quad_form_diag`-command). Therefore, we first define the type (`corr_matrix` and `cov_matrix` for correlation and covariance matrices, respectively), dimension, and name of the desired quantity, and then indicate how to calculate it. Since both variance components of $\boldsymbol{\Sigma_I}$ are fixed to 1 for identification reasons, the second argument of the respective `quad_form_diag`-command is a simple vector of ones. Additionally, we declare one array of integers `y_rep` and one array of real numbers `logrt_rep`. Both arrays store replicated responses and response times that are sampled from the posterior distribution using the distribution-specific random number generators `bernoulli_logit_rng` and `normal_rng`, using the sampled item and person parameters as inputs. These quantities are necessary to be able to conduct PPCs (Stan Development Team, 2021). Lastly, we declare a vector of length `Nobs` named `log_lik` that stores the log-likelihood calculated for each observation. The calculation is done in a loop over observations, where we first calculate the pointwise log-likelihoods for each outcome variable (`log_lik_y` and `log_lik_logrt`), and add them together to obtain the final pointwise log-likelihood values. This quantity is necessary for model fit evaluations and model comparisons after the sampling process (Stan Development Team, 2021). Please note that the quantity has to be named `log_lik` for further use with the `loo`-package (Vehtari et al., 2019). This procedure is model-specific, i.e. for the model extensions illustrated later in this tutorial, the calculation of the replicated data and the pointwise log-likelihood is based on the likelihood of the respective model (see Appendix C.1 and the specifications of the model extensions in this tutorial).

5.2.1 Code Execution in R

Actual estimation of the model is performed via calling the `stan` function of the `rstan` package, which can be seen in online supplement S1. The complete `stan` code as described above can be seen in online supplement S2. In the arguments of the `stan` function call, we specify the model file (with `.stan` extension), and the data object. We run the sampler with four chains on four cores simultaneously for greater computational efficiency. Users with less than four cores available on their machines, however, should set the `cores` argument accordingly. For all other settings we use `rstan's` default option, which means 2000 draws per chain, with 1000 draws as burn-in (Stan Development Team, 2021). After successful convergence, the results are extracted via the `summary` function. It only requires the `fit-`

Figure 22: *The Generated Quantities Block for the Hierarchical Framework Specification.*

```
generated quantities {
  corr_matrix[4] Omega_item;
  corr_matrix[2] Omega_person;
  cov_matrix[4] Sigma_item;
  cov_matrix[2] Sigma_person;
  array[Nobs] int y_rep;
  array[Nobs] real logrt_rep;
  vector[Nobs] log_lik;

  Omega_item = multiply_lower_tri_self_transpose(L_Omega_item);
  Omega_person = multiply_lower_tri_self_transpose(L_Omega_person);
  Sigma_item = quad_form_diag(Omega_item, tau_item);
  Sigma_person = quad_form_diag(Omega_person, [1,1]);
  y_rep = bernoulli_logit_rng(discrimination[kk] .*
        (ability[ii] - difficulty[kk]));
  logrt_rep = normal_rng(intensity[kk] - sensitivity[kk] .*
          speed[ii], sigma_e[kk]);

  for(n in 1:Nobs){
    vector[Nobs] log_lik_y;
    vector[Nobs] log_lik_logrt;
    log_lik_y[n] = bernoulli_logit_lpmf( y[n] | discrimination[kk[n]] *
                (ability[ii[n]] - difficulty[kk[n]]));
    log_lik_logrt[n] = normal_lpdf( logrt[n] | intensity[kk[n]] -
                sensitivity[kk[n]] * speed[ii[n]], sigma_e[kk[n]]);
    log_lik[n] = log_lik_y[n] + log_lik_logrt[n];
  }
}
```

object; it is, however, possible to manually specify the boundaries of the credibility interval (CI) via the `prob` argument. Depending on which parameter estimates are of interest, results can be easily extracted, regrouped, or reshaped.

5.2.2 Convergence Diagnostics and Model Fit Evaluation

Prior to interpreting the results of the Bayesian analysis, several characteristics of the MCMC sampler have to be checked. The most common characteristics are convergence, effective sample size, and the efficiency of the sampling process. Convergence is indicated by the Rhat diagnostic (Vehtari et al., 2020), which compares the between- and within-chain estimates for model parameters. Usually Rhat values smaller than 1.05 indicate acceptable agreement and convergence of the chains. `rstan` provides a function for visually checking the Rhat values for all model parameters (`stan_rhat`) that requires the stanfit-object as input. The *effective sample size* (ESS) is an indicator for uncertainty in the parameter estimates attributable to autocorrelations within the chains (Geyer, 2011). The ESS serves as an indicator for the number of independent samples with the same estimation power as the actually drawn autocorrelated samples. Thus, it measures the amount of independent information within the autocorrelated chains. ESS should be as large as possible. Different guidelines exist for the minimum ESS required to ensure trustworthy inference. Zitzmann and Hecht (2019) recommend an ESS for individual parameter estimates of $ESS > 400$. When assessing

the quality of the individual parameter estimates of the models in this tutorial, we adopt this criterion as well. Divergent transitions are also indicative of sampling inefficiencies, leading to small ESS. Betancourt (2016) illustrates how they depend on the curvature of the posterior distribution of a given parameter, and how they indicate deviations of the simulated Hamiltonian trajectory from the true trajectory. In other words, it is possible for the sampler to be stuck in regions of the posterior exhibiting low mass, from where it is difficult to get out. Consequently, the sampler spends a lot of time in regions where there is no information about the parameter, thus decreasing the ESS. Moreover, convergence problems and bias in the parameter estimates become more likely. Other diagnostics include the Bayesian Fraction of Missing Information and information about transitions saturating the maximum tree depth (E-BFMI; Betancourt, 2016; Stan Development Team, 2021). While the technical details of both diagnostics are beyond the scope of this tutorial, both diagnostics indicate that the sampler was not able to explore the posterior distribution adequately and efficiently. `rstan` provides applied researchers with several convenience functions that make convergence diagnostics very simple. For example, the `check_hmc_diagnostics()` function only requires the stanfit-object as input and checks for divergent transitions, transitions saturating the maximum tree depth, and the E-BFMI (see the example code in online supplement S1).

Another useful tool for performing convergence diagnostics is plotting the (marginal) posterior distributions. This can, for example, be done using the `bayesplot` package (Gabry & Mahr, 2022). We include an illustration of posterior plotting and posterior predictive checking using the `bayesplot` package in online supplement S1. We illustrate the case for the log response times, since posterior predictive checks are primarily useful for continuous data and distributions. The posterior predictive check uses the replicated responses and response times calculated in the generated quantities block (see Figure 22). The replicated data are extracted from the `stanfit` object, and then visually compared to the observed data. The closer the replicated distributions resemble the distribution of the original data, the better. For an extensive tutorial on plotting posterior distributions and performing posterior predictive checks, see Gabry et al. (2019).

Moreover, research questions often involve the evaluation of the goodness of fit of a model or comparisons of competing models. For instance, for checking whether data quality may be impeded by rapid guessing behavior, researchers may want to compare the basic hierarchical model with the mixture extension accommodating rapid guessing behavior illustrated later in this paper. To conduct model fit evaluations and comparisons it is first necessary to

calculate the pointwise log-likelihood for one or more given models (as illustrated in Figure 22). The pointwise log-likelihood can then be extracted using the `loo`-package to calculate, for instance, the *Widely Applicable Information Criterion* (WAIC; Vehtari et al. (2017)). Leave-One-Out cross-validation also requires the model specification to include the calculation of the pointwise log-likelihood. Additionally to the `loo`-package, the `rstan`-package also includes a `loo`-function. Both packages only require the `stanfit` object(s) to calculate model fit indices such as the *LOO-Information Criterion* (LOOIC; Vehtari et al. (2017)) or to compare different models. We include a simple example in online supplement S1. An extensive and more detailed tutorial on how model comparisons can be conducted using the `loo`-package, and how to properly interpret the information provided, can be found in Vehtari et al. (2017).

5.2.3 Empirical Example

To illustrate the basic hierarchical framework, we utilize the freely available PISA 2018 data set (OECD, 2019a). We use a subset of $I = 500$ test-takers from the Canadian sample and $K = 12$ items from a single booklet from the mathematical literacy domain. Polytomous items are dichotomized for reasons of simplicity. In the PISA 2018 study, response times are defined as time spent on item page. The data set is available in the R package `pisaRT` (Becker, 2020). The analysis can be reproduced using supplement S1 (`R` code) and `S2` (Stan code). With the chosen set-up, running the four chains in parallel, model estimation of the hierarchical framework required roughly 1.50 minutes.

Convergence diagnostics indicated no problems during the sampling process, thus warranting a substantial interpretation of the results. We found a negative correlation between speed and ability (-.50; 95% CI [-.59, -.39]) indicating that test-takers with lower ability tended to generate responses faster. Means, standard deviations, and correlations of the item parameters are given in Table 9, where, for instance, it can be seen that more difficult items also tended to be more time intensive. Note, however, that due to the small number of items the credibility intervals are rather broad and that with one exception no correlation between item parameters is credibly different from zero.

The stanfit-object also includes the estimates of the individual model parameters. Thus, it is possible to assess the properties and quality of individual items. Table 10 illustrates the discrimination, difficulty, time intensity and speed sensitivity parameters of the individual items, along with their 95% CIs. As can be seen, all items had acceptable discriminations and difficulties. For example, item 1 was very easy (with a difficulty of -2.62), while Item 11 was

Table 9: *Means, Standard Deviations, and Correlations of the Item Parameters of the Basic Hierarchical Response Time Model*

	$\ln(a)$	b	$\ln(\phi)$	λ
$\ln(a)$	0.37			
	[0.22, 0.59]			
b	.45	1.33		
	[.01, .82]	[0.87, 2.06]		
$\ln(\phi)$	.05	.44	0.33	
	[-.45, .55]	[-.03, .78]	[0.20, 0.54]	
λ	.39	.39	.39	0.54
	[-.15, .78]	[-.10, .75]	[-.10, .75]	[0.36, 0.83]
$\boldsymbol{\mu_K}$	0.19	-.22	-1.19	4.44
	[-0.07, 0.44]	[-1.03, 0.62]	[-1.38, -1.00]	[4.07, 4.75]

Note: $\ln(a)$: log item discrimination; b: item difficulty; $\ln(\phi)$: log speed sensitivity; λ: time intensity. Standard deviations and correlations are given in the diagonal and off-diagonal, respectively. 95% CIs are given in squared brackets.

very difficult (with a difficulty of 2.30). Item 11 also exhibited the highest discriminatory power. The average workload of the items ranged from 38.64 seconds (item 1) to 169.02 seconds (item 5), with their speed sensitivities ranging from 0.16 (item 1) to 0.47 (item 5). The complete `R` code to extract the individual item parameters can be seen in online supplement S1.

5.3 Model Extensions

5.3.1 Modeling the Difficulty-Distance Hypothesis for Non-Cognitive Data

The Basic Idea. The 3PLN model presented above is well suited for modeling response times for cognitive constructs, for example educational achievement testing. However, conceptual limitations arise when this model is applied to non-cognitive data, for example motivational constructs or attitudes. For non-cognitive constructs, response times and the focal trait are often assumed to be more closely intertwined. A common hypothesis is the distance-difficulty hypothesis, which states that the time spent on an item depends on the distance between the difficulty parameter of an item and the person trait. Conceptually, this means that persons who either strongly agree or disagree with a statement can quickly decide on a suitable response option, while persons for whom it is difficult to decide whether or not they agree with a statement need more time for their decision. The model by Ferrando and Lorenzo-Seva (2007) allows to explicitly address this hypothesis. Their model is very similar to the 3PLN model, except that it does not contain a slope parameter. Instead, it includes a

Table 10: *Individual Item Parameter Estimates of the Basic Hierarchical Response Time Model*

Item	a	b	λ	ϕ
1	0.85	-2.62	3.65	0.16
	[0.58, 1.16]	[-3.58, -1.96]	[3.61, 3.69]	[0.12, 0.2]
2	0.77	-1.27	3.67	0.22
	[0.53, 1.03]	[-1.82, -0.87]	[3.64, 3.71]	[0.18, 0.26]
3	1.19	-0.87	5	0.34
	[0.89, 1.51]	[-1.17, -0.63]	[4.95, 5.05]	[0.29, 0.38]
4	1.13	-1.46	4.15	0.25
	[0.86, 1.46]	[-1.87, -1.13]	[4.1, 4.19]	[0.2, 0.29]
5	1.18	1.34	5.13	0.47
	[0.85, 1.52]	[1.04, 1.72]	[5.07, 5.2]	[0.41, 0.53]
6	0.82	-0.78	4.29	0.33
	[0.58, 1.06]	[-1.15, -0.48]	[4.25, 4.34]	[0.29, 0.37]
7	1.56	-0.37	4.73	0.29
	[1.2, 1.99]	[-0.56, -0.2]	[4.68, 4.77]	[0.26, 0.34]
8	0.95	-0.17	3.96	0.43
	[0.69, 1.24]	[-0.41, 0.06]	[3.9, 4.01]	[0.38, 0.47]
9	1.6	0.52	4.39	0.33
	[1.21, 2.08]	[0.34, 0.7]	[4.35, 4.44]	[0.29, 0.37]
10	1.73	0.38	4.95	0.34
	[1.34, 2.22]	[0.22, 0.54]	[4.9, 4.99]	[0.3, 0.38]
11	2.19	2.3	4.83	0.33
	[1.37, 3.23]	[1.86, 2.94]	[4.77, 4.89]	[0.27, 0.39]
12	1.35	0.08	4.46	0.26
	[1.03, 1.7]	[-0.1, 0.25]	[4.42, 4.51]	[0.22, 0.3]

Note: a: item discrimination; b: item difficulty; λ: time intensity; ϕ: speed sensitivity. 95% CIs are given in squared brackets.

distance-difficulty parameter $\delta_{ik} = \sqrt{a_k^2(\theta_i - b_k)^2}$ that is determined by the person and item parameter estimates of the focal latent construct. The full response time model equation is

$$\ln RT_{ik} = \lambda_k - \zeta_i + \beta\delta_{ik} + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}), \tag{55}$$

where β is defined as a regression parameter that is constant across items. Note that when modeling the distance-difficulty hypothesis, the response time model cannot be estimated independently from the response model, as its item and ability parameters are required for the estimation of the latent speed factor. This also leads to conceptual differences regarding the interpretation of the speed dimension. Response time differences attributable to differences in the latent trait, for example a respondent answering quickly because all items are far from his/her trait level versus a respondent answering slowly because all items are close to his/her

trait level, are captured by β and not the latent speed dimension. This means that persons can have different observed response times but the same latent speed level.

What's Different? Compared to the basic hierarchical framework, there are three major differences in the model specification (see Figure 23). (a) The response time model does not include `phi`, the speed sensitivity parameter. Thus, the length of the item-related row-vectors and hyper-parameter vectors is reduced from four to three. (b) Moreover, because of the missing speed sensitivity parameter, the variance of the speed parameter does not have to be fixed to one for model identification. (c) The regression coefficient `beta` and the distance-difficulty parameters `delta` parameters are added. `beta` is equipped with a diffuse normal prior.

Consequently, in the parameters block, we additionally declare the regression coefficient `beta` as a single real-valued parameter. Furthermore, we explicitly declare a single real-valued scale parameter `tau_speed` with a lower bound of zero. In the transformed parameters block, we declare a vector of length `Nobs` containing the distance-difficulty parameters `delta`. Furthermore, we calculate the distance-difficulty parameters. In the model block, we first give `tau_speed` a half-Cauchy prior distribution, and then consider `tau_speed` when obtaining the covariance matrix for the multivariate normal distribution for the person parameters. Other than that, only the specification of the response time model as part of the likelihood differs. Here, the log response times are assumed to follow a normal distribution, governed by Equation 55.

Empirical Example. To illustrate the non-cognitive model suggested by Ferrando and Lorenzo-Seva (2007), we made use of the `extraversion` data set from the R package `diffIRT` (Molenaar, Tuerlinckx, & van der Maas, 2015). The data comprise binary item responses and response times for 146 subjects to 10 extraversion items, asking subjects to indicate whether adjectives such as active or noisy are applicable to their personalities. The analysis can be reproduced using supplement S3 (`R` code) and S4 (`Stan` code). With the chosen set-up, model estimation required approximately two minutes.

Again, convergence diagnostics indicated no problems during the sampling process, thus warranting a substantial interpretation of the results. In line with the difficulty-distance hypothesis, β was negative (-0.10; 95% CI [-0.14, -0.06]), indicating that response time decreases with an increasing distance between the item's difficulty and the person's location on the extraversion variable. The correlation between person variables was -.15, however, with

Figure 23: *The Parameters, Transformed Parameters and Model Blocks for the Non-Cognitive Extension of the Hierarchical Framework.*

```
parameters{
  row_vector[2] person[I];
  row_vector[3] item[K];
  vector[3] mu_item;
  real beta;
  vector<lower=0>[K] sigma_e;
  real<lower=0> tau_speed;
  vector<lower=0>[3] tau_item;
  cholesky_factor_corr[3] L_Omega_item;
  cholesky_factor_corr[2] L_Omega_person;
}

transformed parameters{
  vector<lower=0>[K] discrimination;
  vector[K] difficulty;
  vector[K] intensity;
  vector[I] ability;
  vector[I] speed;
  vector[Nobs] delta;

  discrimination = exp(to_vector(item[,1]));
  difficulty = to_vector(item[,2]);
  intensity = to_vector(item[,3]);
  ability = to_vector(person[,1]);
  speed = to_vector(person[,2]);

  delta = sqrt(square(discrimination[kk]) .*
    square(ability[ii] - difficulty[kk]));
}

model{
  target += lkj_corr_cholesky_lpdf(L_Omega_person | 1);
  target += cauchy_lpdf(tau_speed | 0,2);
  target += multi_normal_cholesky_lpdf(person | [0,0],
           diag_pre_multiply([1,tau_speed], L_Omega_person);

  target += lkj_corr_cholesky_lpdf(L_Omega_item | 1);
  target += normal_lpdf(mu_item | 0,5);
  target += cauchy_lpdf(tau_item | 0,2);
  target += multi_normal_cholesky_lpdf(item | mu_item,
        diag_pre_multiply(tau_item, L_Omega_item);
  target += cauchy_lpdf(sigma_e | 0,2);

  target += normal_lpdf(beta | 0,5);

  target += bernoulli_logit_lpmf( y | discrimination[kk] .*
      (ability[ii] - difficulty[kk]));
  target += normal_lpdf( logrt | intensity[kk] - speed[ii] +
      beta * delta, sigma_e[kk]);
}
```

a 95% CI of [-.40, .14] not credibly different from zero.

5.3.2 Modeling Conditional Dependence of Response Times and Accuracy

The Basic Idea. The assumption of stochastic local independence (all responses and response times are independent given the latent hierarchical structure of the model) incorporated in the basic hierarchical framework has received considerable criticism (e.g., Bolsinova, 2016; Bolsinova, Tijmstra, et al., 2017; Ranger & Ortner, 2012a). In this regard, Bolsinova, de Boeck, and Tijmstra (2017) propose to explicitly model conditional dependencies between responses and response times. Such conditional dependencies may, for instance, arise when-

ever there are heterogeneous (i.e., qualitatively different) response processes. Bolsinova, de Boeck, and Tijmstra (2017) propose to explicitly take into account conditional dependencies by letting the response model depend on the standardized residual response times z_{ik}, defined as

$$z_{ik} = \frac{\ln RT_{ik} - \lambda_k - \zeta_i}{\sigma_{\epsilon_k}}. \tag{56}$$

The parameters of the response model depend on z_{ik} as follows:

$$\ln(a_{ik}) = ln(a_{0k}) + \ln(a_{1k})z_{ik}, \tag{57}$$

and

$$b_{ik} = b_{0k} + b_{1k}z_{ik}. \tag{58}$$

The resulting response model is defined by Bolsinova, de Boeck, and Tijmstra (2017) as follows:

$$P(y_{ik} = 1|\theta_i, a_k, b_k) = \frac{\exp(a_{ik}\theta_i + b_{ik})}{1 + \exp(a_{ik}\theta_i + b_{ik})}. \tag{59}$$

Note that this parameterization deviates from Equation 51 as b_{ik} is not a difficulty but an easiness parameter. Also, b_{ik} and a_{ik} vary over persons, as they depend on the response time residuals. The parameters a_{0k} and b_{0k} are the baseline discrimination and easiness parameters, while the parameters a_{1k} and b_{1k} indicate how strongly the discrimination and easiness are influenced by z_{ik} for each item. If b_{1k} is positive, relatively fast responses are more often correct than slow responses (item easiness b_{ik} increases); if $a_{1k} > 1$, relatively fast responses contain less information about the ability in question (discrimination a_{ik} decreases). If $a_{1k} = 1$ and $b_{1k} = 0$, there are no conditional dependencies. Note that the response time model does not include the speed sensitivity parameter. The model can be used to answer research questions as: If a person takes more time on a specific item than expected, does this increase or decrease the probability of a correct response? And likewise, if a person takes more time on a specific item than expected, are such responses more strongly or weakly related to the person's ability?

What's Different? The following differences to the basic hierarchical framework are worth noting (see Figure 24). First, the response time model does not include `phi`. Second, because of the item parameters (`lnslope0`, `lnslope1`, `intercept0`, and `intercept1`), and because of moving the residual `sigma_e` into the joint multivariate distribution of the item parameters, the dimensions of the item-related row-vectors and hyper-parameter vectors are increased

from four to six. The variance of the speed parameter can be freely estimated. The most important difference to the basic hierarchical framework and the previous extension is found in the `model` block. Here, we first declare three local vectors, each of length `Nobs`, containing the standardized residual response times z_{ik} and the discrimination a_{ki} and easiness b_{ki} parameters of the response model. The hyperprior specification is the same as in the previous extension. In the loop over observations, we first specify the response time model, followed by the calculation of the standardized residual log response times and the discrimination and easiness parameters of the response time model as in Equations 56 to 59.

Empirical Example. We illustrate modeling conditional dependencies of response times and accuracy based on the PISA 2018 data set introduced earlier. The analysis can be reproduced using supplement S3 (`R` code) and S5 (`Stan` code). Model estimation required approximately seven minutes.

Again, no problems were encountered during the sampling process. Table 11 illustrates the core results of the analysis. First of all, we observe that the items have a moderate average baseline discrimination a_0 (1.16; 95% CI [0.94, 1.43]), with an average baseline easiness b_0 of 0.01 (95% CI [-0.87, 0.89]), implying a mean item easiness typical for PISA assessments. Second, the average effect of the residual log-response time on the slope a_{1k} is 0.95 (95% CI [0.83, 1.11]), while its average effect on the easiness b_{1k} is 0.16 (95% CI [-0.05, 0.36]). The credibility intervals, however, include one and zero, respectively. Thus, we can conclude that the residual log-response time has no effect on either the item discrimination or the item easiness. In other words, the informativeness of an item and the probability of a correct response do not depend on the relative speed of a test-taker on a specific item. Note, however, that due to the small number of items all credibility intervals are rather broad. The conditional dependence model, while quite robust in smaller samples, requires larger number of items for more conclusive results (Bolsinova, de Boeck, & Tijmstra, 2017).

5.3.3 Modeling Qualitative Differences in Response Behavior

The Basic Idea. Mixture extensions of the hierarchical framework are quite popular for modeling and investigating qualitative differences in response strategies (Molenaar et al., 2016; Ulitzsch et al., 2019a; C. Wang & Xu, 2015). We focus on the hierarchical latent response model for disengaged rapid guessing behavior by Ulitzsch et al. (2019a) to exemplify how such mixture extensions can be implemented. This mixture modeling approach allows for response processes to vary on the item-by-test-taker level, governed by an unobserved response

Figure 24: *The Parameters, Transformed Parameters, and Model Blocks for the Local Dependence Extension of the Hierarchical Framework.*

```
parameters{
  row_vector[2] person[I];
  row_vector[6] item[K];
  vector[6] mu_item;
  real<lower=0> tau_speed;
  vector<lower=0>[6] tau_item;
  cholesky_factor_corr[2] L_Omega_person;
  cholesky_factor_corr[6] L_Omega_item;
}

transformed parameters{
  vector[K] intensity;
  vector<lower=0>[K] sigma_e;
  vector[K] lnslope0;
  vector[K] lnslope1;
  vector[K] intercept0;
  vector[K] intercept1;
  vector[I] ability;
  vector[I] speed;

  intensity  = to_vector(item[,1]);
  sigma_e = sqrt(exp(to_vector(item[,2])));
  lnslope0 = to_vector(item[,3]);
  lnslope1 = to_vector(item[,4]);
  intercept0   = to_vector(item[,5]);
  intercept1   = to_vector(item[,6]);
  ability = to_vector(person[,1]);
  speed   = to_vector(person[,2]);
}

model{
  vector[Nobs] zki;
  vector[Nobs] aki;
  vector[Nobs] bki;

  target += lkj_corr_cholesky_lpdf(L_Omega_person | 1);
  target += cauchy_lpdf(tau_speed | 0,2);
  target += multi_normal_cholesky_lpdf(person | [0,0],
        diag_pre_multiply([1,tau_speed], L_Omega_person);

  target += lkj_corr_cholesky_lpdf(L_Omega_item | 1);
  target += normal_lpdf(mu_item | 0,5);
  target += cauchy_lpdf(tau_item | 0,2);
  target += multi_normal_cholesky_lpdf(item | mu_item,
        diag_pre_multiply(tau_item, L_Omega_item);

  for(n in 1:Nobs){
    target += normal_lpdf(logrt[n] | intensity[kk[n]] -
        speed[ii[n]], sigma_e[kk[n]]);
    zki[n] = (logrt[n] - (intensity[kk[n]] - speed[ii[n]])) / sigma[kk[n]];
    aki[n] = exp(lnslope0[kk[n]] + lnslope1[kk[n]] * zki[n]);
    bki[n] = intercept0[kk[n]] + intercept1[kk[n]] * zki[n];
    target += bernoulli_logit_lpmf(y[n] | aki[n] * ability[ii[n]] +
        bki[n]);
  }
}
```

process status Δ_{ik}. The model assumes different data-generating processes underlying item responses and response times associated with solution ($\Delta_{ik} = 1$) and rapid guessing behavior ($\Delta_{ik} = 0$). Item responses and response times stemming from solution behavior are modeled according to Equations 51 and 52. Probability correct for responses stemming from rapid guessing behavior is assumed to correspond to the guessing parameter c, while log response times are governed by a common normal distribution with mean μ_d and variance σ_d^2 that is

Table 11: *Means, Standard Deviations, and Correlations of the Item Parameters of the Conditional Dependence Model.*

	λ	σ_ϵ	a_{0k}	a_{1k}	b_{0k}	b_{1k}
λ	0.59					
	[0.40, 0.85]					
σ_ϵ	.15	0.42				
	[-.25, .52]	[0.29, 0.62]				
a_{0k}	.28	-.09	0.38			
	[-.14, .64]	[-.48, .33]	[0.22, 0.61]			
a_{1k}	.11	.31	-.06	.23		
	[-.32, .51]	[-.18, .69]	[-.53, .41]	[0.11, 0.39]		
b_{0k}	-.31	-.41	-.18	-.11	1.83	
	[-.64, .07]	[-.71, -.03]	[-.56, .24]	[-.52, .31]	[1.27, 2.61]	
b_{1k}	.18	.16	-.16	.38	.03	0.39
	[-.24, .56]	[-.27, .56]	[-.57, .27]	[-.09, .76]	[-.40, .45]	[0.24, 0.59]
$\boldsymbol{\mu_K}$	4.42	0.19	1.16	0.95	0.01	0.16
	[4.16, 4.70]	[0.16, 0.24]	[0.94, 1.43]	[0.83, 1.11]	[-0.87, 0.89]	[-0.05, 0.36]

Note: λ: time intensity; σ: residual variance of log response times; a_{0k}: baseline discrimination; a_{1k}: effect of the residual log-response time on the discrimination; b_{0k}: baseline easiness; b_{1k}: effect of the residual log-response time on the easiness. Standard deviations and correlations are given in the diagonal and off-diagonal, respectively. The last row contains the mean vector $\boldsymbol{\mu_K}$ of the joint item parameter distribution. 95% credibility intervals are given in square brackets.

unaffected by person or item characteristics. Further, the model incorporates the assumption that guessing generally requires less time than solution behavior by setting $\lambda_k = \mu_d + \lambda_k^*$ and constraining the time intensity offset parameters λ_k^* (indicating how much longer test-takers require to generate an engaged compared to a disengaged response) to positive values. Item-by-test-taker-specific mixing proportions $P(\Delta_{ik} = 1)$ are modeled employing an IRT model governed by test-taker's engagement ψ_i and the item's engagement difficulty ι_k, that is

$$P(\Delta_{ik} = 1|\psi_i, \iota_k) = \frac{\exp(\psi_i - \iota_k)}{1 + \exp(\psi_i - \iota_k)}. \tag{60}$$

The person $\boldsymbol{\psi}$ and item parameters $\boldsymbol{\iota}$ of this latent response model are considered in the joint distributions of person and item parameters, respectively. Note that as λ_k^* is required to be positive, $\ln \boldsymbol{\lambda}^*$ is considered in the joint multivariate normal distribution of item parameters.

What's Different? Figure 25 gives the parameters, transformed parameters, and model blocks implementing the code modifications required for specifying the mixture model for rapid guessing behavior by Ulitzsch et al. (2019a). In the parameters block, the dimensions of the arrays of person and item parameters (`person`, `item`), the Cholesky factors of the

Figure 25: *The Parameters, Transformed Parameters, and Model Blocks for the Mixture Extension of the Hierarchical Framework.*

```
parameters{
  row_vector[3] person[I];
  row_vector[5] item[K];
  vector[5] mu_item;
  vector<lower=0>[5] tau_item;
  real<lower=0> tau_engagement;
  vector<lower=0>[K] sigma_e;
  real<lower=0> sigma_d;
  real mu_d;
  real<lower=0,upper=1> guess;
  cholesky_factor_corr[3] L_Omega_person;
  cholesky_factor_corr[5] L_Omega_item;
}
transformed parameters{
  vector<lower=0>[K] discrimination;
  vector[K] difficulty;
  vector<lower=0>[K] sensitivity;
  vector[K] intensityStar;
  vector[K] engdifficulty;
  vector[I] ability;
  vector[I] speed;
  vector[I] engagement;

  discrimination = exp(to_vector(item[,1]));
  difficulty = to_vector(item[,2]);
  sensitivity = exp(to_vector(item[,3]));
  intensityStar = exp(to_vector(item[,4]));
  engdifficulty = to_vector(item[,5]);
  ability = to_vector(person[,1]);
  speed = to_vector(person[,2]);
  engagement = to_vector(person[,3]);
}
model{
  target += lkj_corr_cholesky_lpdf(L_Omega_person | 1);
  target += cauchy_lpdf(tau_engagement | 0,2);
  target += multi_normal_cholesky_lpdf(person | [0,0,0],
        diag_pre_multiply([1,1,tau_engagement], L_Omega_person);

  target += lkj_corr_cholesky_lpdf(L_Omega_item | 1);
  target += normal_lpdf(mu_item | 0,5);
  target += cauchy_lpdf(tau_item | 0,2);
  target += multi_normal_cholesky_lpdf(item | mu_item,
        diag_pre_multiply(tau_item, L_Omega_item);


  target += normal_lpdf(mu_d | 0,5)
  target += cauchy_lpdf(sigma_d | 0,2);
  target += cauchy_lpdf(sigma_e | 0,2);
  target += beta_lpdf(guess | 1,1);

  for(n in 1:Nobs){
    target += log_mix(1/(1 + exp(-engagement[ii[n]] +
                  engdifficulty[kk[n]])),
      bernoulli_logit_lpmf(y[n]|discrimination[kk[n]]*
        (ability[ii[n]]-difficulty[kk[n]])) +
      normal_lpdf(logrt[n]|mu_d+intensityStar[kk[n]]-sensitivity[kk[n]]*
        speed[ii[n]],sigma_e[kk[n]]),
      bernoulli_lpmf(y[n]|guess)+
      normal_lpdf(logrt[n]|mu_d,sigma_d));
  }
}
```

person and item correlation matrices (`L_Omega_person`, `L_Omega_item`), as well as the vector of item parameter means and standard deviations (`mu_item`, `sigma_item`) are adjusted to accommodate the fact that the model considers three types of person parameters and five types of item parameters in the respective joint distributions. Further, additional parameters for

the standard deviation of person engagement (`sigma_engagement`), the guessing parameter for guessed responses (`guess`), as well as the common mean and standard deviation of the disengaged log response time distribution (`mu_d`,`sigma_d`) are declared. The transformed parameters block again stores the item and person parameters in separate vectors. In the model block, all standard deviations are equipped with half-Cauchy priors, and an uninformative beta prior is employed for the guessing parameter. The key element of this modified code block is the employment of the `log_mix` function in the `model` block. This function can be used for specifying models with two mixture components. The first element (line 51) gives the mixing proportion, which is given by Equation 60. The second element (lines 52 and 53) gives the first component model, i.e., the models for engaged responses and response times, and the third element (lines 54 and 55) gives the second component model, i.e., the models for rapid guesses and the associated response times. Note that Stan does not estimate the unobserved engagement indicators Δ_{ik}, but solely provides engagement probabilities $P(\Delta_{ik} = 1)$, as it marginalizes over discrete parameters.

As becomes evident from this example specification, mixture extensions pose a powerful and easy-to-adapt tool for incorporating beliefs on qualitative differences in data-generating processes underlying responses and response times.

Empirical Example. Again, we illustrate the model based on the PISA 2018 data set. The analysis can be reproduced using supplement S3 (`R` code) and S6 (`Stan` code). Model estimation of the hierarchical framework required 45 minutes. Convergence diagnostics indicated no problems during the sampling process, thus warranting a substantial interpretation of the results. We found an average engagement probability (i.e. average $P(\Delta_{ik} = 1)$) of .95 (95% CI [.85, .99]), corresponding to an expected rate of disengaged responses of 5%. The probability to provide a correct response under rapid guessing was .09 (95% CI [.05, .14]). The correlation between speed and ability remained negative (-.23; 95% CI [-.35, -.09]), indicating that test-takers showing lower levels of ability when providing an engaged response tended to do so faster. Engagement was highly positively related to ability (.72; 95% CI [.50, .87]), indicating that test-takers approaching the test with higher levels of engagement tended to display higher levels of ability on items they solved in an engaged manner, and showed a weakly negative association with speed (-.23; 95% CI [-.47, .02]), indicating that test-takers approaching the test with higher levels of engagement tended to take more time on items they solved in an engaged manner. The joint distribution of item parameters is summarized in Table 12, where it can be seen that items with higher difficulty and items requiring more

time to be solved in an engaged manner exhibited higher engagement difficulties. However, again, credibility intervals for item parameter correlations were very broad.

Table 12: *Means, Standard Deviations, and Correlations of the Item Parameters of the Rapid Guessing Mixture Model.*

	$\ln(a)$	b	$\ln(\phi)$	$\ln(\lambda^*)$	ι
$\ln(a)$	0.41				
	[0.23, 0.71]				
b	.47	1.47			
	[.00, .81]	[0.96, 2.30]			
$\ln(\phi)$	-.07	.25	0.25		
	[-.57, .45]	[-.24, .67]	[0.14, 0.45]		
$\ln(\lambda^*)$	.38	.36	.18	0.79	
	[-.10, .76]	[-.12, .73]	[-.34, .62]	[0.46, 1.39]	
ι	.03	.32	.29	.40	1.35
	[-.43, .52]	[-.18, .74]	[-.25, .77]	[-.12, .79]	[0.67, 2.37]
$\boldsymbol{\mu_K}$	0.10	-.33	-1.51	-0.15	-9.41
	[-0.17, 0.35]	[-1.22, 0.53]	[-1.68, -1.34]	[-0.78, 0.31]	[-11.08, -6.38]

Note: $\ln(a)$: log item discrimination; b: item difficulty; $\ln(\phi)$: log speed sensitivity; λ^*: time intensity offset; ι: engagement difficulty. Standard deviations and correlations are given in the diagonal and off-diagonal, respectively. The last row contains the mean vector $\boldsymbol{\mu_K}$ of the joint item parameter distribution. 95% credibility intervals are given in square brackets.

5.4 Discussion

This tutorial illustrated the implementation of the basic hierarchical framework of van der Linden (2007) in the Bayesian software Stan and showcased its flexibility and ease of adaption on the basis of three extension. The flexibility of the basic hierarchical framework makes it easy to develop such extensions, facilitating further insights into the relationship between accuracy (or non-cognitive constructs) and speed, into how individuals approach assessment situations, and how they allocate their cognitive resources during the response process. The extensions illustrated in this tutorial are by no means exhaustive. For example, Bezirhan et al. (2021) utilize the basic hierarchical framework to analyze item-revisiting behavior in high-stakes testing by adding the Rasch Poisson Counts Model (RPCM, Wright & Masters, 1982) measuring the number of revisits and relating them to response times and accuracy. Other sources of information to explain response behavior are possible. The flexibility of the basic hierarchical framework is accompanied by the flexibility of its implementation in Stan, which oftentimes requires adaptations of only a few lines of code that mirrors adaptations of the model.

As already mentioned, response time models based on the lognormal distribution are used very frequently. Alternative distributions such as the gamma distribution, the ex-Gaussian distribution, or the Weibull distribution may also be employed. With the Bayesian hierarchical approach outlined in this tutorial it is easy to use other distributions than the lognormal by simply changing the likelihood of the respective component model within the hierarchical framework. Furthermore, Stan offers the possibility for researchers to build custom distributions and use them in their modified hierarchical framework.

As many of the latest developments in response time modeling draw on Bayesian modeling techniques (e.g., J. Lu et al., 2021; Sinharay & Johnson, 2019), this tutorial aimed at providing readers with an introduction to the latest developments with regard to Bayesian hierarchical models. The Bayesian hierarchical approach outlined here offers a unique and flexible framework for estimating recent response time models, but implementation can be time consuming and cumbersome for researchers who have little background in using Bayesian estimation techniques. For applied researchers already familiar with lme4 or general multilevel syntax, the `brms`-package (Bürkner, 2018, 2021) is a valuable alternative for implementing, for example, the basic hierarchical framework of van der Linden (2007). In the online repository accompanying this tutorial, we include a `brms` implementation of the basic hierarchical framework (see online supplement S9). Since `brms` is a kind of general-purpose software, and the model specification is not tailored towards response time modeling, sampling from the joint posterior distribution takes considerably longer than with the model specification illustrated in this tutorial. Moreover, the implementation of model adaptions may oftentimes be markedly less straightforward. Posterior predictive checking and model fit assessment is, however, easier with `brms`, since the fit-objects can be directly used to calculate the WAIC criterion or to conduct leave one out cross validation (Vehtari et al., 2017) with built-in functions, and without modifications to the model code.

The model specifications in this tutorial reflect our latest knowledge with regard to weakly informative hyperprior specifications (Gelman, 2006; König et al., 2020; Polson & Scott, 2012). The utility of the modeling approach lies in the fact that the models illustrated in this tutorial, which are arguably quite complex with large numbers of parameters, can be estimated with relatively small test lengths and sample sizes. To increase the power to detect, for instance, dependencies between accuracy and response times further, informative prior distributions would be required. This additional information could stem, for example, from pre-calibrated item parameters and their standard errors (that is where the actual

information is introduced) that can be used for model estimation. This would, however, lead to distinct changes in the model specification. First, a hierarchical prior structure for the items would not be necessary in the case that item parameters for both the 2PL and 3PLN model are available. Second, when only item parameters for either the 2PL or the 3PLN model are available, the dimensions of the multivariate normal distribution for the item parameters would reduce; moreover, only the grand means and standard deviations of the item discriminations and difficulties (or time intensities and sensitivities) have to be sampled. Third, although seldom the case, it is possible that researchers have information about the grand means and standard deviations of certain item parameters. This would imply removing the grand means and standard deviations as actively sampled parameters, and putting the actual means and standard deviations into the sampling statement of the multivariate normal distribution. A last possibility to utilize informative prior distribution without changing the hierarchical structure is related to the `beta` parameter in the non-cognitive extension. Prior information could come from previous studies on similar samples about the difficulty-distance hypothesis in order to update the estimate and to get more certainty about its magnitude.

5.4.1 Concluding Remarks

To conclude, with this accessible tutorial we hope that we have given non-specialists and applied researchers a better understanding of the use and utility of the basic hierarchical response time framework and Bayesian hierarchical modeling of response time models. Due to its modular nature, it is easy to adapt the framework to researchers' needs, as illustrated by the three extensions included in our tutorial. We hope that the tutorial furthers the application of innovative response time models for answering novel substantive research questions in various assessment contexts.

6 Automated Test Assembly in R: The eatATA Package

Published as: Becker*, B., Debeer*, D., Sachse, K. A., & Weirich, S. (2021). Automated Test Assembly in R: The eatATA Package. *Psych, 3*(2), 96–112. https://doi.org/10.3390/psych3020010[28]

*Both authors contributed equally to the manuscript.

Abstract: Combining items from an item pool into test forms (test assembly) is a frequent task in psychological and educational testing. Although efficient methods for automated test assembly exist, these are often unknown or unavailable to practitioners. In this paper we present the `R` package `eatATA`, which allows using several mixed integer linear programming solvers for automated test assembly in `R`. We describe the general functionality and the common work flow of `eatATA` using a minimal example. We also provide four more elaborate use cases of automated test assembly: (a) The assembly of multiple test forms for a pilot study; (b) the assembly of blocks of items for a multiple matrix booklet design in the context of a large-scale assessment; (c) the assembly of two linear test forms for individual diagnostic purposes; (d) the assembly of multi-stage testing modules for individual diagnostic purposes. All use cases are accompanied with example item pools and commented `R` code.

[28]The indices for persons and items have been adapted to the notation of this thesis.

6.1 Theoretical Background

In psychological or educational testing, assembling test forms from an existing item pool is a frequent challenge (van der Linden, 2005). Sometimes, a single test form is constructed, which has to be maximally informative for a certain classification decision. Sometimes, test security is a concern and therefore multiple, parallel test forms are created to prevent test-takers from answer-copying and from sharing test content across test sessions (Luecht & Sireci, 2011). In large-scale assessments, multiple test forms are used to cover a broader range of test content and to increase measurement precision (Kuhn & Kiefer, 2015; OECD, 2019b). In *multi-stage testing* (MST), test modules with different ability target groups are created (Yan et al., 2016), whereas in *computer adaptive testing* (CAT), after every item, a new item is added to the test (van der Linden & Glas, 2000). Because assembling test forms by hand can be cumbersome and error prone, *automated test assembly* (ATA) methods have been developed, which rely on mathematical programming techniques such as *mixed integer linear programming* (MILP). In many, if not all cases where items should not just be randomly selected, ATA can help, and will likely lead to better solutions than manual test assembly.

In ATA, test specifications, similar to the number of items per test form or item type distributions across test forms, are formulated as mathematical constraints. The optimization goal (for instance, maximizing test information at a certain ability level) is formulated as a mathematical objective function. Mathematical programming solvers can be used to find an optimal solution for the given combination of mathematical constraints and objective, meaning that the optimal item-to-test-form assignment for the test assembly problem can be found. A general introduction to automated test assembly is, for example, available in the work of van der Linden (2005).

Unfortunately, in practice, ATA approaches are not utilized as often as they could. Due to conceptual and technical barriers, practitioners frequently opt for manual trial and error approaches instead, leading to sub-optimal solutions. With the `R` package `eatATA` (*educational assessment tools: Automated Test Assembly*) and this tutorial we try to give easy access to ATA to more practitioners. The paper is structured as follows: First, we give a short introduction to why an `R` package is suitable in this context. We then give an overview of the functionalities of `eatATA` and which solvers are accessible via the package, and illustrate the general work flow when using `eatATA` for automated test assembly with a minimal example. Subsequently, we provide four practical use cases alongside detailed and commented `R` code

to illustrate the package functionality in depth.

6.2 eatATA

In the context of psychological and educational testing, `R` (R Core Team, 2023) is a common tool for psychometric and statistical analyses. `R` is an open source and free software environment and its extensive and actively maintained libraries offer tools for a rich diversity of data analysis use cases. Furthermore, a variety of mathematical programming solvers are available through `R`, including both open source and commercial solvers. These solvers can usually be accessed via packages that function as APIs (*application programming interfaces*) to a specific solver, and the solver itself is often included directly in the respective package (e.g., `lpSolveAPI` is an API to the solver `lpSolve`). This, in principal, enables researchers to use `R` for test assembly purposes. For example, a short tutorial on how ATA can be used in `R` using `lpSolveAPI` (Konis & Schwendinger, 2020) can be seen Diao and van der Linden (2011). However, while such an implementation is possible, the translation of test specifications into mathematical constraints can be an interesting challenge for some, but a cumbersome task for others.

The mathematical programming solvers and APIs are often very flexible and applicable to a wide range of optimization problems that go far beyond ATA. Although valuable, this flexibility also increases the complexity for users, especially because the APIs, the solvers, and the available documentation are generally not targeting ATA applications. Hence, for researchers without a background in mathematical programming, applying the solvers to ATA problems is far from straightforward. In addition, there exist considerable differences between the APIs of different solvers. Therefore, when an educational measurement practitioner has invested the time to become familiar with one specific solver API, switching to a different solver will likely require an additional effort. From our experience, we believe that there are two main reasons why ATA methods are highly underused in practice: First, a lot of practitioners are not aware that there are efficient assembly techniques that could be helpful in their daily practice. Second, practitioners that are aware of the efficient assembly techniques often lack the time to work out all the operational details.

Through the `R` package `eatATA` and this tutorial paper we want to promote ATA methods and provide easier access to ATA for measurement practitioners. The package facilitates the access to mathematical programming and its potential for ATA-problems without worrying about how test specifications are formulated mathematically. In the spirit of the `R` program-

ming language, the functionality of the package is based on functions. Every test specification can be expressed by a function, thereby enabling a work flow in `R` that will feel familiar for practitioners with `R` experience.

Available solvers within the package are `GLPK` (Makhorin, 2018), `lpSolve` (Berkelaar et al., 2016), `SYMPHONY` (Ladanyi et al., 2019), and `Gurobi` (Gurobi Optimization, LLC, 2021a). These solvers are used via the R package APIs `Rglpk` (Theussl & Hornik, 2019), `lpSolve` (Berkelaar & Csárdi, 2020), `Rsymphony` (Harter et al., 2020), and `gurobi` (Gurobi Optimization, LLC, 2021b). For a general overview of different available open source and commercial MILP solvers see, for example, Donoghue (2015) and Luo (2020).

6.2.1 Work Flow

Regardless of the context of a specific test assembly problem, the common challenge when assembling test forms is that a variety of requirements for the resulting test form(s) have to be fulfilled, otherwise known as the test specifications (van der Linden, 2005). The schematic work flow of automatically assembling test forms and incorporating the required test specifications using `eatATA` is the following:

1. Item Pool: A `data.frame` including all information on the item pool is loaded or created. If the items have already been calibrated (e.g., based on data from a pilot study) this will include the calibrated item parameters.

2. Test Specifications: Usually a combination of: (a) Typically one objective function and (b) multiple constraints.

   (a) Objective Function: Usually a single object corresponding to the optimization goal, created via one of the *objective function* functions. This refers to a test specification where we have no absolute criterion, but where we want to minimize or maximize something.

   (b) Further Constraints: Further constraint objects, created using various constraint functions. These refer to test specifications with a fixed value or an upper and/or lower bound.

3. Solver Call: The `useSolver()` function is called using the constraint objects to find an optimal solution.

4. Solution Processing: The solution can be inspected using the `inspectSolution()` and `appendSolution()` functions.

To illustrate this general work flow, we provide a small illustrative example. More complex extensions are discussed in the specific use cases later in the paper.

6.2.2 Minimal Example

In this minimal example, the goal is to assemble a single test form with maximum *test information function* (TIF) for a medium ability level. Furthermore, the test form should consist of exactly ten distinct items and have an average test time of approximately 8 min. The complete R syntax for this example can be found in Supplement S0 and the mathematical formulation of the test assembly problem is described in Appendix D.1. Further details regarding mathematical formulations of MILP problems and how MILP solvers operate can be found in van der Linden (2005).

(1) Item Pool. For the illustrative example, we use a small simulated item pool of 30 items, which is included in the `eatATA` package (`items_mini`). In general, item pool information should be stored in a single `data.frame` with each row representing an item. In the example item pool, items are characterized by their format ("`format`"), average response times ("`time`"), and a difficulty parameter ("`difficulty`"), based on a calibration according to a Rasch model (Rasch, 1960). To calculate the *item information function* (IIF) we use the `calculateIIF()` function (see Figure 26). Alternatively, the `calculateIIF()` function could be used to calculate the IIF for the item parameters from the 2 and 3 parameter logistic models[29]. We provide the item parameters and one or multiple ability points (`theta`) at which the item information function should be calculated. In our case, we are interested only in the information function at a medium ability, so we set `theta = 0` and append the IIF to our item pool `data.frame`. The resulting first five rows of the item pool can be seen in Table 13.

Figure 26: *Calculate Item Information Function.*

```
items_mini$IIF_0 <- calculateIIF(B = items_mini$difficulty, theta = 0)
```

(2a) Objective Function. As a first object, we define the objective function. This corresponds to a test specification where we have no absolute criterion that needs to be exactly fulfilled, but where the goal is to minimize or maximize something. This can vary greatly

[29] In principal, any response model can be used within `eatATA` and there exist various `R` packages to calculate IIFs for a wide range of response models.

depending on the goal of the test assembly. Common optimization goals include maximizing the TIF of a test form, minimizing test time or test length, or minimizing differences in TIF between test forms.

Table 13: *First Five Items of the Simulated Item Pool.*

Item	Format	Time	Difficulty	IIF_0
1	mc	27.79	-1.88	0.11
2	mc	15.45	0.84	0.45
3	mc	31.02	1.12	0.33
4	mc	29.87	0.73	0.50
5	mc	23.13	-0.49	0.61

In our example, we seek to maximize the TIF at a medium ability level using the `maxObjective()` function (see Figure 27). This is achieved via maximizing the sum of the IIFs of the items in the test form[30]. Note that item identifiers should be supplied to all objective function and constraint functions. This guarantees that all constraints relate to the same set of items and provides a more readable solver output. Other available functions for defining optimization goals are: `minObjective()`, `maximinObjective()`, `minimaxObjective()`, and `cappedMaximinObjective()`.

Figure 27: *Define Objective Function: Maximize Item Information Function at Average Ability.*

```
testInfo <- maxObjective(nForms = 1, itemValues = items_mini$IIF,
                         itemIDs = items_mini$item)
```

(2b) Constraints. In the next step, we translate our further test specifications for the test assembly into function calls. These constraints are not optimization goals but specifications with fixed target values or upper and/or lower bounds. For this, we create multiple constraint objects (see Figure 28). In our example we want to fix the number of items in the test form to exactly ten items, which is performed by the `itemsPerFormConstraint()` function. This function specifically serves the purpose of setting the test length for the test assembly. Using the `operator` and the `targetValue` arguments we can set a fixed target value or an upper or lower bound. The total test time is constrained to approximately eight minutes using the `itemsValuesDeviationConstraint()` function. This function belongs to a family of functions that can be used to set constraints using numerical item values. By setting

[30] As a Rasch model has been used for calibration, maximizing the TIF at ability level 0 corresponds to minimizing the difference of the average item difficulty from 0.

`allowedDeviaton = 5` we allow the testing time to vary between 7 min and 55 s and 8 min and 5 s. Note that setting the test time to be exactly eight minutes would be overly restrictive and not necessary from a practical stand point. An overview over the available constraint functions and their functionality can be found here: https://CRAN.R-project.org/package=eatATA/vignettes/overview.html.

Figure 28: *Define Constraints: Number of Items in the Test Form, Number of Times an Item can be Used, and Total Average Testing Time.*

```
# Number of items (test length)
itemNumber <- itemsPerFormConstraint(nForms = 1, operator = "=",
                                     targetValue = 10,
                                     itemIDs = items_mini$item)

# Test time
testTime <- itemValuesDeviationConstraint(nForms = 1,
                                          itemValues = items_mini$time,
                                          targetValue = 8 * 60,
                                          allowedDeviation = 5,
                                          relative = FALSE,
                                          itemIDs = items_mini$item)
```

(3) Solver Call. Finally, we collect all constraint objects defined above in a `list` and hand these to the solver of our choice via the `useSolver()` function (see Figure 29). The order in which the constraints (including the objective function) are created or ordered does not have any impact on the solution of the test assembly problem.

As complex test assembly problems with large numbers of possible solutions can lead to long computation times, it is often reasonable to set a time limit for the solver via the `timeLimit` argument. In cases where the time limit is reached and where at least one feasible solution is found, but the search of the total solution space is incomplete, the function returns the best available solution. In most practical applications, the quality of this solution will be absolutely sufficient. As this illustrative example is a very simple ATA problem for which `GLPK` finds a solution almost instantly, it is not necessary to set a time limit for the solver. The solver used for ATA can be specified via the `solver` argument, with "`GLPK`" being the default.

Figure 29: *Solve MILP problem.*

```
solver_out <- useSolver(list(itemNumber, testTime, testInfo),
                        solver = "GLPK")
```

Note that sometimes combinations of constraints can lead to infeasibility issues. That is, it is possible that for a given set of test specifications for a specific item pool no feasible solution

exists. If this is the case, `useSolver()` will issue a corresponding message. To identify which (combination of) constraints causes the infeasibility, it can be helpful to remove constraints from the ATA problem step by step until feasibility is achieved. Alternatively, constraints can be added to the ATA problem step by step starting with just the objective function until the problem becomes infeasible (Spaccapanico Proietti et al., 2020).

(4) Solution Processing. `eatATA` provides two functions to process the output of `useSolver()`: `inspectSolution()` to directly view the assembled test forms presented in a list that only contains the items in the assembled test form(s), and `appendSolution()`, which appends the assignment matrix containing 0 (item not in this test form) and 1 (item in this test form) to the item pool `data.frame` (see Figure 30).

Figure 30: *Inspect the Solver Solution and Append it to the Item Pool `data.frame`.*

```
inspectSolution(solver_out, items = items_mini, idCol = "item")
item_mini_out <- appendSolution(solver_out, items = items_mini,
                                idCol = "item")
```

6.3 Use Cases

In the following section, we present four different applications of the `eatATA` package to test assembly problems. The use cases were chosen to cover a broad range of contexts for ATA application: (1) A pilot study setting in which we assemble multiple test forms while depleting the item pool (without prior item calibration), (2) a typical large-scale assessment situation, in which calibrated items are assembled to blocks for a multiple matrix booklet design, (3) the assembly of multiple parallel test forms for a high-stakes assessments from a calibrated item pool, and (4) the assembly of modules from a calibrated item pool for a multi-stage assessment. To illustrate the accessibility of `eatATA` compared to plain solver API's use case (3) and (4) correspond to two of the problems used in the tutorial paper by Diao and van der Linden (2011). Because the solver calls and the solution processing do not differ much between the minimal example and the different use cases, we primarily focus on how the constraint and objective function definitions have to be altered from application to application. Complete syntaxes for all use cases can be found in the corresponding supplementary files at https://www.mdpi.com/article/10.3390/psych3020010/s1.

6.3.1 Pilot Study

Usually, when conducting a pilot study, little is known about the empirical characteristics of the item pool. Instead, the goal of a pilot study is to gather such information (e.g., response times, and missing rates) and calibrate the items. Hence, the test specifications for pilot studies often deviate substantially from test assembly specifications for operational tests. For this use case, we use a simulated item pool `items_pilot`, which is included in the `eatATA` package. The item pool consists of 100 items with various characteristics, for example the expected response times in seconds ("`time`"), the item format ("`format`"), and a rough estimate of the item difficulty ("`diffCategory`"), grouped into five categories. The first five items of the item pool can be seen in Appendix D.2. From this item pool, we want to assemble test forms that meet the following requirements: (1) each item should appear in exactly one test form (this implies no item overlap between test forms), (2) all items should be used (*item pool depletion*), (3) the expected test form response times should be as close to 10 minutes as possible, (4) the number of test forms should be determined accordingly, (5) item difficulty categories and items formats should be distributed as evenly as possible across test forms, (6) each content domain should be at least once in each test form, and (7) item exclusions should be incorporated.

The definition of the objective function and all constraints can be seen in Figure 31. Among the test specifications listed above, (3) is the specification most suitable for formulation as an objective function. This means that we want to optimize the test takers' mean test taking time and keep it as close as possible to 10 min per test form. In order to achieve this, we first transform the expected item response times to minutes. Then we calculate the ideal number of test forms by dividing the sum of all expected item response times (which is 74 min) by ten. As we prefer test forms below our target test time to test forms above our target test time, we choose the next integer above via the `ceiling()` function, resulting in eight test forms. The actual objective function is defined via the `minimaxObjective()` function, which allows us to specify a `targetValue`. The maximum difference of test form times from this target value is then minimized.

Second, we implement test specifications (1) and (2) (item pool should be depleted and no item overlap) as constraints using a single function call to `itemUsageConstraint()`. The operator argument is set to "`=`" which means that every item will occur exactly once across all test forms. Test specification (5) refers to the difficulty column "`diffCategory`" as well as to the item format column "`format`". For item difficulty, we define the column

Figure 31: *Define Constraints for Pilot Study Test Assembly.*

```
# Determine number of test forms
items_pilot$time_in_min <- items_pilot$time / 60
nForms   <- ceiling(sum(items_pilot$time_in_min) / 10 )

# Objective function (response times)
timeCons <- minimaxObjective(nForms = nForms,
            itemValues = items_pilot$time_in_min,
            targetValue = 10, itemIDs = items_pilot$item)

# Item pool depletion
noItemOverlap <- itemUsageConstraint(nForms, targetValue = 1,
                 operator = "=", itemIDs = items_pilot$item)

# Difficulty and format
items_pilot$diffCategory <- as.factor(items_pilot$diffCategory)
equal_diff <- autoItemValuesMinMaxConstraint(nForms = nForms,
                              itemValues = items_pilot$diffCategory,
                              itemIDs = items_pilot$item)
equal_format <- autoItemValuesMinMaxConstraint(nForms = nForms,
                 itemValues = items_pilot$format,
                              itemIDs = items_pilot$item)

# Content categories
domainCons <- itemCategoryMinConstraint(nForms = nForms,
              itemCategories = items_pilot$domain,
              itemIDs = items_pilot$item, min = c(1, 1, 1))

# Exclusions
exclusionTuples <- itemTuples(items_pilot, idCol = "item",
                   infoCol = "exclusions", sepPattern = ",␣")
excl_constraints <- itemExclusionConstraint(nForms = nForms,
                   itemTuples = exclusionTuples,
                   itemIDs = items_pilot$item)
```

"`diffCategory`" to be a factor variable, as we do not want the numerical mean value to be equal across test forms but the distribution of distinct difficulty levels. We use the function `autoItemValuesMinMaxConstraint()` to determine the required `targetValues` automatically, after which the function directly calls the respective constraint functions using the calculated `targetValues`. By default, the function returns the resulting minimum and maximum levels. For example, for item difficulty, items of difficulty category 1 will occur once or twice in each test form. Alternatively, for item formats, the "`cmc`" format will occur four or five times in each test form. Test specification (6) requires that each domain occurs at least once in each test form. Using the `itemCategoryMinConstraint()` and the `min` argument, we define for each of the three categories (levels) of `domain` ("`listening`", "`reading`", "`writing`") the minimum occurrence frequency.

Finally, we implement test specification (7), the item exclusion constraints that are captured in the "`exclusions`" column of the `items_pilot data.frame`. The column contains item exclusions as a single character string for each item. The items in the data set have either no exclusions (`NA`), only one exclusion (e.g., "`76`"), or multiple exclusions (e.g., "`70, 64`"). As there are items with multiple exclusions, we need to separate the string into discrete item identifiers via the function `itemTuples()`, which produces pairs (*tuples*) of exclusive items

(also called *enemy items*). Using the `sepPattern` argument in the `itemTuples()` function, the user must specify the pattern, which separates the item identifiers within the string. These tuples can be used to define exclusion constraints in the `itemExclusionConstraint()` function. The complete code for the pilot study use case, including the solver call and the solution inspection, can be seen in Supplement S1.

6.3.2 LSA Blocks for Multiple Matrix Booklet Designs

Many *large-scale assessment* (LSA) test forms (typically referred to as booklets) consist of multiple item blocks (also referred to as clusters). In the test assembly process, first, item blocks are assembled from the item pool and later these item blocks are combined into test forms according to so called multiple matrix booklet designs (Gonzalez & Rutkowski, 2010). Examples of this approach can be found in the PISA studies (OECD, 2019b) or are described by Kuhn and Kiefer (2015) for the Austrian Educational Standards Assessment. The present use case illustrates the first step - assembling test items to eight item blocks that fit in multiple matrix booklet designs. The second step - combining item blocks to test forms – is currently beyond the scope of `eatATA` and the reader is referred to the literature on booklet designs (Frey et al., 2009; Pokropek, 2011).

For this purpose, we assume that a pilot study has been conducted and that all required parameter estimates from an item calibration are available. We use a simulated item pool of 209 items with typical properties, which is included in the `eatATA` package (`items_lsa`). The first 10 items of this item pool can be seen in Appendix D.3. The assembled item blocks should conform to the following test specifications: (1) blocks should contain as many well-fitting items as possible, (2) hierarchical stimulus item structures should be incorporated, (3) no item overlap, (4) a fixed set of anchor items has to be included in the block assembly[31], (5) the average item block times should be around 20 min, (6) difficulty levels should be distributed evenly across item blocks, (7) all blocks should contain at least three different item formats, and (8) maximally two items per block should have an average proportion of correct responses below 8 or above 92 percent.

The definition of the objective function and all constraints can be seen in Figure 32. Test specification (1) is chosen as the objective function. The infit (weighted MNSQ) is among the most widely used diagnostic Rasch fit statistics (OECD, 2014) and can be found in col-

[31] If LSAs intend to measure trends between different times of measurement, new assessment cycles partially reuse items from former studies, so-called *anchor items*, to establish a common scale (Kolen & Brennan, 2014). Usually, anchor items are chosen beforehand based on their advantageous psychometric properties.

umn "`infit`". As we are only interested in absolute deviations from 1 (otherwise positive and negative deviations could cancel each other out) we create a new variable, `infitDev`. The deviation of this variable from 0 is then minimized using the `minimaxObjective()` function.

Figure 32: *Define Constraints for LSA Test Assembly.*

```
# Objective function (infit)
infitDev <- abs(items_lsa$infit - 1)
infitCons <- minimaxObjective(nForms = 8, itemValues = infitDev,
                              targetValue = 0, itemIDs = items_lsa$item)

# Shared stimuli
incluTup <- stemInclusionTuples(items_lsa, "item", "testlet")
incluCons <- itemInclusionConstraint(nForms = 8, itemTuples = incluTup,
                                     itemIDs = items_lsa$item)

# Item overlap
overlapCons <- itemUsageConstraint(nForms = 8, targetValue = 1,
                                   operator = "<=",
                                   itemIDs = items_lsa$item)

# Anchor items
anchorCons <- itemUsageConstraint(nForms = 8, targetValue = 1,
                            operator = "=",
                            whichItems=items_lsa$item[items_lsa$anchor==1],
                            itemIDs = items_lsa$item)

# Block times
timeCons <- itemValuesDeviationConstraint(nForms = 8,
                               itemValues = items_lsa$time,
                               targetValue = 1170, allowedDeviation = 150,
                               relative = FALSE, itemIDs = items_lsa$item)
# Difficulty
diffLevels <- as.factor(items_lsa$level)
levelCons <- itemCategoryMinConstraint(nForms = 8, diffLevels,
                             itemIDs = items_lsa$item,
                             min = c(1,2,2,1))

# Item format
formatLv <- as.factor(ifelse(grepl("complex",items_lsa$format)|
                         grepl("matching",items_lsa$format),
                         "mix",ifelse(grepl("open",items_lsa$format)|
                                        grepl("sentence",items_lsa$format),
                           "open", "closed")))
formatCons <- itemCategoryMinConstraint(nForms = 8, formatLv,
                              itemIDs = items_lsa$item,  min = c(2,2,2))

# Proportion correct
freqLv <- as.factor(ifelse(items_lsa$frequency > .92, "above",
                      ifelse(items_lsa$frequency < .08, "below", "okay")))
freqCons <- itemCategoryMaxConstraint(nForms = 8, freqLv,
                            itemIDs = items_lsa$item,
                            max = c(2, 2, 200))
```

Test specification (2) is a common challenge for cognitive tests in LSAs, where items are usually not distinct units. Instead, multiple items share a common stimulus (e.g., a text, a picture or an auditive stimulus). Such item sets are often called *testlets*. In general, testlet structures can be dealt with in different ways: (a) In the assembly, testlets can be treated as fixed structures and used as the actual units in the test assembly, (b) testlet structures can be incorporated using fixed inclusion constraints (e.g., whenever item A is chosen, items B and C that belong to the same stimulus have to be chosen, too), (c) hierarchical structures can

be incorporated in the test assembly (see chapter 7 in van der Linden, 2005)). In the `eatATA` package, options (a) and (b) are implemented and option (b) is chosen for this specific use case. Option (b) can indeed be implemented very similarly to the item exclusion constraints that were introduced in the pilot study use case. Inclusion tuples are built using the function `stemInclusionTuples()` and then provided to the `itemInclusionConstraint()` function.

Test specification (3) is implemented similarly as in the previous cases using the `itemUsageConstraint()` function. The "less than or equal" operator "`<=`" is used, because complete depletion of the item pool is not required. Test specification (4) refers to the forced inclusion of certain items in the block assembly, which can also be implemented using the `itemUsageConstraint()` function. In this specific case we specify the `whichItems` argument, which lets us choose to which items this constraint should apply to. For this specification, the `operator` argument is set to "`=`" as the items have to appear once across the blocks.

The further test specifications are implemented in line with similar constraints in previous examples: block times, referring to test specification (5), are constrained using the `itemValuesDeviationConstraint()` function. Test specifications (6), (7), and (8) are implemented by transforming the respective variables to factors so we can apply the `itemCategoryMinConstraint()` or the `itemCategoryMaxConstraint()` functions. As every block should contain at least some items at the intermediate difficulty levels and also in each block at least one item at the adjacent difficulty levels, we set the `min` argument for this test specification to `c(1, 2, 2, 1)`. For test specification (7), item formats are grouped into three different groups, which then are constrained by setting the minimum number of items of each group per block to two. In some LSA studies, items are flagged that have empirical proportions correct below and/or above a certain value (cf., test specification (8)). Therefore, we limit the inclusion of items that range below 8 percent and above 92 proportion correct to a maximum of two items per category per block.

The complete code for the LSA use case, including the solver call and the solution inspection, can be seen in Supplement S2. Note that for this test assembly problem, the GLPK Simplex Optimizer finds a feasible solution very quickly but the complete integer optimization process takes a substantial amount of time, due to the large item pool, multiple assembled item blocks, and various constraints. This showcases that often setting a time limit and using a feasible but not optimal solution is sufficient in practice.

6.3.3 High-Stakes Assessment

This use case corresponds to Problem 1 in the paper by Diao and van der Linden (2011). Because the item pool used in Diao and van der Linden (2011) is not freely available, an item pool was generated with similar characteristics (`items_diao` in the `eatATA` package). The item pool consists of 165 items following the three-parameter logistic model (3PL). Each item belongs to one of six content categories. The first five items of the generated item pool can be seen in Appendix D.4. In this example, the goal is to assemble two parallel test forms with the following test specifications: (1) absolute target values for the TIFs set as $T_\theta = 5.4, 10, 5.4$ at $\theta = -1.5, 0, 1.5$; minimize the distances of the TIFs of the two new forms with respect to the target at these ability values, (2) distribute the number of items per content category evenly across test forms; Appendix D.5 presents the numbers of items per content category that are available in the complete item pool as well as the numbers required in each of the two forms, (3) no overlapping items, and (4) each test form should contain exactly 55 items. These specifications are directly copied from Diao and van der Linden (2011). The code for calculating the IIF, and setting up the minimax objective function as well as the other constraints can be seen in Figure 33.

To implement test specification (1) we create minimax objects at each of the specified ability values, and combine them in one objective function. Hence, when solving the ATA problem, the solver will try to minimize the maximal distance between the target and the two forms at the three ability values. The implementation of the further constraints directly corresponds to formulations of test specifications in the use cases above. Therefore, further explanations on these specifications are omitted. The complete code for the high-stakes assessment use case, including the solver call and the solution inspection, can be seen in Supplement S3.

6.3.4 Multi-Stage Testing

This use case covers the case of multi-stage testing and corresponds to Problem 3 in Diao and van der Linden (2011). The use case uses the same items as use case (3), but the item pool is doubled: all items are duplicated. Hence, the item pool in this example contains 330 items following the 3PL model. Here, the goal is to assemble a two-stage multi-stage test with one routing module in the first stage and three modules in the second stage. The test specifications are: (1) the TIF of the first-stage module is required to be relatively uniform between $\theta = -1$ and $\theta = 1$, (2) the TIFs of the second-stage modules are required to be

Figure 33: *Constraint Definitions for High-Stakes Assessment Test Assembly.*

```
# Theta values
theta_values <- c(-1.5, 0, 1.5)

# Calculate item information for each theta-value
items_diao[, paste0("IIF_", theta_values)] <- calculateIIF(A = items_diao$a,
                                                           B = items_diao$b,
                                                           C = items_diao$c,
                                                           theta = theta_values)

# Specify target values for each theta-value
target_values <- structure(c(5.4, 10, 5.4),
                           names = paste0(theta_values))

# Objective function: minimize maximum difference between the TIF and
#  the target values
minimaxTif <- combineConstraints(lapply(theta_values,
                                        function(theta_value) {
  minimaxObjective(
    nForms = 2,
    itemValues = items_diao[, paste0("IIF_", theta_value)],
    targetValue = target_values[as.character(theta_value)],
    itemIDs = items_diao$item
  )
}))

# Other constraints
contentConstraints <- itemCategoryConstraint(
  nForms = 2,
  itemCategories = items_diao$Category,
  operator = ">=",
  targetValues = c(9, 9, 7, 9, 9, 11),
  itemIDs = items_diao$item)

noOverlap <- itemUsageConstraint(
  nForms = 2,
  itemIDs = items_diao$item)

testLength <- itemsPerFormConstraint(
  nForms = 2,
  operator = "=",
  targetValue = 55,
  itemIDs = items_diao$item)
```

single-peaked at $\theta = -1$, $\theta = 0$ and $\theta = 1$, respectively, (3) the number of items per content category should be evenly distributed across the test forms according to Appendix D.5, (4) no item overlap, (5) for the first-stage module the test length should be 30 items, and (6) for each of the second-stage modules the test length should be 20 items.

Original Approach. Diao and van der Linden (2011) split this assembly problem in two separate problems. In a first step, the routing module for the first stage is assembled. Thereafter, the three modules for the second stage are assembled using only the remaining items in the pool. In order to assemble the routing module with a uniform relative target, a maximin approach is used—that is, the minimum value of the TIF at the three ability values is maximized. At the same time, the TIF values at the three ability values are required to be close to each other, in order to create a TIF with a relatively flat plateau. More specifically, the TIF at the three ability values is required to be within a distance of 0.5 of each other. Hence,

the `allowedDeviation` is set to 0.5. The syntax for the implementation of the objective function and the constraints can be found in Appendix D.6.

Combined Capped Approach. Using the `eatATA` package, it is possible to assemble the modules for the two stages in one combined assembly. Especially in situations with multiple stages, a simultaneous assembly may prevent infeasibility at later stages—that is, when the modules for the stages are assembled sequentially, the assembly of the first stages may deplete the item pool so that it becomes impossible to meet certain test specifications at later stages. In addition, from a practitioner's perspective, a simultaneous assembly may also be easier, as the item pool does not need to be adjusted after every assembly step.

The `R` syntax for the combined capped approach can be seen in Figure 34. To implement test specification (1) we specify the maximization of the minimum TIF values at the ability values for the routing module in the first stage. Note that we do not use the original maximin approach but rather the capped maximin approach (Luo, 2020). The capped maximin approach does not require to set a maximally allowed deviation, it combines maximizing the minimal TIF with minimizing the maximal difference between the TIFs. For the modules at the second stage (test specification (2)) the capped maximin approach is also used. To combine these constraints, knowing that the obtained TIF values in the first stage and the obtained TIF values at the second stage do not need to be in the same range, we can set a weight for the TIF values. In this case, the minimal TIF in both stages can be considered equally important. Hence, the weights are set to 1 (which is the default). Because the other test specifications in Figure 34 correspond to test specifications illustrated earlier, further explanations are omitted. The complete code for the multi-stage assessment use case, both for the original two-stage as well as the new combined assembly, with the solver call and the solution inspection, can be found in Supplement S4.

6.4 Discussion

In 2005, van der Linden published his seminal book on automated test assembly. Since then, additional tutorial papers and illustrations have been written to make ATA methods more accessible to practitioners (e.g., Diao & van der Linden, 2011; Donoghue, 2015). However, we believe that hurdles are still quite substantial for practitioners who want to utilize ATA methods: Besides the conceptual challenges of formulating test specifications as concrete constraints, one has to become familiar with formulating mathematical constraints and the intricacies of specific solver APIs. The `eatATA` package and this tutorial paper have been

Figure 34: *Constraint Definitions for Multi-Stage Assessment Module Assembly.*

```
# Objective function (TIF stage 1)
maximinTIF1 <- combineConstraints(lapply(theta_values,
                                         function(theta_value)
{
  cappedMaximinObjective(
    nForms = 4,
    itemValues = items_diao2[, paste0("IIF_", theta_value)],
    weight = 1,
    whichForms = 1,
    itemIDs = items_diao2$item)
}))

# Objective function (TIF stage 2)
maximinTIF2 <- combineConstraints(lapply(theta_values,
                                         function(theta_value)
{
  cappedMaximinObjective(
    nForms = 4,
    itemValues = items_diao2[, paste0("IIF_", theta_value)],
    weight = 1,
    whichForms = which(theta_values == theta_value) + 1,
    itemIDs = items_diao2$item)
}))

# Content categories stage 1 and 2
contentConstraints1 <- itemCategoryConstraint(
  nForms = 4,
  itemCategories = items_diao2$Category,
  operator = ">=",
  targetValues = c(4, 4, 3, 4, 4, 5),
  whichForms = 1,
  itemIDs = items_diao2$item)
contentConstraints2 <- itemCategoryConstraint(
  nForms = 4,
  itemCategories = items_diao2$Category,
  operator = ">=",
  targetValues = c(3, 3, 2, 3, 3, 4),
  whichForms = 2:4,
  itemIDs = items_diao2$item)

# No item overlap
noOverlap2 <- itemUsageConstraint(
  nForms = 4,
  itemIDs = items_diao2$item)

# Test length stage 1 and 2
testLength1 <- itemsPerFormConstraint(
  nForms = 4,
  operator = "=",
  targetValue = 30,
  whichForms = 1,
  itemIDs = items_diao2$item)
testLength2 <- itemsPerFormConstraint(
  nForms = 4,
  operator = "=",
  targetValue = 20,
  whichForms = 2:4,
  itemIDs = items_diao2$item)
```

written to promote ATA methods and make them more accessible to practitioners and researchers. We have provided a short overview of the basic ideas of ATA and an illustration of the typical `eatATA` work flow. Using a small illustrative example and four different, more realistic use cases, we demonstrated how the package can be used to implement ATA in `R`. By choosing a wide range of different ATA applications with diverse test specifications we hope to spark interest in ATA methods in a broad audience.

6.4.1 Limitations

There are currently a few limitations when using `eatATA` to solve ATA problems. As mentioned in use case (2), hierarchical item stimulus structures cannot be implemented as flexibly as suggested by van der Linden (2005) with optional item selection. However, in practice, item sets are often treated as fixed units anyway, as altering the item set that is presented alongside a stimulus might have undesirable effects on the psychometric properties of the individual items. Furthermore, item overlap specifications between test forms cannot be specified directly in `eatATA`. Generally, a direct implementation of item overlap constraints drastically increases the complexity of the mathematical programming problem, resulting in high computing times. Moreover, often item overlap specifications can be met indirectly, for instance by first selecting a set of items that can serve as overlap items, and then constraining the number of overlap items per test form, as well as how many times the overlap items can appear across the test forms. Therefore, direct overlap constraints are deliberately not included in `eatATA`. Finally, solver selection in `eatATA` is limited to the solvers mentioned in the introduction. For example, `CPLEX` and `XPRESS` are potent commercial alternatives to `Gurobi`. Another potentially promising open source solver, unfortunately currently without an `R` API is `SCIP` (Luo, 2020). However, we do believe that for many ATA contexts the available selection of solvers is more than sufficient.

6.4.2 Alternatives

It is noteworthy that while for most data handling procedures or statistical methods a wide variety of `R` packages exists, this is not the case for ATA methods. More precisely, we are only aware of four other `R` packages on CRAN that have some ATA functionality implemented, of which only two (`TestDesign`, `RSCAT`) seem to be under active development. Indeed, `TestDesign` (Choi & Lim, 2020) provides access to the same selection of solvers as `eatATA` but has a strong focus on adaptive testing. This is illustrated by the fact that `TestDesign` is not suited for the assembly of multiple parallel test forms[32]. In a similar vein, `RSCAT` provides functionality specific to the shadow-test approach in computerized adaptive testing (Jiang, 2020). However, other testing approaches, such as multi-stage testing or linear testing, are not supported in that package. Finally, `Rata` (Luo, 2019a) and `xxIRT` (Luo, 2019b) also implement ATA methods in `R`. Yet both packages only provide access to the `Rglpk` and `lpSolve` solvers. In addition, although both packages in general have a similar

[32]Readers of the present thesis should note that in the meantime a tutorial paper on using TestDesign for fixed-form test assembly as well as CAT has been published (Choi et al., 2021).

work flow compared to `eatATA`, their functionality is more limited compared to `eatATA` (e.g., no specific categorical item constraint functions, no automatic calculation of target values, no item inclusions constraints).

Furthermore, alternative approaches to MILP have been proposed in the past, e.g., heuristic algorithms, which are also capable of solving automated test assembly problems. Examples include genetic algorithms (T.-Y. Chang & Shiu, 2012; Sun et al., 2008; Verschoor, 2007) or simulated annealing (Veldkamp, 1999). For a short but comprehensive overview see van der Linden (2005). As these algorithms do not search the entire solution space they are not guaranteed to find the optimal solution to the optimization problem but may be computationally faster than the classic MILP solvers that are used by `eatATA`. Another potential benefit of heuristic algorithms is that some allow the introduction of soft constraints, which might be helpful for dealing with feasibility issues. However, to our knowledge, only limited ATA applications of these algorithms exist. For example, we are not aware of a single ATA application of heuristic algorithms using `R`. Furthermore, it can be argued that for most practical ATA applications MILP solvers perform sufficiently well from a computational stand point (van der Linden & Li, 2016).

6.4.3 Conclusions

We believe that `eatATA` can be a helpful tool for researchers and practitioners that want to assemble test forms. It is applicable in a wide range of scenarios and its user interface should be rather intuitive for `R` users. By providing this tool we hope to promote automated test assembly methods, which are almost always superior to manual test assembly approaches.

7 Discussion

As assessments are important cornerstones of modern educational systems, their fairness and validity are prerequisites for the fairness and the effectiveness of the educational systems themselves. Speededness has frequently been identified as a potential threat to the validity and fairness of assessments in a wide range of different contexts. The validity of power tests with time limits is threatened as speededness can introduce construct-irrelevant variance into an assessment, thereby changing the nature of the measured construct. If a certain degree of speededness is desired in an assessment, speededness has to be carefully controlled to avoid under-representation of constructs. The fairness of assessments is threatened as speededness can be expected to affect test-takers very differently depending on their working speed. This threat is amplified by the fact that different subgroups (e.g., gender and ethnicity subgroups) are differentially affected by speededness (e.g., Lawrence, 1993; Voyer, 2011). While there is plenty of research on the topic of speededness in general, to this day there is still little research on (a) how speededness can be properly controlled in assessments and (b) how speededness may affect the fairness of assessments when multiple, parallel test forms are used. Furthermore, (c) there are currently only very limited practical tools and tutorials for controlling speededness in assessments.

One of the only published approaches for controlling speededness so far is the approach proposed by van der Linden (2011a, 2011b). However, it is limited to the rather restrictive 2PLN model, which assumes equal speed sensitivities across all items. Furthermore, van der Linden (2011a, 2011b) argues that controlling the speededness on the test-form-level is sufficient to generate parallel test forms. However, as already Sax and Cromack (1966) as well as Leary and Dorans (1985) noted, some item orders can be more beneficial than others if a test is speeded. The present thesis builds upon the ideas of van der Linden (2011a, 2011b) to develop a more general and flexible approach for controlling speededness, which can accommodate the 3PLN response time model. Furthermore, the thesis builds upon the ideas of Leary and Dorans (1985) to develop simple-to-implement approaches mitigating differential effects of speededness due to different item orders. Finally, the thesis seeks to supply practitioners with all the necessary tools to implement the proposed approaches in practice. In the following, the research work presented in Chapters 2-6 is summarized.

Chapters 2 and 3 focused on illustrating how the approach for controlling speededness proposed by van der Linden (2011a, 2011b) falls short in frequently encountered empirical situations. The approach by van der Linden (2011a, 2011b) makes use of a restricted lognor-

mal response time model, the 2PLN model, in which all speed sensitivities (factor loadings) are set to 1. However, so far, no substantial arguments in favor of this assumption have been presented in the literature. Empirically, this assumption seems overly restrictive as well, as whenever it has been tested on empirical data, the assumption of equal speed sensitivities did not hold. This was also shown in the empirical data analyses provided in Chapter 2 and 3. Based on this argument, Chapter 2 illustrated how differential speed sensitivities across test forms can lead to substantially unfair test forms. Chapter 3 presented a more general approach for controlling speededness in assessments, which makes use of the 3PLN model instead of the 2PLN model.

Chapter 4 demonstrated how speededness can negatively affect the fairness of parallel test forms. As it can be assumed that speededness mainly affects items at the end of the test, properties of the items placed at the end of test forms are crucial for the impact of speededness. In case a test is speeded for a test-taker, the test becomes less challenging when items are sorted by ascending difficulty (easy to hard) than when items are sorted by descending difficulty (hard to easy). This is due to the fact that the test-taker automatically spends most of their time on items they have a high solution probability on, namely the easy items at the beginning of the test. Therefore, Chapter 4 proposed that using identical and time intensive items at the end of interchangeably used test forms is a sensible and easy-to-implement approach to overcome such fairness issues.

Chapters 5 and 6 provided assessment practitioners and researchers interested in applying the approaches presented in Chapters 2-4 with some theoretical and practical guidelines for doing so. Chapter 5 illustrated how the response time models discussed in the Chapters before can be implemented in the general-purpose Bayesian estimation software `stan`. Various extensions of the basic hierarchical framework were discussed as well. The suggested model implementations, convergence checks, and model comparisons were, for instance, applied in Chapter 3. Chapter 6 demonstrated how general automated test assembly can be performed using the R package `eatATA`, which was created in the scope of this thesis as well. The R package `eatATA` is, for instance, in use for the German Vergleichsarbeiten or the international TIMSS study 2023 (von Davier & Mullis, 2022).

In the following, important limitations of the research presented in Chapters 2-6 will be discussed. These limitations refer to specific assumptions of the used response time models and the test assembly approaches. Second, alternative approaches for controlling speededness in assessments are presented. Third, practical implications of the presented research for

different testing contexts are discussed. The thesis closes with an outlook on how future research could build upon the presented research.

7.1 Limitations

Potential limitations of the presented research mainly pertain to assumptions of the response time modeling approaches used and general assumptions made regarding test-taking behavior. Specific limitations to the presented research work and proposed approaches were already briefly discussed in Chapters 2-6. Beyond these brief discussions, this section focuses in-depth on two of the most relevant assumptions: The conditional independence assumption, which is mainly of relevance for Chapters 2 and 3, and the stable item order assumption, which is mainly of relevance for Chapter 4.

7.1.1 Conditional Independence Assumption

A central assumption of the 2PLN and the 3PLN model is that of local stochastic independence. It is assumed that, given the latent speed factor, no residual covariances exist between manifest response times of the different test items (van der Linden, 2006). This assumption can also be reframed as a *stable speed assumption*. This assumption is violated, if a test-taker works with varying speed levels, for instance if they realize after the first half of the test that it is necessary to speed up to respond to all items within the time limit. As a consequence, residual correlations will occur within items of the first half and within items of the second half. Indeed, especially in the context of the hierarchical framework, the stable speed assumption has frequently been questioned (e.g., Bolsinova & Tijmstra, 2016; Coomans et al., 2016; Domingue et al., 2021; van der Linden et al., 2010). For instance, some researchers argue that response processes typically differ qualitatively and should be categorized into fast and slow processes (e.g., Coomans et al., 2016; Molenaar & de Boeck, 2018). In contrast, Fox and Marianti (2016) argue that test-takers may simply change their working speed throughout the test, for instance due to decreasing motivation for low-ability test-takers. Violations of the independence assumption can occur at two different steps in the assessment cycle: (a) when piloting or pre-testing items or (b) during the operational test.

Pre-Test. When controlling the speededness of an assessment as proposed in Chapter 3, test designers rely on unbiased and reliable estimates of item parameters from pilot studies. If model assumptions of the respective response time models are violated, this can lead to bias in item parameter estimates and thereby to biased and unfair test forms during test assembly.

This applies both for response time modeling (used for controlling the speededness of test forms) and for response modeling (for controlling, for instance, the TIF of test forms). In fact, in the context of IRT, a comparable assumption to the stable speed assumption is made: That of a constant ability level that does not change throughout the test (van der Linden & Glas, 2010b). Research investigating violations of this assumption, however, usually focuses on *item position effects* (e.g., Debeer & Janssen, 2013; Nagy et al., 2018; Ong & Pastor, 2022; Weirich et al., 2017; Q. Wu et al., 2019). This discrepancy regarding framing (ability differences on the person side vs. difficulty differences on the item side) is due to the fact that instability of test-taker behavior across a test can either be attributed to the ability of a person (e.g., the displayed ability is decreasing over the course of the test due to fatigue) or to the difficulty of the items in the test (e.g., items positioned at the end of the test have a higher effective difficulty compared to if they are placed at the beginning of the test as test-takers are already fatigued). On the first glance it may seem that the assumption of stability of both speed and ability is more restrictive than the assumption of just speed stability. However, it can be argued that violations of the constant speed assumption will indirectly lead to violations of the constant ability assumption due to the speed-accuracy trade-off anyway. Nevertheless, the issue of biased item parameters from pilot studies is a serious concern. Ideal piloting conditions are therefore discussed in a designated section in "7.3 Practical Implications".

Operational Test. In general, it should be emphasized that the proposed approach for controlling speededness in Chapter 3 is to be used before any response data is collected from an operational assessment. This means that the response time models discussed are applied to pre-test data, not to the response times of the operational assessment (at least not for the purpose of controlling speededness when designing the assessment). Nevertheless, in an ideal world, test-takers would adopt the exact speed level that is required from them to just finish in time. Any deviation from this speed level (e.g., initially working too slow and then guessing on or not answering to items at the end of the test) can be expected to result in lower scores for test-takers (Tijmstra & Bolsinova, 2018). However, such deviations from ideal test-taking behavior are not an issue of test assembly but arguably an issue of unclear test instructions or too complex requirements regarding test-wiseness. One could argue that additional measures (e.g., more fine-grained time limits such as item time limits; Goldhammer, 2015; Goldhammer et al., 2017) may be sensible. While this is not an issue of controlling speededness in assessments, future research directions related to this are discussed

in “7.4 Directions for Future Research”.

7.1.2 Stable Item Order Assumption

Chapter 4 discussed how speededness and item order may interact, resulting in unfair test forms. In speeded conditions, test forms containing identical item sets can lead to different ability estimates depending on how items are ordered. This is, however, mainly an issue if it is assumed that (a) test-takers are affected by speededness mainly at the end of a test and (b) that test-takers work on tests in a linear fashion. In fact, there is overwhelming consensus among researchers as well as empirical evidence that both assumptions hold in practice. This includes research on not reached items, rapid guessing, or test speededness, which frequently assumes and shows that all these processes affect items located at the end tests (e.g., Glas & Pimentel, 2008; Lindner et al., 2019; Nagy et al., 2022; Pastor et al., 2019; Pohl & Carstensen, 2013; Pohl et al., 2019; Pools, 2022; Rose et al., 2017; Schnipke & Scrams, 1997; Tijmstra & Bolsinova, 2018; Ulitzsch et al., 2020; Wise et al., 2009). This assumption is also in line with research on decreasing test taking engagement, which focuses on how engagement declines throughout the test based on item positions (e.g., List et al., 2017). Furthermore, the aforementioned research on item position effects finds that items positioned at the end of a test are typically more difficult than items positioned earlier in a test (e.g., Debeer & Janssen, 2013; Nagy et al., 2018; Ong & Pastor, 2022; Weirich et al., 2017; Q. Wu et al., 2019). Additionally, research on statistically detecting speededness or accommodating for speededness frequently finds that test-takers speed up at a certain point in a test or increase their working speed throughout the test (Bolt et al., 2002; De Boeck et al., 2011; Goegebeur et al., 2008; Jin & Wang, 2014; Kahraman et al., 2013; Suh et al., 2012; Williams, 2017; Wollack et al., 2003; Yamamoto, 1995). Lastly, Lee and Haberman (2016) investigated the order in which test-takers work on the items of a test and confirm that they do so predominantly in the presented order. Note that similar findings of test-takers speeding up and performing worse at later points in the test were also confirmed in Chapter 5 of this thesis. The only exceptions to this overwhelming evidence can be found in the context of formula scoring.

Formula Scoring. *Formula scoring* is a scoring approach for multiple-choice tests that penalizes incorrect responses to discourage test-takers from randomly guessing on items to which they do not know the answer (Lord, 1975; Thurstone, 1919). The penalty is defined as $-\frac{1}{c-1}$, with c denoting the number of response options. It is plausible to assume that

formula scoring causes test-takers to skip harder items and leave them for later. Doing so allows test-takers to later decide if they want to answer to the harder items at all. In the context of formula scoring, Y.-W. Chang et al. (2014) and J. Chang et al. (2016) propose speeded IRT models, which assume that test-takers work on the items in the order of increasing difficulty. The researchers term this test-taking strategy the *leave-the-harder-till-later* approach. Thereby, these models account for the fact that due to speededness test-takers may not answer to all items and these items are most likely the hardest ones in the test. See Bejar (1985) and Cao and Stokes (2008) for similar ideas but with a focus on detecting the speededness of a test and partial rapid guessing, respectively.

Formula scoring is generally viewed critically by the research community as personality traits and test wiseness are likely to affect test scores, thereby threatening the validity and fairness of a test (Budescu & Bo, 2015; Rowley & Traub, 1977). Yet, while a leave-the-harder-till-later test-taking strategy is possible in paper-based assessments, it can be expected to be less probable in computer-based assessments where navigation between items and tasks is often less convenient. Additionally, test administrators can restrict the possibility to navigate backwards to already seen items in computer-based assessments. Nevertheless, if formula scoring is used in an assessment and navigating between items is freely possible, considering differential effects of item order may be less relevant.

Finally, the approaches suggested in Chapter 4 can be considered low-risk approaches. Even if test-takers were to work on the presented items in a different order, presenting the same, most time intensive items at the end of the test for all test-takers will hardly have any negative consequences. It is plausible to assume that at least some of the test-takers will work on a test in a linear fashion. For them, eliminating unfair effects of differential item ordering is vital. For all other test-takers, fixing the most time intensive items at the end of all test forms will not be harmful. The small loss in test security should have very little practical impact (see Chapter 4 for more detailed explanations regarding test security and test wiseness).

7.2 Alternative Approaches to Dealing with Test Speededness

While the approach of van der Linden (2011a, 2011b) and the proposed generalization of it to the 3PLN model (see Chapters 2 & 3) are among the only approaches for controlling speededness in achievement assessments, other researchers have proposed conceptually different approaches regarding how to deal with test speededness. Five approaches which, at least in

some situations, can be viable alternatives or suitable additions to controlling speededness in test assembly, will be discussed: (a) maximizing information per time unit, (b) explicit scoring rules, (c) investigating the speed-ability trade-off experimentally, (d) statistical modeling approaches, and (e) item time limits. As these approaches have not yet been discussed in Chapters 2 and 3, detailed discussions to which degree they may substitute controlling speededness during test assembly are warranted.

7.2.1 Maximizing Information per Time Unit

In the CAT context, measurement efficiency is of major interest. Frequently, measurement efficiency is defined as the number of items which have to be administered to achieve sufficient measurement precision. Others have argued that efficiency should instead be measured in terms of required testing time. For this purpose, Fan et al. (2012) introduced the *maximizing information per time unit criterion* (MITC). They build upon the lognormal response time model from van der Linden (2006) and use it to predict response times for all not yet administered items in the item pool. Based on these expected response times, item selection strategies are modified such as the *maximum information criterion* (MIC) or the *a-stratified item selection method*. The goal is to no longer select the most informative item overall but the item that provides the most amount of information per time unit in CAT. This is achieved by using the expected response time $E(RT_k)$ given a preliminary speed estimate as a weighting factor. For instance, the MITC of an item k for ability level θ and speed level ζ is defined as

$$I_k(\theta, \zeta) = \frac{I_k(\theta)}{E(RT_k)}. \tag{61}$$

A number of studies have built upon the ideas of Fan et al. (2012) using different response time models (Patton, 2015), modifying the MITC approach (Cheng et al., 2017; Choe et al., 2018), or applying it to CDMs (Finkelman et al., 2014). An important generalization of the MITC approach, the *generalized MITC* (GMITC), has been presented by Choe et al. (2018). The model is generalized in that it allows centering and weighting the response times. Thereby the model allows not only minimizing the testing time, but minimizing the difference to a target test time v:

$$I_k(\theta, \zeta) = \frac{I_k(\theta)}{(E(RT_{ik}|\zeta_i) - v)^w}. \tag{62}$$

The MITC is a special case of the GMITC with $v = 0$ and $w = 1$. Despite the popularity of the MITC or related approaches in research it should be noted that MITC approaches have some important practical limitations compared to the more flexible approach for controlling speededness presented in Chapter 3. First, they are only applicable in CAT, as no speed estimates are available in fixed-form testing. An extension to MST seems possible, but has apparently not yet been proposed. Second, MITC approaches in themselves provide no framework for incorporating further test specifications (e.g., T. Wu et al., 2022). In contrast, the approach proposed in Chapter 3 is based on ATA, which easily allows for incorporating further test specifications (Huang, 2019). Therefore, GMITC approaches can be valuable tools for reducing or controlling the speededness in CAT applications with a limited amount of additional test specifications, but are not applicable in a wider range of scenarios.

7.2.2 Scoring Rules

In explicit scoring rule approaches, speed is incorporated directly into the scoring of responses. For instance, Maris and van der Maas (2012) propose the so called *signed residual time* (SRT) scoring rule. The scoring rule places larger rewards on fast and correct responses than on slow and correct responses as well as larger penalties on fast and incorrect responses than on slow and incorrect responses. An advantage of this approach is that test-takers are informed about the scoring method beforehand and can incorporate this knowledge into their test-taking strategy. For an extension of the originally proposed model, see also van Rijn and Ali (2018). One of the benefits of the SRT scoring rule is that information can even be gained from comparatively easy items with high solution probabilities, namely how fast test-takers can produce the correct answer. However, the SRT scoring rule has been criticized for ignoring that test-takers can differ in their chosen speed-ability trade-off (Tijmstra & Bolsinova, 2021). Furthermore, it can be argued that such scoring rules always imply a confounded measurement of speed and ability and make the measurement of a pure ability dimension impossible. Currently, the SRT scoring rule (or comparable scoring rules) are mainly used in the context of *computerized adaptive practicing* (e.g., Klinkenberg et al., 2011). In computerized adaptive practicing, learners are typically presented with easy items (expected probability of correct responses above .75) in order to motivate the learners and continue the learning process. In addition, the typical skills in these environments are skills that benefit from drill and practice, such as early numerical skills or basic reading skills. It seems that scoring rules are well suited for these applications but may not be applicable in

the wider assessment context.

7.2.3 Experimentally Varied Speed-Ability Trade-Off

Especially in the context of experimental psychology, researchers have argued that rather than focusing only on ability or speed, measuring the speed-accuracy trade-off provides a more holistic picture (Goldhammer et al., 2017; Wickelgren, 1977). In line with this, one could argue that assessments could experimentally vary the speed-ability trade-off to assess the complete shape of the trade-off for all test-takers. This would mean gaining substantially more information than simply measuring a compound of speed and ability. However, current research indicates that the interpretation of speed-ability trade-offs can be challenging especially in unspeeded low-stakes settings. For instance, Domingue et al. (2022) present an extensive review of the SAbT among a wide variety of data sets and find inconsistent results. Certain features of assessments, such as time limits or different stakes may confound findings. For example, in low-stakes assessments, fast responses could be motivated responses, when test-takers move fast and efficiently through the test. However, fast responses could also be produced by unmotivated test-takers who do not take sufficient time and care in answering the items. Ranger et al. (2021) reason that motivational processes can indeed lead to different empirical speed-ability trade-offs.

Regardless of conceptual issues, experimentally varied SAbT assessments would be very challenging to implement in practice. Such experimental manipulations would deviate strongly from current assessment practice and would complicate the test administration as well as the analysis of tests. Furthermore, it is unlikely that test-takers can always perfectly control their speed-ability trade-off. In current assessment practice, usually only a single speed-ability manipulation is implemented (i.e., a global time limit is set). Still, test-takers often struggle to find the appropriate speed-ability trade-off that enables them to complete all items without producing missing responses or resorting to rapid guessing (Tijmstra & Bolsinova, 2018).

7.2.4 Statistical Modeling

Other researchers have proposed that ability could be statistically purified of the effects of speed and speededness. For instance, speed and ability could simply be modeled separately in the hierarchical framework by van der Linden (2007) and be reported as separate latent constructs (e.g. Pohl et al., 2019, 2021; Ulitzsch et al., 2020). Others have suggested mixture modeling approaches to separate solution behavior and rapid guessing behavior due to test

speededness (Boughton & Yamamoto, 2007; Meyer, 2010; C. Wang & Xu, 2015). Probably the most straightforward approach is treating not-reached items as not administered (e.g., Pohl & Carstensen, 2012). This way, test-takers can work at their maximum capacity and the test is simply reduced in its test length appropriately for each test-taker.

While such approaches have the advantage that they can be applied after the assessment has been administered, for instance if speededness was not considered during test design and assembly, they have various weaknesses: First, those approaches suffer from similar weaknesses as approaches that try to measure and quantify the speededness of a test. They frequently rely on symptoms of test speededness which simply may not be present for test-wise test-takers, such as rapid guesses or missing responses. Second, these approaches make specific assumptions about the underlying speed-ability relationship (e.g., linearity) that may not always hold in practice. Third, even if we know the speed and ability level at which a person has worked, we can never know the maximum ability level at which a person would be able to work from a single speededness condition (Goldhammer et al., 2017; Tijmstra & Bolsinova, 2018). Fourth, individual person parameter estimates will depend on specific sample dynamics. For instance, assume a joint model for ability and speed is estimated to account for speed differences. If, in the sample, high-ability test-takers usually work fast, other test-takers will also be rewarded simply for working fast. If in another sample, high-ability test-takers work slowly but carefully, other test-takers will be punished for working fast. Fifth, especially in the context of high-stakes assessments, such purification approaches could be abused or "gamed" if test-takers are aware that specific response strategies are rewarded. Sixth, in low-stakes settings, speededness can sometimes be difficult to disentangle from (other) sources of construct irrelevant variance, such as test-taker disengagement (e.g., Wise & Kingsbury, 2022; Wise & Kuhfeld, 2021).

These concerns are in line with concerns voiced by other researchers who have expressed general skepticism about purification approaches (Robitzsch & Lüdtke, 2022; Tijmstra & Bolsinova, 2018). Partchev et al. (2011) have shown that a test which was administered without speededness for all test-takers may be turned into a compound measure of speed and ability after its administration via *posterior time limits*. However, a test with undefined levels of speededness for test-takers cannot be purified of its potential speededness via statistical modeling. Nevertheless, there certainly is merit in approaches for detecting speededness and purifying ability estimation of the effects of speededness. However, none of these approaches can substitute for controlling the speededness of test forms in the first place. Instead, statis-

tical modeling may provide a useful addition to approaches for controlling speededness, for instance if test-takers in low-stakes assessment work too fast due to motivational issues.

7.2.5 Item Time Limits

Goldhammer (2015) proposes using time limits on item level to prevent test-takers from working at different speed levels during an assessment. In contrast to common overall time limits for complete tests, item time limits restrict how much time a test-taker can spend on each individual item in the test. This approach is specifically targeted at issues arising from test-takers choosing different speed-ability trade-offs in the assessment, thereby distorting ability differences between test-takers (Goldhammer, 2015; Tijmstra & Bolsinova, 2018). However, item time limits themselves do not prevent validity issues due to speededness or fairness issues due to differential speededness. If item time limits are chosen too strictly, speededness can confound a pure ability measurement just as much as test time limits. If different test forms with distinct items are used for different test-takers, time limits and workload still need to be chosen to be fair across test-takers. Therefore, similar to statistical modeling, item time limits may be a suitable addition to the proposed approach for controlling speededness in assessments but cannot substitute it. In fact, the approach for controlling speededness in Chapter 3 may be useful in setting appropriate time limits on the item level. Finally, van Rijn et al. (2021) report that item time limits can have negative effects on the overall performance of test-takers compared to test time limits. Future research could investigate both the practical advantages and disadvantages of item time limits and whether the approach presented in Chapter 3 is suitable for setting item time limits.

7.3 Practical Implications

There is currently little guidance for assessment practitioners on how to control speededness and the consequences of speededness in assessments in the literature. This section aims at helping to fill this gap, providing some practical advice and further discussion. First, the practical relevance of the proposed approaches for different testing contexts, namely high-stakes and low-stakes assessments, and different levels of assessment adaptivity will be discussed. Second, it will be discussed how piloting conditions should be designed to obtain reliable and valid item parameter estimates for test assembly procedures. Finally, the current lack of transparent reporting on how speededness is controlled during test assembly in almost all testing programs is discussed and suggestions are made on how testing programs could

improve their reporting.

7.3.1 Relevance for Different Assessment Contexts

Chapters 2 - 4 were mainly focused on speededness in the context of fixed-form linear high-stakes assessment. However, as was already pointed out in Chapter 1, it can be argued that controlling speededness is relevant in most testing contexts. In the following, it is argued that especially Chapters 2 and 3 are relevant irrespective of the stakes for test-takers and the level of adaptivity used.

Assessment Stakes. When stakes are high for individual test-takers it is apparent that test forms must be fair (i.e., interchangeable) for each individual taking the test. In low-stakes assessments, for instance large-scale assessments, results are usually only reported on group level (e.g., country; Kirsch et al., 2013). Therefore, strict requirements for fairness on the individual test-taker level often do not exist. If test forms are distributed evenly across the different groups, which should be compared, differences between the test forms simply even out across groups. However, this does not mean that controlling speededness is not relevant in the context of low-stakes assessments. While speededness may not be a fairness issue on the level of individual test-takers in low-stakes assessments, it should be considered a validity issue in both high-stakes and low-stakes assessments. If an educational large-scale assessment seeks to measure knowledge in geometry and is involuntarily speeded, it is unclear, which policy decisions should be based on poor results in such a test: Is the geometry curriculum flawed? Or were the students simply not answering test questions in a fast and efficient manner? Other assessments seek to measure a compound construct, as in the case of reading speed which can be considered an essential part of reading literacy. In such cases, the construct may not be correctly represented if test forms are not speeded enough or too strongly speeded. Furthermore, speededness can still be a fairness issue if subgroups differ in their working speed. In such instances, a speeded test can favor certain subgroups leading to fairness issues on the group level.

Additionally, controlling speededness should be considered relevant from a practical perspective. For instance, differentially speeded test forms can have a negative impact on test-taking behavior, as low-stakes assessments are often administered in a classroom context. If test forms are differentially speeded, meaning that some test forms contain more workload than others, this will lead to some test-takers finishing ahead of other test-takers. These test-takers may then get bored and distract other test-takers from the test. Alternatively, if

some test-takers realize they are proceeding slower compared to other test-takers, they may feel the urge to work faster (and less accurately) on the test.

Adaptivity. While controlling speededness is crucial during the assembly of fixed-form linear test-forms, one can argue that it is of even greater importance in the context of adaptive assessment modes. Speededness is an inherent concern in CAT and MST, as the time intensity of items is often correlated with their difficulty (Bridgeman & Cline, 2004; van der Linden, 2009b). This means that high-ability test-takers will often systematically receive more time intensive items than low-ability test-takers both in CAT and MST (Davey & Lee, 2011; van der Linden & Xiong, 2013). Furthermore, CAT and MST are frequently administered in fixed time slots with fixed time limits as well (Bridgeman & Cline, 2004). In a certain sense, high-ability test-takers are therefore punished for performing well in adaptive testing situations, an issue stated both in the literature for CAT (Bridgeman & Cline, 2004; Huang, 2019; van der Linden & Xiong, 2013) as well as MST (Davey & Lee, 2011).

Fortunately, all presented approaches for controlling speededness during test assembly in Chapter 3 are applicable to CAT and MST, just as they are to fixed-form linear tests. In MST, speededness constraints can simply be added during the assembly of modules[33]. For a review on how ATA methods blend with MST see, for example, Zheng et al. (2016). For a practical implementation see Chapter 6 of this thesis. In CAT, speededness constraints can be added within the shadow-test framework (van der Linden, 2005). The latter was already demonstrated for the original approach using the 2PLN model by van der Linden and Xiong (2013). Therefore, an extension to the 3PLN model as demonstrated in Chapter 3 seems straightforward.

7.3.2 Piloting Conditions

When items are piloted to inform the assembly of operational test forms, the goal is to obtain valid and reliable information for these items, such as IRT item parameters or response time model parameters. Such valid and reliable response time model parameter estimates are, for instance, vital for the approach to controlling speededness presented in Chapter 3. But how should pilot studies be designed to obtain valid and reliable item parameters? In fact, this question is rarely addressed in the literature so far. Chapter 3 briefly discussed the following two approaches: (1) Piloting conditions are chosen so they are as similar as possible

[33] Alternatively, parallelism can also be implemented on the higher *panel* or *pathway* level, following a *top-down* approach.

to operational testing conditions or (2) piloting conditions are chosen to minimize bias and construct-irrelevant influences on parameter estimates. In the following, the advantages and disadvantages of both approaches are discussed.

Similar Conditions. Choosing piloting conditions that closely mimic the conditions of the operational test aims at preventing item parameters to substantially differ between the pilot and the operational test. Indeed, if items are piloted in a low-stakes setting and are later administered in a high-stakes setting, test-takers may invest less effort in the pilot test compared to the operational test (Wise & DeMars, 2005). Thereby, item difficulty may be overestimated and item time intensity may be underestimated based on the piloting data. However, it is not always possible to exactly mimic operational testing conditions in pilot tests. In low-stakes assessments, such as educational large-scale assessments, this is often easy to implement, as both the pre-test as well as the operational test are low-stakes, anyway. However, for high-stakes assessments, this is more challenging, as it is unethical and/or difficult from a practical perspective to associate high-stakes for individuals with the outcome of a pilot test. To circumvent this problem, some high-stakes assessments pilot items as part of their operational test so test-takers perform and behave similarly both in the operational test and the pilot test (e.g., Educational Testing Service, 2010).

Ideal Conditions. In contrast, it can also be argued that, if items are piloted under operational test conditions (e.g., speeded), these conditions could lead to bias in item parameter estimates. As discussed before, conventional response and response time models assume that test-takers work with constant ability and speed throughout a (pilot) test. Yet, processes such as test speededness, fatigue, or declining effort can impact the ability and speed with which test-takers work throughout the test. A prominent phenomenon often researched in this context are the aforementioned item position effects (e.g., Debeer & Janssen, 2013; Weirich et al., 2017). To counter such effects, four measures have been proposed in Chapter 4 which shall be reiterated to emphasize their relevance for practical applications: (a) test-takers should be given sufficient time on the pilot test, (b) the pilot test should be sufficiently short so test-takers do not experience fatigue or a decrease in effort, (c) item positions should be balanced in the pilot test design so item position effects average out, and (d) models which take item position effects or varying speed levels into account could be used for the estimation of item parameters (e.g., Fox & Marianti, 2016; Hong et al., 2020).

Choosing appropriate piloting conditions is often also limited by practical constraints.

For instance, piloting items for high-stakes assessments in low-stakes settings can lead to confidentiality issues. Furthermore, research exists that seeks to reduce the necessity of piloting items in general. For instance, Baldwin et al. (2021) propose an approach suitable for predicting response times based on item surface properties. In theory, such approaches could be extended to predict the item parameters of the 3PLN model so these can be used in ATA, even if the items have not been actually piloted.

7.3.3 Speededness Control Reporting

In this thesis, it has been repeatedly argued that controlling the speededness of a test is crucial for the validity and fairness of an assessment. This emphasis is in line with various research on the topic (e.g., Cintron, 2021; Jurich, 2020; Kane, 2020; Y. Lu & Sireci, 2007). Unfortunately however, it stands in stark contrast to the common practice of testing programs. In fact, testing programs rarely if ever describe their test assembly strategies, including how they control the speededness of the assessment. In technical reports, information on test assembly is often either omitted completely or reduced to a few conceptual statements (e.g., College Board, 2015; Educational Testing Service, 2010; OECD, 2013, 2016a). It can be argued that testing programs and assessments in general should explicitly define their measured constructs and how they are related to speed as well as how speededness is controlled in test assembly. More precisely it is suggested that assessment designers explicitly communicate:

1. Is the construct their assessment measures related to speed (i.e., is speed an intentional parameter or nuisance factor)?
2. If speed is seen as an intentional parameter, to which degree should speed be part of the measured construct?
3. How are pilot studies conducted to inform the assembly of test forms (e.g., time limit, stakes for test-takers, instructions regarding time use)?
4. How is test speededness controlled during (automated) test assembly (e.g., van der Linden approach or approach presented in Chapter 3)?
5. If multiple, parallel test forms are used, are items balanced at the end of test forms to prevent differential effects of speededness (see recommendations provided in Chapter 4)?

Only if these aspects of test designs are communicated, researchers can determine whether an assessment appropriately considered speededness. Furthermore, transparent reporting of these aspects will certainly foster further research in the area of test speededness.

7.4 Directions for Future Research

In the present thesis, various perspectives on how to control the speededness of an assessment and how differential effects of speededness can be controlled were discussed. Future research could build upon the presented research in a variety of ways. In the following, an extensive (but not exhaustive) list of suggestions is given.

7.4.1 Defining Response Times

During large parts of this thesis (Chapters 2, 3, & 5) it was assumed that response times are available from item pre-testing. Indeed, this is a crucial limitation of the present thesis. If, for instance, items are pre-tested using paper-based assessments or response times are not collected via the testing software, the respective approaches for controlling speededness are simply not feasible. Unfortunately, even if response times are collected, their definition can vary from assessment to assessment. As described in Chapter 1, some researchers define response times as time-on-task (e.g., Goldhammer et al., 2020; OECD, 2022), others define response times as time until the last answer-change was performed (e.g., Kröhne & Goldhammer, 2018; Li et al., 2017). Even though the topic of response time modeling is trending in the psychometric literature (Becker et al., 2022; van Rijn & Sinharay, 2023), there is surprisingly little research on this technical definition of response times. Future research should investigate, which response time operationalization is most appropriate for psychometric purposes. Furthermore, certain administration conditions (e.g., preventing test-takers from revisiting items) might benefit clear operationalizations of response times. Such constraints might be especially attractive for pilot tests in which the goal is not necessarily a valid and reliable person parameter estimation but investigating the properties of items.

7.4.2 Defining Speededness

In his work, van der Linden (2011a, 2011b) defined test speededness as an interaction of the workload of a test, its time limit and the working speed of a test-taker. This definition was adopted throughout this thesis. More precisely, van der Linden (2011a, 2011b) defined the

degree of speededness[34] as the probability π of a test-taker to run out of time as[35]

$$\begin{aligned} \pi &= Pr\left\{ \sum_{j=1}^{k} RT_k > RT_{lim} \mid \zeta, \lambda, \phi, \sigma_\epsilon^2 \right\} \\ &= 1 - F_{RT_{tot}}(RT_{lim} \mid \zeta, \lambda, \phi, \sigma_\epsilon^2). \end{aligned} \tag{63}$$

$F_{RT_{tot}}$ denotes the cumulative distribution function of RT_{tot}, RT_{lim} the time limit set on a test. However, this speededness definition may be unsatisfactory for assessment practitioners as it (a) focuses on the displayed speed level of the test-taker and (b) does not provide a clear categorical definition of speededness. Both aspects warrant a further discussion and should be investigated in future research.

Displayed Speed Level. Assume a test-taker is administered a time intensive test with a strict time limit. If the test-taker chooses a fast and appropriate initial working speed, they are able to finish the test in time. Following the definition of van der Linden (2011b), the test-taker in this case was not exposed to a speeded test administration, as the test-taker was not running out of time. Yet, if the test administration was not speeded, the test-taker would probably have chosen a different, slower speed level leading to greater displayed ability. Therefore, an extended definition of speededness is proposed: Test speededness is defined as an interaction of the workload of a test, its time limit and the *maximum working speed of a test-taker at which they still achieve their maximum ability*. Ranger et al. (2021) refer to this maximum ability level of a test-taker as a test-taker's *capability*. If a test-taker has to increase their working speed to a point at which their displayed ability is reduced to finish the test within the time limit, a test administration should be considered speeded. Or in other words: If a test-taker had scored higher on a test had they been given more time, the test administration for this test-taker must be considered speeded.

This speededness definition implies that the only sufficient approaches for detecting speededness are test-retest approaches (Harik et al., 2018). Test-retest approaches refer to the administration of multiple, parallel test forms with different time limits to compare individual test-taker performance. If a shorter time limit leads to a lower ability estimate compared to a more lenient time limit for a test-taker, this means that the stricter time limit led to a speeded test administration for this specific test-taker. However, such a comparison does

[34] It can be argued that *degree of speededness* is unfortunate wording as this term could also refer to the amount of time pressure a test-taker experiences (i.e., how strongly do they have to adjust their speed level).

[35] Note that the item parameter ϕ has been added to Equation 63 as this thesis advocates for the use of the 3PLN model instead of the restricted 2PLN model. Also note that π simply constitutes the inverse probability of the quantile at RT_{lim}.

not imply that the more lenient time limit led to an unspeeded test administration. An unspeeded test administration can only be proven by a configuration in which additional time does not lead to additional gains in ability.

In contrast, Bridgeman (2020) suggests experiments with random assignment as the gold standard, such as performed in the studies by Evans and Reilly (1972), Bridgeman, Trapani, and Curley (2004), or Bridgeman, Cline, and Hessinger (2004). Unfortunately, experimental approaches have an important pitfall: They investigate whether a test is speeded for the average test-taker. However, this can mean little for an individual, slow test-taker, who will experience test speededness even if a test is not speeded for the average test-taker.

This newly proposed definition of speededness brings up a variety of research questions which should be addressed by future research, such as: Can test-takers accurately report whether they have experienced a speeded test administration (i.e., do they have insight into their own speed-ability trade-off)? How can pure power tests determine how time limits must be set to guarantee an unspeeded test administration for all test-takers? Can specific instructions during pilot testing (e.g., "Take as much time as you need.") substitute cost intensive test-retest approaches? And how can test-retest approaches be designed efficiently to predict test speededness for future test-takers?

Probabilistic Speededness Definition. Assume a test-taker with a fixed speed level could be presented the same test 100,000 times. As it is assumed that item response times are manifestations of the latent speed level confounded with measurement error, every test administration will yield a slightly different total test time. The resulting test time distribution may look like one of the distributions in Figure 9 (Chapter 3), with the majority of the test time distribution located between 8,500 and 10,000 seconds. Furthermore, assume that the time limit of the test is set at 9,300 seconds. Should the test be considered speeded for the specific speed level of the hypothetical test-taker? Empirically, the test is speeded for the test-taker approximately 50% of the time. If a test-retest approach is used to determine whether the test is speeded for this specific test-taker, measurement error can lead to substantially different conclusions. This may be unsatisfactory for assessment practitioners. In that sense, two approaches to determining speededness could be considered: (a) Repeated test-retest approaches, which allow to estimate how frequently a test is speeded for a certain speed level or (b) statistical modeling of the response times, for instance, via the 3PLN model. While repeated test-retest approaches seem hardly possible in practice, statistical modeling requires that researchers are able to instruct test-takers to work at the exact speed level that

allows them to work at their maximum capacity.

Future research should investigate the practical feasibility of such a statistical modeling approach for determining the speededness of a test administration for a specific speed level. Furthermore, future research could investigate whether this probabilistic notion of speededness may be translated into a categorical definition for practical applications. For instance, it could be argued that it suffices that running out if time is highly unlikely (e.g., $< 5\%$) for a speed level to consider a test unspeeded for this speed level.

7.4.3 Setting Speededness based on Risk Probabilities

In his work, van der Linden (2011b, p. 46) emphasizes that arguably the most relevant use case for being able to control speededness is that "...we may select a new test form to realize a desired level of risk π for a given time limit RT_{lim}." This could mean, for example, to assemble a test on which slow test-takers with $\zeta = -1$ should have a 90% probability of finishing the test within the time limit of 30 minutes. Alternatively, if a test is meant to be substantially speeded, test-takers with moderate working speed ($\zeta = 0$) should have a 50% probability of finishing the test within the time limit. However, both van der Linden (2011b) and Chapter 3 of this thesis focused on applications in which the cumulants of the desired response time distribution are already given, either because the aim is to create an additional test form (parallel to an existing test form) or to modify an existing test form. Furthermore, in his previous work, van der Linden (2011a) focused on determining an appropriate time limit RT_{lim} given a fixed test (containing items with item parameters λ and σ_ϵ^2) and a desired risk π for a specific speed level ζ to run out of time. Indeed, Equation 63 can be reformulated into Equation 64 to get the appropriate time limit RT_{lim} for a test (van der Linden, 2011a):

$$RT_{lim} = F^{-1}_{RT_{tot}}(1 - \pi \mid \zeta, \lambda, \phi, \sigma_\epsilon^2). \tag{64}$$

However, a frequent application is that a fixed time limit RT_{lim} is given (e.g., a certain time slot is already allocated for the assessment, or the assessment has been using a certain time limit before), alongside with a (maximum) desired risk π for test-takers to run out of time. Unfortunately, it is not straightforward to derive the cumulants of a respective test time distribution given Equations 63 and 64. Future research could investigate how a desired level of risk and a given time limit can, for a specific speed level, be translated into the first two cumulants of a response time distribution (mean and variance), which then can be constrained in ATA as illustrated in Chapter 3.

7.4.4 Understanding Speed Sensitivity

In Chapters 2 and 3, there was a strong emphasis on the fact that, empirically, items often differ regarding their speed sensitivity. However, there is currently little research on which item properties determine the speed sensitivity of items. For an exception, see the work of Vista and Alahmadi (2022) who compare speed sensitivities across sub-domains of cognitive ability. While item developers and test designers may have an intuitive understanding of factors influencing the difficulty (e.g., complexity, constructed response formats) or time intensity of items (e.g., text length), this is less trivial for slope or discrimination parameters. However, practically speaking, in some situations it could be beneficial for item developers to have insights into how the speed sensitivity of items can be manipulated.

It should be noted that, while items which discriminate well regarding ability are usually preferred to poorly discriminating items, the same may not be true for speed sensitivity. If items are highly sensitive for speed differences, this means that some test-taker need very little time on an item, while others need substantially larger amounts of time. In a lot of practical situations, test designers will prefer to have test-takers require similar amounts of time instead.

7.4.5 Different Response Time Models in ATA

The present thesis has focused on utilizing lognormal response time models, namely the 2PLN and 3PLN models in ATA. However, in the psychometric literature, a wide variety of different response time models is available (De Boeck & Jeon, 2019). Furthermore, by conventional, frequentist CFA standards, the lognormal response time model frequently exhibits poor model fit in practical applications (van Rijn & Sinharay, 2023). However, this is rarely explicitly investigated in the literature. Future research could investigate how different response time models can be used in the ATA framework to control the speededness of test forms. The central requirements for using other response time models in ATA are that (a) it is possible to calculate the cumulants of the expected response time distributions and that (b) a conditional independence assumption allows for these cumulants to be additive.

An alternative to approaches based on response time models is using predicted response times for item selection. For instance, in CAT, information is collected on the test-taker during the test and available during later stages of the test assembly process. In such applications, machine learning approaches may outperform classic response time modeling approaches in predicting response times.

7.4.6 Differential Item Functioning

A major concern for the validity and fairness of assessments is *differential item functioning* (DIF; e.g., de Ayala, 2022). DIF refers to items functioning differently in different subpopulations, for instance due to differential familiarity with specific item content. Usually, DIF focuses on differential difficulty of items. However, it seems plausible to assume that items can be differentially time intensive as well. For instance, research has shown that reading speed substantially differs between different languages and across the life span (Brysbaert, 2019). Therefore, international assessments which are administered in different languages or assessments administered across different age groups may be prone to the effect of differential item functioning regarding the time intensity of items (for similar ideas, see Lee & Haberman, 2016). For instance, Shin et al. (2020) illustrate using PISA data that measurement invariance does not hold across assessment languages and countries based on IRT analysis of categorized response times. If a test is differentially time intensive for test-takers, this can lead to differential speededness and unfair test administrations. Note that DIF regarding ability and speed may be often correlated in practice. If test-takers are familiar with the content of a specific item (e.g., test-takers are asked about the capital of Germany leading to DIF for German test-takers in comparison to other nationalities) this will often have an impact both on the speed and accuracy test-takers show on the item. Future research could investigate the prevalence and the consequences of DIF regarding time intensity and its relation to conventional DIF.

7.4.7 Extended Testing Time

A frequent challenge for test administrators is providing test accommodations related to time limits. A common practice is providing, for instance, *students with disabilities* (SWD) with extended testing times both in high-stakes assessments (Lovett, 2010) as well as low-stakes assessments (e.g., OECD, 2016b). This is mostly done out of fairness reasons. If a test is supposed to be unspeeded or only slightly speeded, SWD test-takers working at substantially lower speed levels would otherwise be at a disadvantage. In practice, these time accommodations are mostly categorical (e.g., standard time vs. 50% additional time; Cahan et al., 2016). However, it seems unlikely that the underlying construct (i.e., speed) is categorical (Gernsbacher et al., 2020). Therefore, researchers have argued that the current practice of test time accommodations is arbitrary and threatens the validity of assessments, as non-SWD test-takers would frequently benefit from extended test time as well (Cahan

et al., 2016).

An illustrative example of how arbitrary these test time accommodations are and how easily the current system can be gamed is the *2019 College Entrance Scandal* (United States Department of Justice, 2019). Essential part of this fraudulent scheme was non-SWD test-takers getting (undeserved) test time accommodations. However, these test time accommodations themselves were achieved without any bribes by simply repeatedly requesting the accommodations or seeking out lenient medical practitioners. Future research should investigate how more fine-grained test time accommodations could be provided to test-takers according to their true needs. For instance, while experimental investigations of the speed-ability trade-off may be too time-consuming and resource-intensive for all test-takers, such an approach could be suitable for determining test time accommodations.

7.4.8 Item Order in Tests with no Item Overlap

Chapter 4 focused on the effects of item order in test forms with identical items (i.e., full item overlap). In such instances, keeping the order of items constant across test forms (at least for specific parts of the test forms) is a viable solution. However, item order and speededness can also lead to unfair test forms if the test forms have little or no item overlap at all. For instance, test form A could still have mainly easy items at the end of the test while test form B could have mainly difficult items at the end of the test. Unfortunately, if test forms have no overlapping items, keeping identical items at the end of the test constant is not a viable option. Future research should embrace this challenge, for instance by using item pairs with comparable properties or assembling parallel test subsets which can be used at the end of tests. Examples for the concept of item pairs can already be found in the literature: Samejima (1977) refers to test forms consisting only of such parallel item pairs as *strongly parallel* test forms. Clause et al. (1998) propose that additional test forms can be constructed by specifically creating matching items, a procedure which the authors term *item-cloning*. Armstrong et al. (1992) and P.-H. Chen (2016) suggest that new test forms can be assembled parallel to an existing reference form using *pairwise item matching*. Various distance measures are proposed as means to identifying item pairs. Already Gulliksen (1950) proposed a method called *matched random subtests*, in which item pairs are created based on their difficulty and discrimination. For an adaptation of the method to MILP see van der Linden and Boekkooi-Timminga (1988). While none of these approaches were designed to deal with speededness issues it seems plausible that these could easily be extended in that

direction.

7.5 Conclusion

Controlling the speededness of test forms and the impact of such speededness is crucial for the fairness and validity of assessments. In this thesis, it has been demonstrated that the state-of-the-art approach for controlling speededness by van der Linden (2011a, 2011b) neglects that items usually differ in their speed sensitivity and that this can be detrimental to the fairness of test forms. A generalized approach incorporating differing speed sensitivities via the 3PLN model and thereby overcoming this limitation has been proposed. Moreover, it has been demonstrated how response time models can be estimated flexibly in `stan` as well as how ATA can be performed in the R package `eatATA`. Additionally, it has been illustrated how item order can be crucial if multiple, speeded test forms are used. With the provided tools, including a new software package as well as extensive tutorial material, assessment practitioners should be able to use the proposed approaches in practice.

References

Aamodt, M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management*, *21*(2), 151–160. https://doi.org/10.1177/009102609202100203

American Educational Research Association, American Psychological Association, & National Council on Educational Measurement. (2014). *Standards for educational and psychological testing*. AERA.

Armstrong, R. D., Belov, D., & Weissman, A. (2005). Developing and assembling the law school admission test. *Interfaces*, *35*(2), 140–151. https://doi.org/10.1287/inte.1040.0123

Armstrong, R. D., Jones, D. H., & Wu, I.-L. (1992). An automated test development of parallel tests from a seed test. *Psychometrika*, *57*(2), 271–288. https://doi.org/10.1007/bf02294509

Baldwin, P., Yaneva, V., Mee, J., Clauser, B. E., & Ha, L. A. (2021). Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, *58*(1), 4–30. https://doi.org/10.1111/jedm.12264

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*(4), 1281–1311.

Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. *Journal of Educational Psychology*, *32*(4), 285–296. https://doi.org/10.1037/h0061115

Becker, B. (2020). *pisaRT: Small example response and response time data from PISA 2018* [R package version 2.0.1]. https://CRAN.R-project.org/package=pisaRT

Becker, B., Debeer, D., Sachse, K. A., & Weirich, S. (2021). Automated test assembly in R: The eatATA package. *Psych*, *3*(2), 96–112. https://doi.org/10.3390/psych3020010

Becker, B., Debeer, D., Weirich, S., & Goldhammer, F. (2021). On the speed sensitivity parameter in the lognormal model for response times and implications for high-stakes measurement practice. *Applied Psychological Measurement*, *45*(6), 407–422. https://doi.org/10.1177/01466216211008530

Becker, B., Neuendorf, C., & Jansen, M. (2022). *Nutzung von Logdaten in der empirische Bildungsforschung - Eine Bedarfsanalyse* (KonsortSWD Working Paper No. 2022-3). https://doi.org/10.5281/zenodo.7030995

Bejar, I. I. (1985, June). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (ETS Research Report Series No. RR·85-11). Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00096.x

Belov, D. I., & Chen, P.-H. (2014). Review of automated test assembly. In Y. Cheng & H.-H. Chang (Eds.), *Advancing methodologies to support both summative and formative assessments* (pp. 3–19). Information Age Publishing.

Berkelaar, M., & Csárdi, G. (2020). *lpSolve: Interface to 'Lp_solve' v. 5.5 to solve linear/integer programs* [R package version 5.6.15]. https://CRAN.R-project.org/package=lpSolve

Berkelaar, M., Eikland, K., & Notebaert, P. (2016). *Lp_solve* (Software; Version 5.5.2.5). http://lpsolve.sourceforge.net/5.5/

Bernardi, R. A., Baca, A. V., Landers, K. S., & Witek, M. B. (2008). Methods of cheating and deterrents to classroom cheating: An international study. *Ethics & Behavior*, *18*(4), 373–391. https://doi.org/10.1080/10508420701713030

Bertling, M., & Weeks, J. P. (2018). Using response time data to reduce testing time in cognitive tests. *Psychological Assessment*, *30*(3), 328–338. https://doi.org/10.1037/pas0000466

Betancourt, M. (2016). *Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo*. arXiv Preprint. https://doi.org/10.48550/arXiv.1604.00695

Bezirhan, U., von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: The hierarchical speed–accuracy–revisits model. *Educational and Psychological Measurement*, *81*(2), 363–387. https://doi.org/10.1177/0013164420950556

Bolsinova, M. (2016). *Balancing simple models and complex reality* [Doctoral dissertation, Utrecht University]. https://dspace.library.uu.nl/handle/1874/340002

Bolsinova, M., de Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bolsinova, M., & Tijmstra, J. (2015). Can response speed be fixed experimentally, and does this lead to unconfounded measurement of ability? *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 165–168. https://doi.org/10.1080/15366367.2015.1105080

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, *41*(2), 123–145. https://doi.org/10.3102/1076998616631746

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13–38. https://doi.org/10.1111/bmsp.12104

Bolsinova, M., Tijmstra, J., Molenaar, D., & Boeck, P. D. (2017). Conditional dependence between response time and accuracy: An overview of its possible sources and directions for distinguishing between them. *Frontiers in Psychology*, *8*, 1–6. https://doi.org/10.3389/fpsyg.2017.00202

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). Springer. https://doi.org/10.1007/978-0-387-49839-3_9

Bridgeman, B. (2020). Relationship between testing time and testing outcomes. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 59–72). Routledge. https://doi.org/10.4324/9781351064781-5

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement*, *41*(2), 137–148. https://doi.org/10.1111/j.1745-3984.2004.tb01111.x

Bridgeman, B., Cline, F., & Hessinger, J. (2004). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education*, *17*(1), 25–37.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*(4), 291–310.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 1–30. https://doi.org/10.1016/j.jml.2019.104047

Budescu, D. V., & Bo, Y. (2015). Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, *80*(4), 1105–1122. https://doi.org/10.1007/s11336-014-9425-x

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Cahan, S., Nirel, R., & Alkoby, M. (2016). The extra-examination time granting policy: A reconceptualization. *Journal of Psychoeducational Assessment*, *34*(5), 461–472. https://doi.org/10.1177/0734282915616537

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*(2), 209–230. https://doi.org/10.1007/s11336-007-9045-9

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Carroll, J. B. (1993, January). Abilities in the domain of cognitive speed. In J. B. Carroll (Ed.), *Human cognitive abilities* (pp. 440–509). Cambridge University Press. https://doi.org/10.1017/cbo9780511571312.012

Chang, J., Tsai, H., Su, Y.-H., & Lin, E. M. H. (2016). A three-parameter speeded item response model: Estimation and application. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 27–38). Springer Proceedings in Mathematics & Statistics (Vol. 167). https://doi.org/10.1007/978-3-319-38759-8_3

Chang, T.-Y., & Shiu, Y.-F. (2012). Simultaneously construct IRT-based parallel tests based on an adapted CLONALG algorithm. *Applied Intelligence*, *36*(4), 979–994. https://doi.org/10.1007/s10489-011-0308-x

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika*, *79*(2), 255–274. https://doi.org/10.1007/s11336-013-9336-2

Chen, H., Boeck, P. D., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence*, *69*, 16–23. https://doi.org/10.1016/j.intell.2018.04.001

Chen, H. (2012). The moderating effects of item order arranged by difficulty on the relationship between test anxiety and test performance. *Creative Education*, *3*(3), 328–333. https://doi.org/10.4236/ce.2012.33052

Chen, P.-H. (2016). Three-element item selection procedures for multiple forms assembly. *Applied Psychological Measurement*, *40*(2), 114–127. https://doi.org/10.1177/0146621615605307

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior research methods*, *49*(2), 502–512. https://doi.org/10.3758/s13428-016-0712-6

Chidomere, R. C. (1989). Test item arrangement and student performance in principles of marketing examination: A replication study. *Journal of Marketing Education*, *11*(3), 36–40. https://doi.org/10.1177/027347538901100307

Chirumamilla, A., Sindre, G., & Nguyen-Duc, A. (2020). Cheating in e-exams and paper exams: The perceptions of engineering students and teachers in Norway. *Assessment & Evaluation in Higher Education*, *45*(7), 940–957. https://doi.org/10.1080/02602938.2020.1719975

Chittka, L., Dyer, A. G., Bock, F., & Dornhaus, A. (2003). Bees trade off foraging speed for accuracy. *Nature*, *424*(6947), 388–388. https://doi.org/10.1038/424388a

Choe, E. M., Kern, J. L., & Chang, H.-H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*(2), 135–158. https://doi.org/10.3102/1076998617723642

Choi, S. W., & Lim, S. (2020). *TestDesign: Optimal test design approach to fixed and adaptive test construction* [R package version 1.1.3]. https://CRAN.R-project.org/package=TestDesign

Choi, S. W., Lim, S., & van der Linden, W. J. (2021). TestDesign: An optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika*, *49*(2), 191–229. https://doi.org/10.1007/s41237-021-00145-9

Cintron, D. W. (2021). *Methods for measuring speededness: Chronology, classification, and ensuing research and development* (ETS Research Report Series No. RR–21-22). Educational Testing Service. https://doi.org/10.1002/ets2.12337

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology*, *51*(1), 193–208. https://doi.org/10.1111/j.1744-6570.1998.tb00722.x

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119–130. https://doi.org/10.2307/27798099

College Board. (2015). *Test specifications for the redesigned SAT*. https://satsuite.collegeboard.org/media/pdf/test-specifications-redesigned-sat-1.pdf

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H. L. J., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy - response time data. *PLOS ONE*, *11*(5), e0155149. https://doi.org/10.1371/journal.pone.0155149

Craigmile, P. F., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika*, *75*(4), 613–632. https://doi.org/10.1007/s11336-010-9172-6

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In E. Care, P. Griffin, & B. McGaw (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). Springer.

Davey, T., & Lee, Y.-H. (2011, December). *Potential impact of context effects on the scoring and equating of the multistage GRE-revised general test* (ETS GRE Board Research Report No. GREB–08-01). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02262.x

Davis, D. B. (2017). Exam question sequencing effects and context cues. *Teaching of Psychology*, *44*(3), 263–267. https://doi.org/10.1177/0098628317712755

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*(8), 583–603. https://doi.org/10.1177/0146621611428446

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*, 1–11. https://doi.org/10.3389/fpsyg.2019.00102

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd Ed.). Guilford Press.

Debeer, D., Ali, U. S., & van Rijn, P. W. (2017). Evaluating statistical targets for assembling parallel mixed-format test forms. *Journal of Educational Measurement*, *54*(2), 218–242. https://doi.org/10.1111/jedm.12142

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

Debelak, R., Gittler, G., & Arendasy, M. (2014). On gender differences in mental rotation processing speed. *Learning and Individual Differences*, *29*, 8–17. https://doi.org/10.1016/j.lindif.2013.10.003

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. Oxford University Press.

Devine-Eller, A. (2012). Timing matters: Test preparation, race, and grade level. *Sociological Forum*, *27*(2), 458–480. https://doi.org/10.1111/j.1573-7861.2012.01326.x

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398–409. https://doi.org/10.1177/0146621610392211

Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory Into Practice*, *55*(2), 153–159. https://doi.org/10.1080/00405841.2016.1148989

Dodeen, H. (2008). Assessing test-taking strategies of university students: Developing a scale and estimating its psychometric indices. *Assessment & Evaluation in Higher Education*, *33*(4), 409–419. https://doi.org/10.1080/02602930701562874

Dolin, J., Black, P., Harlen, W., & Tiberghien, A. (2018, October). Exploring relations between formative and summative assessment. In J. Dolin & R. Evans (Eds.), *Transforming assessment: Contributions from science education research (vol. 4)* (pp. 53–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-63248-3_3

Domingue, B. W., Kanopka, K., Stenhaug, B., Soland, J., Kuhfeld, M., Wise, S., & Piech, C. (2021). Variation in respondent speed and its implications: Evidence from an adaptive testing scenario. *Journal of Educational Measurement*, *58*(3), 335–363. https://doi.org/10.1111/jedm.12291

Domingue, B. W., Kanopka, K., Stenhaug, B., Sulik, M. J., Beverly, T., Brinkhuis, M., Circi, R., Faul, J., Liao, D., McCandliss, B., Obradovic, J., Piech, C., Porter, T., Projekt ILEAD Consortium, Soland, J., Weeks, J., Wise, S. L., & Yeatman, J. (2022). Speed–accuracy trade-off? Not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real-world settings. *Journal of Educational and Behavioral Statistics*, *47*(5), 576–602. https://doi.org/10.3102/10769986221099906

Donoghue, J. R. (2015). *Comparison of integer programming (IP) solvers for automated test assembly (ATA)* (Research Report No. RR-15-05). Educational Testing Service. https://doi.org/10.1002/ets2.12051

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*(2), 129–151. https://doi.org/10.2307/1165318

Educational Testing Service. (2010). *TOEFL iBT test framework and test development volume 1 (1st ed.)* in TOEFL iBT Research Insight Series 1.

Educational Testing Service. (2020). *TOEFL iBT test framework and test development volume 1 (3rd ed.)* in TOEFL iBT Research Insight Series.

Ellis, A. P., & Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-efficacy. *Journal of Applied Social Psychology*, *33*(12), 2607–2629. https://doi.org/10.1111/j.1559-1816.2003.tb02783.x

Elton, L. (2004). A challenge to established assessment practice. *Higher Education Quarterly*, *58*(1), 43–62. https://doi.org/10.1111/j.1468-2273.2004.00259.x

Entink, R. H. K., Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 621–640. https://doi.org/10.1348/000711008x374126

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment*, *25*(3), 179–197. https://doi.org/10.1080/10627197.2020.1804353

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*(2), 123–131. https://doi.org/10.1111/j.1745-3984.1972.tb00767.x

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*(5), 655–670. https://doi.org/10.3102/1076998611422912

Fenton, L. F. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, *8*(1), 57–67. https://doi.org/10.1109/tcom.1960.1097606

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525–543. https://doi.org/10.1177/0146621606295197

Finch, H. (2008). Estimation of Item Response Theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Finkelman, M., de la Torre, J., & Karp, J. A. (2020). Cognitive diagnosis models and automated test assembly: An approach incorporating response times. *International Journal of Testing*, *20*(4), 299–320. https://doi.org/10.1080/15305058.2020.1828427

Finkelman, M., Kim, W., Weissman, A., & Cook, R. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing*, *2*(3), 59–76. https://doi.org/10.7333/1412-0204059

Förster, J., Higgins, E., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organizational Behavior and Human Decision Processes*, *90*(1), 148–164. https://doi.org/10.1016/s0749-5978(02)00509-5

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Fox, J.-P. (2019). Course LNIRT: Modeling response accuracy and response times. https://www.jean-paulfox.com/wp-content/uploads/2018/01/LNIRT-Demo-2019.pdf

Fox, J.-P., Klein Entink, R., & van der Linden, W. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*(7), 1–14. https://doi.org/10.18637/jss.v020.i07

Fox, J.-P., Klotzke, K., & Entink, R. K. (2021). *LNIRT: Lognormal response time item response theory models* [R package version 0.5.1]. https://CRAN.R-project.org/package=LNIRT

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*(2), 243–262. https://doi.org/10.1111/jedm.12143

Franks, N. R., Dornhaus, A., Fitzsimmons, J. P., & Stevens, M. (2003). Speed versus accuracy in collective decision making. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1532), 2457–2463. https://doi.org/10.1098/rspb.2003.2527

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Frey, A., Spoden, C., & Born, S. (2020). Construction of psychometrically sound written university exams. *Psychological Test and Assessment Modeling*, *65*(4), 472–486.

Gabry, J., & Mahr, T. (2022). *Bayesplot: Plotting for Bayesian models* [R package version 1.9.0]. https://mc-stan.org/bayesplot/

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402. https://doi.org/https://doi.org/10.1111/rssa.12378

Gafni, N., & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation*, *20*(3), 309–19. https://doi.org/10.1016/0191-491x(94)90018-3

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511. https://doi.org/10.1214/ss/1177011136

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman; Hall/CRC.

Gernsbacher, M. A., Soicher, R. N., & Becker-Blease, K. A. (2020). Four empirically based reasons not to administer time-limited tests. *Translational Issues in Psychological Science*, *6*(2), 175–190. https://doi.org/10.1037/tps0000232

Geyer, C. J. (2011). Introduction to markov chain monte carlo. In S. Brooks, A. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 3–48). Chapman & Hall/CRC.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907–922. https://doi.org/10.1177/0013164408315262

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247–261. https://doi.org/10.1177/0146621603027004001

Goegebeur, Y., Boeck, P. D., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*(1), 65–87. https://doi.org/10.1007/s11336-007-9031-2

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*, 133–164. https://doi.org/10.1080/15366367.2015.1100020

Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). Analysing log file data from PIAAC. In D. B. Maehler & B. Rammstedt (Eds.), *Methodology of educational measurement and assessment* (pp. 239–269). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_10

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*(2), 108–119. https://doi.org/10.1016/j.intell.2011.02.001

Goldhammer, F., Kroehne, U., Hahnel, C., & Boeck, P. D. (2021). Controlling speed in component skills of reading improves the explanation of reading comprehension. *Journal of Educational Psychology*, *113*(5), 861–878. https://doi.org/10.1037/edu0000655

Goldhammer, F., & Kröhne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: Experimental time limits, posterior time limits, and response time modeling. *Applied Psychological Measurement*, *38*(4), 255–267. https://doi.org/10.1177/0146621613517164

Goldhammer, F., Steinwascher, M. A., Kröhne, U., & Naumann, J. (2017). Modelling individual response time effects between and within experimental speed conditions: A GLMM approach for speeded tests. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 238–256. https://doi.org/10.1111/bmsp.12099

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet design and parameter recovery in large-scale assessments. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments: Volume 3* (pp. 125–156). IEA-ETS Research Institute.

Gulek, C. (2003). Preparing for high-stakes testing. *Theory Into Practice*, *42*(1), 42–50. https://doi.org/10.1207/s15430421tip4201_6

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley & Sons. https://doi.org/10.1037/13240-000

Gurobi Optimization, LLC. (2021a). Gurobi optimizer reference manual version 9.1. https://www.gurobi.com/wp-content/plugins/hd_documentations/documentation/9.1/refman.pdf

Gurobi Optimization, LLC. (2021b). *Gurobi: Gurobi optimizer 9.1 interface* [R package version 9.1-1]. http://www.gurobi.com

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., Jodoin, M., Walsh, W., & Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*(2), 308–327. https://doi.org/10.1111/jedm.12177

Harter, R., Hornik, K., & Theussl, S. (2020). *Rsymphony: SYMPHONY in R* [R package version 0.1-29]. https://CRAN.R-project.org/package=Rsymphony

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 1–19. https://doi.org/10.3389/fnins.2014.00150

Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, *76*(3), 616–628. https://doi.org/10.1016/j.neuron.2012.08.030

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186–201. https://doi.org/10.1037/h0074579

Hong, M., Rebouças, D. A., & Cheng, Y. (2020). Robust estimation for response time modeling. *Journal of Educational Measurement*, *58*(2), 262–280. https://doi.org/10.1111/jedm.12286

Howell, W. C., & Kreidler, D. L. (1963). Information processing under contradictory instructional sets. *Journal of Experimental Psychology*, *65*(1), 39–46. https://doi.org/10.1037/h0038982

Huang, H.-Y. (2019). Utilizing response times in cognitive diagnostic computerized adaptive testing under the higher-order deterministic input, noisy 'and' gate model. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 109–141. https://doi.org/10.1111/bmsp.12160

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Jiang, B. (2020). *RSCAT: Shadow-test approach to computerized adaptive testing* [R package version 1.1.0]. https://CRAN.R-project.org/package=RSCAT

Jin, K.-Y., & Wang, W.-C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, *51*(2), 178–200. https://doi.org/10.1111/jedm.12041

Jones, L. V., & Thissen, D. (2006). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics - psychometrics* (pp. 1–27). Elsevier. https://doi.org/10.1016/s0169-7161(06)26001-2

Jurich, D. P. (2020). A history of test speededness: Tracing the evolution of theory and practice. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 1–18). Routledge. https://doi.org/10.4324/9781351064781-1

Kahraman, N., Cuddy, M. M., & Clauser, B. E. (2013). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement*, *37*(5), 343–360. https://doi.org/10.1177/0146621613477236

Kane, M. (2020). The impact of time limits and timing information on validity. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 19–31). Routledge. https://doi.org/10.4324/9781351064781-2

Kaneko, H., Tamura, H., Kawashima, T., & Suzuki, S. S. (2006). A choice reaction-time task in the rat: A new model using air-puff stimuli and lever-release responses. *Behavioural brain research*, *174*(1), 151–159. https://doi.org/10.1016/j.bbr.2006.07.020

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues* (9th ed.). Cengage Learning.

Kasli, M., Zopluoglu, C., & Toton, S. L. (2022). A deterministic gated lognormal response time model to identify examinees with item preknowledge. *Journal of Educational Measurement*, Advance online publication. https://doi.org/10.1111/jedm.12340

Kim, S. (2021). Prepping for the TOEFL iBT writing test, Gangnam style. *Assessing Writing*, *49*, 1–11. https://doi.org/10.1016/j.asw.2021.100544

Kingston, N. M., & Kramer, L. B. (2013). High-stakes test construction and test use. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (2nd ed., pp. 189–205). https://doi.org/10.1093/oxfordhb/9780199934874.013.0010

Kippel, G. M. (1985). Use of forms C and D of the California Achievement Test as equivalent forms. *Psychological Reports*, *57*(3), 1049–1050. https://doi.org/10.2466/pr0.1985.57.3f.1049

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013, July). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11). Springer. https://doi.org/10.1007/978-94-007-4629-9_1

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54–75. https://doi.org/10.1037/a0014877

Klinkenberg, S., Straatemeier, M., & van der Maas, H. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Knapp, R. R. (1960). The effects of time limits on the intelligence test performance of Mexican and American subjects. *Journal of Educational Psychology*, *51*(1), 14–20. https://doi.org/10.1037/h0038366

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer New York. https://doi.org/10.1007/978-1-4939-0317-7

König, C., Becker, B., & Ulitzsch, E. (2023). Bayesian hierarchical response time modelling—a tutorial. *British Journal of Mathematical and Statistical Psychology*, Advance online publication. https://doi.org/10.1111/bmsp.12302

König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement*, *44*(4), 311–326. https://doi.org/10.1177/0146621619893786

Konis, K., & Schwendinger, F. (2020). *lpSolveAPI: R interface to 'lp_solve' version 5.5.2.0* [R package version 5.5.2.0-17.7]. https://CRAN.R-project.org/package=lpSolveAPI

Kotz, S., Balakrishnan, N., Read, C. B., & Vidakovic, B. (2005). *Encyclopedia of statistical sciences*. John Wiley & Sons.

Kröhne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application

to questionnaire items. *Behaviormetrika*, *45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

Kuhn, J.-T., & Kiefer, T. (2015). Optimal test assembly in practice. *Zeitschrift für Psychologie*, *221*(3), 190–200. https://doi.org/10.1027/2151-2604/a000146

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-l. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*(2), 179–188. https://doi.org/10.1037/0033-2909.95.2.179

Kyllonen, P., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*, 1–29. https://doi.org/10.3390/jintelligence4040014

Ladanyi, L., Ralphs, T., Menal, G., & Mahajan, A. (2019). *coin-or/SYMPHONY* (Version 5.6.17). https://projects.coin-or.org/SYMPHONY

Lawrence, I. M. (1993, December). *The effect of test speededness on subgroup performance* (ETS Research Report Series No. RR 93-49). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01560.x

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, *55*(3), 387–413. https://doi.org/10.3102/00346543055003387

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359.

Lee, Y.-H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, *16*(3), 240–267. https://doi.org/10.1080/15305058.2015.1085385

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time data as validity evidence: Has it lived up to its promise and, if not, what would it take to do so. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 159–177). Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5_9

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Harvard University Press.

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, *10*, 1–15. https://doi.org/10.3389/fpsyg.2019.01533

List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., & Nagy, G. (2017). Performance decline in low-stakes educational assessments: Different mixture modeling approaches. *Large-scale Assessments in Education*, *5*, 1–15. https://doi.org/10.1186/s40536-017-0049-3

Liu, T., Wang, C., & Xu, G. (2022). Estimating three- and four-parameter MIRT models with importance-weighted sampling enhanced variational auto-encoder. *Frontiers in Psychology*, *13*, 1–19. https://doi.org/10.3389/fpsyg.2022.935419

Llabre, M. M., & Froman, T. W. (1987). Allocation of time to test items: A study of ethnic differences. *The Journal of Experimental Education*, *55*(3), 137–140. https://doi.org/10.1080/00220973.1987.10806446

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*(1), 7–11. https://doi.org/10.1111/j.1745-3984.1975.tb01003.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Information Age Publishing.

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, *80*(4), 611–638. https://doi.org/10.3102/0034654310364063

Lu, J., Wang, C., & Shi, N. (2021). A mixture response time process model for aberrant behaviors and item nonresponses. *Multivariate Behavioral Research*, Advance online publication. https://doi.org/10.1080/00273171.2021.1948815

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37. https://doi.org/10.1111/j.1745-3992.2007.00106.x

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing* (Research Report 2011-12). College Board.

Luo, X. (2019a). *Rata: Automated test assembly* [R package version 0.0.2]. https://CRAN.R-project.org/package=Rata

Luo, X. (2019b). *xxIRT: Item response theory and computer-based testing in R* [R package version 2.1.2]. https://CRAN.R-project.org/package=xxIRT

Luo, X. (2020). Automated test assembly with mixed-integer programming: The effects of modeling approaches and solvers. *Journal of Educational Measurement*, *57*(4), 547–565. https://doi.org/10.1111/jedm.12262

Makhorin, A. (2018). *GNU linear programming kit (GLPK)* (Version 4.65). http://www.gnu.org/software/glpk/glpk.html

Man, K., Harring, J. R., Ouyang, Y., & Thomas, S. L. (2018). Response time based nonparametric Kullback-Leibler divergence measure for detecting aberrant test-taking behavior. *International Journal of Testing*, *18*(2), 155–177. https://doi.org/10.1080/15305058.2018.1429446

Man, K., & Harring, J. R. (2020). Assessing preknowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, *81*(3), 441–465. https://doi.org/10.1177/0013164420968630

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. https://doi.org/10.1007/s11336-012-9288-y

Martin, M. O., Mullis, I. V., & Hooper, M. (2017). *Methods and procedures in PIRLS 2016.* International Association for the Evaluation of Educational Achievement (IEA).

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). *Methods and procedures: TIMSS 2019 technical report.* International Association for the Evaluation of Educational Achievement (IEA).

McDonald, R. P. (2013). Modern test theory. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (2nd ed., pp. 118–143). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934874.013.0007

McKeachie, W. J., Pollie, D., & Speisman, J. (1955). Relieving anxiety in classroom examinations. *The Journal of Abnormal and Social Psychology*, *50*(1), 93–98. https://doi.org/10.1037/h0046560

Melikyan, Z. A., Agranovich, A. V., & Puente, A. E. (2019). Fairness in psychological testing. In G. Goldstein, D. Allen, & J. DeLuca (Eds.), *Handbook of psychological assessment* (4th ed., pp. 551–572). Academic Press.

Melikyan, Z. A., Puente, A. E., & Agranovich, A. V. (2021). Cross-cultural comparison of rural healthy adults: Russian and American groups. *Archives of Clinical Neuropsychology*, *36*(3), 359–370. https://doi.org/10.1093/arclin/acz071

Messick, S. (1993, December). *Foundations of validity: Meaning and consequences in psychological assessment* (ETS Research Report Series No. RR-93-51). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01562.x

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538. https://doi.org/10.1177/0146621609355451

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*(3), 707–726. https://doi.org/10.1177/001316446502500304

Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279–297. https://doi.org/10.1007/s11336-017-9602-9

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). Fitting diffusion item response theory models for responses and response times using the R package diffIRT. *Journal of Statistical Software*, *66*(4), 1–34. https://doi.org/10.18637/jss.v066.i04

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015a). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015b). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219. https://doi.org/10.1111/bmsp.12042

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*(3), 291–315. https://doi.org/10.1007/bf02289044

Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. *Educational and Psychological Measurement*, *20*(2), 223–230. https://doi.org/10.1177/001316446002000203

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *The Journal of Educational Research*, *63*(10), 463–465.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Pearson Education International.

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling*, *60*(2), 165–187.

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*, Advance online publication. https://doi.org/10.1111/jcal.12719

Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, *53*, 1–16. https://doi.org/10.1016/j.lindif.2016.10.002

Neely, D. L., Springston, F. J., & McCann, S. J. H. (1994). Does item order affect performance on multiple-choice exams? *Teaching of Psychology*, *21*(1), 44–45. https://doi.org/10.1207/s15328023top2101_10

OECD. (2013). *Technical report of the survey of adult skills PIAAC (second edition).*

OECD. (2014). *PISA 2012 technical report.*

OECD. (2016a). *PISA 2015 assessment and analytical framework.*

OECD. (2016b). *PISA 2015 technical report.*

OECD. (2019a). *PISA 2018 assessment and analytical framework.*

OECD. (2019b). *PISA 2018 technical report.*

OECD. (2022). *PISA 2018 technical report: Annex K.*

Ong, T. Q., & Pastor, D. A. (2022). Uncovering the complexity of item position effects in a low-stakes testing context. *Applied Psychological Measurement*, *46*(7), 571–588. https://doi.org/10.1177/01466216221108134

Ophoff, J. G., & Cramer, C. (2022, January). The engagement of teachers and school leaders with data, evidence and research in Germany. In C. Brown & J. R. Malin (Eds.), *The Emerald handbook of evidence-informed practice in education* (pp. 175–195). Emerald Publishing Limited. https://doi.org/10.1108/978-1-80043-141-620221026

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200–219. https://doi.org/10.1111/j.1745-3984.1994.tb00443.x

Pachella, R. G., & Pew, R. W. (1968). Speed-accuracy tradeoff in reaction time: Effect of discrete criterion times. *Journal of Experimental Psychology*, *76*(1), 19–24. https://doi.org/10.1037/h0021275

Pant, H. A. (2013). Wer hat einen Nutzen von Kompetenzmodellen? *Zeitschrift für Erziehungswissenschaft*, *16*(S1), 71–79. https://doi.org/10.1007/s11618-013-0388-y

Partchev, I., Boeck, P. D., & Steyer, R. (2011). How much power and speed is measured in this test? *Assessment*, *20*(2), 242–252. https://doi.org/10.1177/1073191111411658

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189–212. https://doi.org/10.1080/10627197.2019.1615373

Patton, J. M. (2015). *Some consequences of response time model misspecification in educational measurement* [Doctoral dissertation, University of Notre Dame]. https://www.proquest.com/docview/1649186920

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology*, *39*(4), 299–307. https://doi.org/10.1037/h0086821

Pettit, K. L., Baker, K. G., & Davis, L. D. (1986). Unconscious biasing of student examination scores: A case of sequential versus random information retrieval. *Journal of Marketing Education*, *8*(3), 20–24. https://doi.org/10.1177/027347538600800306

Plummer, M. (2016). *rjags: Bayesian graphical models using MCMC* [R package version 4-6]. https://CRAN.R-project.org/package=rjags

Plummer, M. (2017). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling* (Version 4.3.0). https://mcmc-jags.sourceforge.io/

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11. https://journal.r-project.org/articles/RN-2006-002/RN-2006-002.pdf

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, *26*(4), 909–950. https://doi.org/10.1002/pam.20292

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report - Scaling the data of the competence tests* (NEPS Working Paper No. 14). Nationales Bildungspanel. https://fis.uni-bamberg.de/handle/uniba/1751

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study - Many questions, some answers and further challenges. *Journal for Educational Research Online*, *5*(2), 189–216. https://doi.org/10.25656/01:8430

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika*, *84*(3), 892–920. https://doi.org/10.1007/s11336-019-09669-2

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science*, *372*(6540), 338–340. https://doi.org/10.1126/science.abd3300

Pokropek, A. (2011). Missing by design: Planned missing-data designs in social science. *Research & Methods*, *20*(1), 81–105.

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902. https://doi.org/10.1214/12-BA730

Pools, E. (2022). Not-reached items: An issue of time and of test-taking disengagement? The case of PISA 2015 reading data. *Applied Measurement in Education*, *35*(3), 197–221. https://doi.org/10.1080/08957347.2022.2103136

Powers, D. E. (1985). Effects of coaching on GRE aptitude test scores. *Journal of Educational Measurement*, *22*(2), 121–136. https://doi.org/10.1111/j.1745-3984.1985.tb01052.x

Powers, D. E. (2017). Understanding the impact of special preparation for admissions tests. In R. E. Bennett & M. von Davier (Eds.), *Methodology of educational measurement and assessment* (pp. 553–564). Springer. https://doi.org/10.1007/978-3-319-58689-2_17

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement*, *36*(2), 93–118. https://doi.org/10.1111/j.1745-3984.1999.tb00549.x

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ranger, J., Kuhn, J.-T., & Pohl, S. (2021). Effects of motivation on the accuracy and speed of responding in tests: The speed-accuracy tradeoff revisited. *Measurement: Interdisciplinary Research and Perspectives*, *19*(1), 15–38. https://doi.org/10.1080/15366367.2020.1750934

Ranger, J., & Ortner, T. (2012a). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*(2), 128–148.

Ranger, J., & Ortner, T. (2012b). A latent trait model for response times on tests employing the proportional hazard model. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 334–349. https://doi.org/10.1111/j.2044-8317.2011.02032.x

Rasch, G. (1960). *Studies in mathematical psychology: Probabilistic models for some intelligence and attainment tests (Volume 1)*. Nielsen & Lydiche.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*(4), 261–270. https://doi.org/10.1111/j.1745-3984.1979.tb00107.x

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85–106. https://doi.org/10.31234/osf.io/hvurb

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules* [R package version 2.8-21]. https://CRAN.R-project.org/package=TAM

Robitzsch, A., & Lüdtke, O. (2022). Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Measurement Instruments for the Social Sciences*, *4*, 1–20. https://doi.org/10.1186/s42409-022-00039-w

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, *12*(3), 247–259. https://doi.org/10.1027/1015-5759.12.3.247

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*(3), 795–819. https://doi.org/10.1007/s11336-016-9544-7

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15–22. https://doi.org/10.1111/j.1745-3984.1977.tb00024.x

Russell, M., Fischer, M. J., Fischer, C. M., & Premo, K. (2003). Exam question sequencing effects on marketing and management sciences student performance. *Journal for Advancement of Marketing Education*, *3*, 1–10.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, *42*(2), 193–198. https://doi.org/10.1007/bf02294048

Sax, G., & Cromack, T. R. (1966). The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement*, *3*(4), 309–311. https://doi.org/10.1111/j.1745-3984.1966.tb00896.x

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50. https://doi.org/10.1016/j.intell.2014.10.003

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Routledge. https://doi.org/10.4324/9781410612250-20

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Scrams, D., & Smith, R. (2010). A practical approach to balancing time demands across test forms [Presented at the annual meeting of the National Council on Measurement in Education, Denver, CO].

Sehmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*(1), 67–81. https://doi.org/10.1111/j.1745-3984.1990.tb00735.x

Shin, H., Kerzabi, E., Joo, S.-H., Robin, F., & Yamamoto, K. (2020). Comparability of response time scales in PISA. *Psychological Test and Assessment Modeling*, *62*(1), 107–135.

Sinharay, S., & Johnson, M. S. (2019). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 397–419. https://doi.org/https://doi.org/10.1111/bmsp.12187

Sireci, S. G., & Botha, S. M. (2020). Timing considerations in test development and administration. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 32–46). Routledge. https://doi.org/10.4324/9781351064781-3

Smith, K. J., Davy, J. A., & Easterling, D. (2004). An examination of cheating and its antecedents among marketing and management majors. *Journal of Business Ethics*, *50*(1), 63–80. https://doi.org/10.1023/b:busi.0000020876.72462.3f

Spaccapanico Proietti, G., Matteucci, M., & Mignani, S. (2020). Automated test assembly for large-scale standardized assessments: Practical issues and possible solutions. *Psych*, *2*(4), 315–337. https://doi.org/10.3390/psych2040024

Spearman, C. (1927). *The abilities of man*. MacMillan.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stan Development Team. (2021). *rstan: The R interface to Stan* [R package version 2.21.3]. http://mc-stan.org/

Steedle, J. T., & Grochowalski, J. (2017). The effect of stakes on accountability test scores and pass rates. *Educational Assessment*, *22*(2), 111–123. https://doi.org/10.1080/10627197.2017.1309276

Steinmayr, R., & Spinath, B. (2019). Why time constraints increase the gender gap in measured numerical intelligence in academically high achieving samples. *European Journal of Psychological Assessment*, *35*(3), 392–402. https://doi.org/10.1027/1015-5759/a000400

Stoevenbelt, A. H., Wicherts, J. M., Flore, P. C., Phillips, L. A. T., Pietschnig, J., Verschuere, B., Voracek, M., & Schwabe, I. (2022). Are speeded tests unfair? Modeling the impact of time limits on the gender gap in mathematics. *Educational and Psychological Measurement*, Advance online publication. https://doi.org/10.1177/00131644221111076

Suh, Y., Cho, S.-J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement*, *49*(3), 285–311. https://doi.org/10.1111/j.1745-3984.2012.00176.x

Sun, K.-T., Chen, Y.-J., Tsai, S.-Y., & Cheng, C.-F. (2008). Creating IRT-based parallel test forms using the genetic algorithm method. *Applied Measurement in Education*, *21*(2), 141–161. https://doi.org/10.1080/08957340801926151

Swensson, R. G., & Edwards, W. (1971). Response strategies in a two-choice reaction task with a continuous cost for time. *Journal of Experimental Psychology*, *88*(1), 67–81. https://doi.org/10.1037/h0030646

Swineford, F. (1956). *Technical manual for users of test analyses* (ETS Statistical Report No. SR-56-42). Educational Testing Service.

Swineford, F. (1974). *The test analysis manual* (ETS Statistical Report No. SR-74-06). Educational Testing Service.

Theunissen, T. (1985). Binary programming and test design. *Psychometrika*, *50*(4), 411–420. https://doi.org/10.1007/bf02296260

Theussl, S., & Hornik, K. (2019). *Rglpk: R/GNU linear programming kit interface* [R package version 0.6-4]. https://CRAN.R-project.org/package=Rglpk

Thurstone, L. L. (1919). A scoring method for mental tests. *Psychological Bulletin*, *16*(7), 235–240. https://doi.org/10.1037/h0069898

Tiffin-Richards, S. P., & Pant, H. A. (2017). Arguing validity in educational assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Methodology of educational measurement and assessment* (pp. 469–485). Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0_27

Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology*, *9*, 1–14. https://doi.org/10.3389/fpsyg.2018.00964

Tijmstra, J., & Bolsinova, M. (2021). Accounting for individual differences in speed in the discretized signed residual time model. *British Journal of Mathematical and Statistical Psychology*, *74*(S1), 176–198. https://doi.org/10.1111/bmsp.12223

Togo, D. F. (2002). Topical sequencing of questions and advance organizers impacting on students' examination performance. *Accounting Education*, *11*(3), 203–216. https://doi.org/10.1080/0963928022000025480

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., von Davier, M., & Pohl, S. (2019a). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level nonresponse. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., von Davier, M., & Pohl, S. (2019b). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, *55*(3), 425–453. https://doi.org/10.1080/00273171.2019.1643699

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A multi-process item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*, *80*(3), 522–547. https://doi.org/10.1177/0013164419878241

United States Department of Justice. (2019). Affidavit in support of criminal complaint. U.S. attorney's office district of Massachusetts. https://www.justice.gov/file/1142876/download

van der Linden, W. J. (2005). *Linear models for optimal test assembly*. Springer. https://doi.org/10.1007/0-387-29054-0

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. https://doi.org/10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2009a). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*(1), 25–41. https://doi.org/10.1177/0146621607314042

van der Linden, W. J. (2011a). Setting time limits on tests. *Applied Psychological Measurement*, *35*(3), 183–199. https://doi.org/10.1177/0146621610391648

van der Linden, W. J. (2011b). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60. https://doi.org/10.1111/j.1745-3984.2010.00130.x

van der Linden, W. J. (2017). Test speededness and time limits. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 249–266). Chapman; Hall/CRC.

van der Linden, W. J. (2022). Two statistical tests for the detection of item compromise. *Journal of Educational and Behavioral Statistics*, *47*(4), 485–504. https://doi.org/10.3102/10769986221094789

van der Linden, W. J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subtests method. *Applied Psychological Measurement*, *12*(2), 201–209. https://doi.org/10.1177/014662168801200210

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Springer. https://doi.org/10.1007/0-306-47531-6

van der Linden, W. J., & Glas, C. A. W. (2010a). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

van der Linden, W. J., & Glas, C. A. W. (2010b). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139. https://doi.org/10.1007/s11336-009-9129-9

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347. https://doi.org/10.1177/0146621609349800

van der Linden, W. J., & Li, J. (2016). Comment on three-element item selection procedures for multiple forms assembly: An item matching approach. *Applied Psychological Measurement*, *40*(8), 641–649.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*(3), 259–270. https://doi.org/10.1177/01466216980223006

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 418–438. https://doi.org/10.3102/1076998612466143

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

Vander Schee, B. A. (2013). Test item order, level of difficulty, and student performance in marketing education. *Journal of Education for Business*, *88*(1), 36–42. https://doi.org/10.1080/08832323.2011.633581

van Rijn, P. W., & Ali, U. S. (2018). A generalized speed-accuracy response model for dichotomous items. *Psychometrika*, *83*(1), 109–131. https://doi.org/10.1007/s11336-017-9590-9

van Rijn, P. W., Attali, Y., & Ali, U. S. (2021). Impact of scoring instructions, timing, and feedback on measurement: An experimental study. *The Journal of Experimental Education*, 1–23. https://doi.org/10.1080/00220973.2021.1969532

van Rijn, P. W., & Sinharay, S. (2023). Modeling item response times. In *International encyclopedia of education(fourth edition)* (pp. 321–330). Elsevier. https://doi.org/10.1016/b978-0-12-818630-5.10040-5

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* [R package version 2.5.1]. https://mc-stan.org/loo

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved Rhat for assessing convergence of MCMC. *Bayesian Analysis*. https://doi.org/10.1214/20-BA1221

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, *36*(3), 253–266. https://doi.org/10.1111/j.1745-3984.1999.tb00557.x

Veldkamp, B. P. (2016). On the issue of item selection in computerized adaptive testing with response times. *Journal of Educational Measurement*, *53*(2), 212–228. https://doi.org/10.1111/jedm.12110

Veldkamp, B. P., Avetisyan, M., Weissmann, A., & Fox, J.-P. (2017). Stochastic programming for individualized test assembly with mixture response time models. *Computers in Human Behavior*, *76*, 693–702. https://doi.org/10.1016/j.chb.2017.04.060

Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, *37*(2), 123–139. https://doi.org/10.1177/0146621612469825

Verschoor, A. J. (2007). *Genetic algorithms for automated test assembly* [Doctoral dissertation, Twente University]. https://www.persistent-identifier.nl/urn:nbn:nl:ui:28-60710

Vista, A., & Alahmadi, M. T. (2022). Differences in discrimination with respect to latent trait and test-taking speed across items measuring sub-domains of cognitive ability. *Journal of Psychoeducational Assessment*, *40*(8), 1000–1016. https://doi.org/10.1177/07342829221118183

von Davier, M., & Mullis, I. (2022). TIMSS 2023: Progress towards a fully digital assessment [Presented at the IEA General Assembly in Split, Croatia]. https://www.iea.nl/sites/default/files/2022-10/GA63_TIMSS%202023.pdf

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis. *Psychonomic Bulletin & Review*, *18*(2), 267–277. https://doi.org/10.3758/s13423-010-0042-0

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501. https://doi.org/10.3102/1076998618767123

Wang, L. (2019). *Does rearranging multiple choice item response options affect item and test performance?* (ETS Research Report Series No. RR-19-02). Educational Testing Service. https://doi.org/10.1002/ets2.12238

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/bf02294627

Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58*(4), 671–701.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*(2), 115–129. https://doi.org/10.1177/0146621616676791

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence*, *30*(6), 537–554. https://doi.org/10.1016/s0160-2896(02)00086-7

Williams, I. (2017). *A speededness item response model for associating ability and speededness parameters* [Doctoral dissertation, Rutgers State University]. https://www.proquest.com/docview/2002597559

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., & Kingsbury, G. G. (2022). Performance decline as an indicator of generalized test-taking disengagement. *Applied Measurement in Education*, *35*(4), 272–286. https://doi.org/10.1080/08957347.2022.2155651

Wise, S. L., & Kuhfeld, M. (2021). A method for identifying partial test-taking engagement. *Applied Measurement in Education*, *34*(2), 150–161. https://doi.org/10.1080/08957347.2021.1890745

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242. https://doi.org/10.1207/s15324818ame0803_3

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, *40*(4), 307–330. https://doi.org/10.1111/j.1745-3984.2003.tb01149.x

Wright, B., & Masters, G. (1982). *Rating scale analysis*. MESA Press.

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, *7*, 1–21. https://doi.org/10.1186/s40536-019-0073-6

Wu, T., Guo, S., & Chang, H.-H. (2022). Modeling student's response time in an attribute balanced cognitive diagnostic adaptive testing. In M. Wiberg, D. Molenaar, J. Gonzalez, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology* (pp. 299–312). Springer Proceedings in Mathematics & Statistics (Vol. 393). https://doi.org/10.1007/978-3-031-04572-1_23

Yamamoto, K. (1995, June). *Estimating the effects of test length and test time on parameter estimation using the hybrid model* (TOEFL Technical Report No. TR-10). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01637.x

Yan, D., Von Davier, A. A., & Lewis, C. (2016). *Computerized multistage testing: Theory and applications*. CRC Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Springer. https://doi.org/10.1007/978-0-387-85461-8_18

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 262–286. https://doi.org/10.1111/bmsp.12114

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H.-H. (2016, April). Overview of test assembly methods in multistage testing. In D. Yan, A. A. Von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 125–138). CRC Press. https://doi.org/10.1201/b16858-16

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 646–661. https://doi.org/10.1080/10705511.2018.1545232

A Appendix to Chapter 2

A.1 Derivation of the (Model Implied) Correlation of Item Response Times

In the following, λ_k and ϕ_k are the time intensity parameter and the speed sensitivity parameter of item k, respectively, $\sigma^2_{\epsilon_k}$ is the residual variance parameter of item k and ζ_i is the speed parameter of person i. We assume that over persons, speed is normally distributed $\zeta \sim \mathcal{N}(0, \sigma^2_\zeta)$. Further, if X is a log normally distributed variable with parameters μ and σ, then $X = \exp(Y)$ where $Y \sim \mathcal{N}(\mu, \sigma^2)$, which is the same as $ln(X) \sim \mathcal{N}(\mu, \sigma^2)$. Then, from Johnson et al. (1994), the expectation of X is:

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}[\exp(Y)] \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right). \end{aligned} \tag{65}$$

In addition, the expectation of X^2 is:

$$\begin{aligned} \mathrm{E}(X^2) &= \mathrm{E}[(\exp(Y))^2] \\ &= \mathrm{E}[\exp(2Y)] \\ &= \mathrm{E}[\exp(Y^*)] \\ &= \exp\left(2\mu + 2\sigma^2\right), \end{aligned} \tag{66}$$

where $Y^* = 2Y$ and hence $Y^* \sim \mathcal{N}(2\mu, 4\sigma^2)$.

Based on the 3PLN lognormal measurement model for response times RT_{ik}, the expectation of the response time variable $\mathrm{E}(RT_k)$ of an item k can be written as

$$\begin{aligned} \mathrm{E}(RT_k) &= \mathrm{E}\left[\mathrm{E}(RT_{ik}|\zeta_i)\right] \\ &= \mathrm{E}\left[\exp\left(\lambda_k - \phi_k\zeta_i + \frac{\sigma^2_{\epsilon_k}}{2}\right)\right] \\ &= \mathrm{E}\left[\exp(Z)\right] \\ &= \exp\left(\lambda_k + \frac{\sigma^2_{\epsilon_k}}{2} + \frac{\phi^2_k\sigma^2_\zeta}{2}\right), \end{aligned} \tag{67}$$

where Z is a normally distributed: $Z \sim \mathcal{N}(\lambda_k + \frac{\sigma^2_{\epsilon_k}}{2}, \phi^2_k\sigma^2_\zeta)$.

Using similar steps in the derivation, the expectation of the squared response time variable $\mathrm{E}(RT_k^2)$ can be written as

$$\begin{aligned}\mathrm{E}[RT_k^2] &= \mathrm{E}\left[\mathrm{E}(RT_{ik}^2|\zeta_i)\right] \\ &= \mathrm{E}\left[\exp\left(2\lambda_k - 2\phi_k\zeta_i + 2\sigma_{\epsilon_k}^2\right)\right] \\ &= \exp\left(2\lambda_k + 2\sigma_{\epsilon_k}^2 + 2\phi_k^2\sigma_\zeta^2\right).\end{aligned} \tag{68}$$

Therefore, the variance of the response time variable of item k, $\mathrm{Var}(RT_k)$, can be denoted as

$$\begin{aligned}\mathrm{Var}(RT_k) &= \mathrm{E}(RT_k^2) - \mathrm{E}(RT_k)^2 \\ &= \exp\left(2\lambda_k + 2\sigma_{\epsilon_k}^2 + 2\phi_k^2\sigma_\zeta^2\right) - \exp\left(2\lambda_k + \sigma_{\epsilon_k}^2 + \phi_k^2\sigma_\zeta^2\right) \\ &= \exp\left(2\lambda_k\right)\exp\left(\sigma_{\epsilon_k}^2 + \phi_k^2\right)\exp\left(\sigma_{\epsilon_k}^2 + \phi_k^2\right) - \exp\left(2\lambda_k\right)\exp\left(\sigma_{\epsilon_k}^2 + \phi_k^2\right) \\ &= \exp\left(2\lambda_k + \sigma_{\epsilon_k}^2 + \phi_k^2\sigma_\zeta^2\right)\left(\exp(\sigma_{\epsilon_k}^2 + \phi_k^2\sigma_\zeta^2) - 1\right).\end{aligned} \tag{69}$$

For the expectation of the distribution of the product of the response times of items k and l, $\mathrm{E}(RT_kRT_l)$, this gives

$$\begin{aligned}\mathrm{E}(RT_kRT_l) &= \mathrm{E}\left[\exp\left(N(\lambda_k + \lambda_l - (\phi_k + \phi_l)\zeta, \sigma_{\epsilon_k}^2 + \sigma_{\epsilon_l}^2)\right)\right] \\ &= \exp\left(\lambda_k + \lambda_l + \frac{\sigma_{\epsilon_k}^2 + \sigma_{\epsilon_l}^2}{2} + \frac{(\phi_k + \phi_l)^2\sigma_\zeta^2}{2}\right).\end{aligned} \tag{70}$$

Therefore, the covariance of the response time variables of two items k and l, $\mathrm{Cov}(RT_k, RT_l)$ and the respective product of the variances of these two items can be denoted as:

$$\begin{aligned}&\mathrm{Cov}(RT_k, RT_l) = \mathrm{E}(RT_kRT_l) - \mathrm{E}(RT_k)\mathrm{E}(RT_k) \\ &= \exp\left(\lambda_k + \lambda_l + \frac{\sigma_{\epsilon_k}^2 + \sigma_{\epsilon_l}^2}{2} + \frac{(\phi_k + \phi_l)^2\sigma_\zeta^2}{2}\right) - \exp\left(\lambda_k + \lambda_l + \frac{\sigma_{\epsilon_k}^2 + \sigma_{\epsilon_l}^2}{2} + \frac{(\phi_k^2 + \phi_l^2)\sigma_\zeta^2}{2}\right) \\ &= \exp\left(\lambda_k + \lambda_l + \frac{\sigma_{\epsilon_k}^2 + \sigma_{\epsilon_l}^2}{2} + \frac{(\phi_k^2 + \phi_l^2)\sigma_\zeta^2}{2}\right)\left[\exp\left(\phi_k\phi_l\sigma_\zeta^2\right) - 1\right]\end{aligned} \tag{71}$$

$$
\begin{aligned}
&\mathrm{Var}(RT_k)\mathrm{Var}(RT_l) \\
&\quad = \exp\left(2\lambda_k + \sigma^2_{\epsilon_k} + \phi^2_k\sigma^2_\zeta\right)\left(\exp(\sigma^2_{\epsilon_k} + \phi^2_k\sigma^2_\zeta) - 1\right) \\
&\qquad \exp\left(2\lambda_l + \sigma^2_{\epsilon_l} + \phi^2_l\sigma^2_\zeta\right)\left(\exp(\sigma^2_{\epsilon_l} + \phi^2_l\sigma^2_\zeta) - 1\right) \\
&\quad = \exp\left(2\lambda_k + \sigma^2_{\epsilon_k} + \phi^2_k\sigma^2_\zeta\right)\exp\left(2\lambda_l + \sigma^2_{\epsilon_l} + \phi^2_l\sigma^2_\zeta\right) \\
&\qquad \left(\exp(\sigma^2_{\epsilon_k} + \phi^2_k\sigma^2_\zeta) - 1\right)\left(\exp(\sigma^2_{\epsilon_l} + \phi^2_l\sigma^2_\zeta) - 1\right) \\
&\quad = \exp\left(2\lambda_k + 2\lambda_l + \sigma^2_{\epsilon_k} + \sigma^2_{\epsilon_l} + \phi^2_k\sigma^2_\zeta + \phi^2_l\sigma^2_\zeta\right) \\
&\qquad \left(\exp(\sigma^2_{\epsilon_k} + \phi^2_k\sigma^2_\zeta) - 1\right)\left(\exp(\sigma^2_{\epsilon_l} + \phi^2_l\sigma^2_\zeta) - 1\right)
\end{aligned} \tag{72}
$$

This gives

$$
\begin{aligned}
\rho_{RT_k,RT_l} &= \frac{\mathrm{Cov}(RT_k, RT_l)}{\sqrt{\mathrm{Var}(RT_k)\mathrm{Var}(RT_l)}} \\
&= \frac{\left[\exp\left(\phi_k\phi_l\sigma^2_\zeta\right) - 1\right]}{\sqrt{\left(\exp(\sigma^2_{\epsilon_k} + \phi^2_k\sigma^2_\zeta) - 1\right)\left(\exp(\sigma^2_{\epsilon_l} + \phi^2_l\sigma^2_\zeta) - 1\right)}}.
\end{aligned} \tag{73}
$$

This is the model implied correlation of the two response time distributions of items k and i under the 3PLN model. Under the 2PLN model $phi_k = phi_l = 1$, therefore the respective correlation is

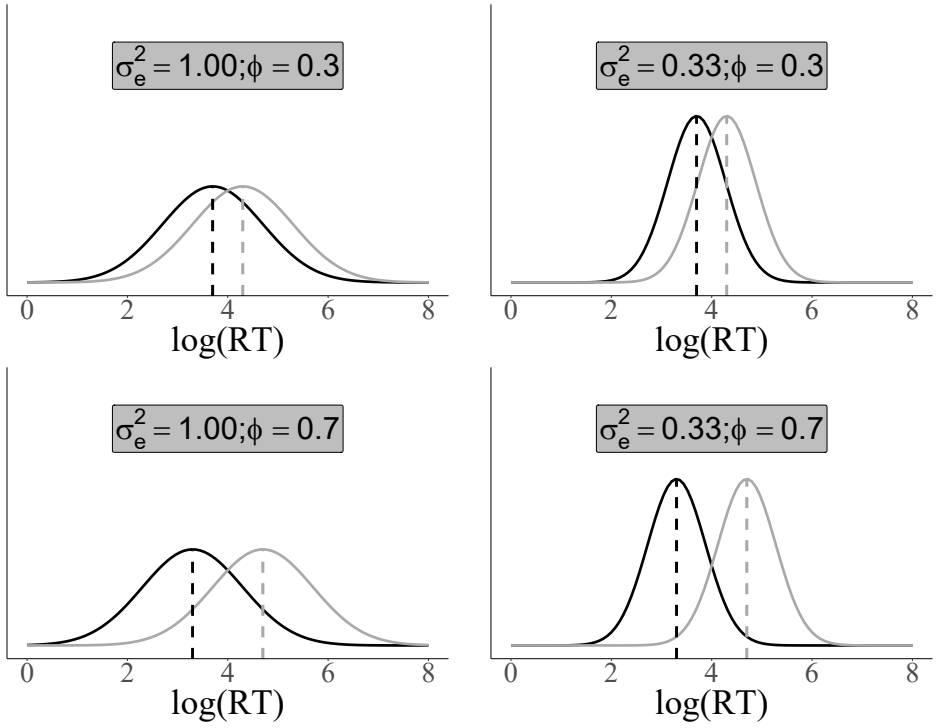
$$
\frac{\left[\exp\left(\sigma^2_\zeta\right) - 1\right]}{\sqrt{\left(\exp(\sigma^2_{\epsilon_k} + \sigma^2_\zeta) - 1\right)\left(\exp(\sigma^2_{\epsilon_l} + \sigma^2_\zeta) - 1\right)}}. \tag{74}
$$

References

Johnson Norman, L., Kotz, S., & Balakrishnan, N. (1994). Lognormal distributions. In *Continuous univariate distributions (Vol. 1, Ed. 2)*. Wiley.

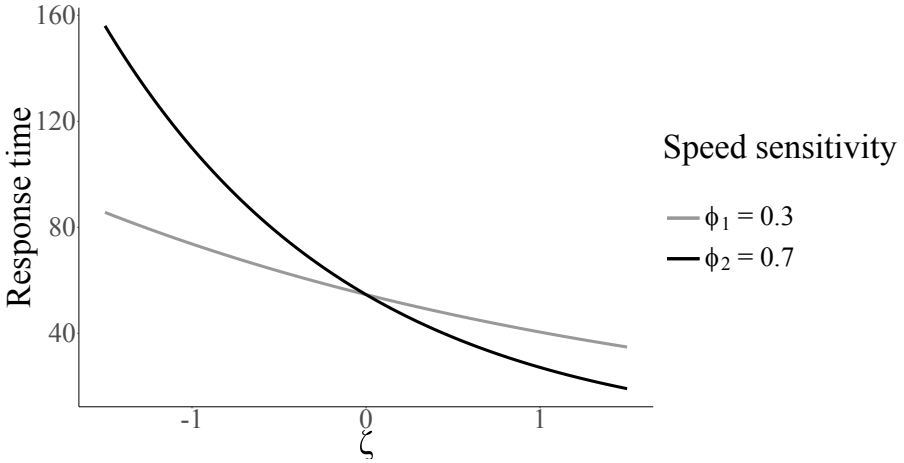
A.2 Item Log Response Time Distributions

Figure 35: *Expected Log Response Time Distributions of a Fast Person with $\zeta_1 = 1$ (Black Line) and a Slow Person with $\zeta_2 = -1$ (Grey Line) on Four Different Items, all with $\lambda_k = 4$. Dashed Lines Indicate the Medians of the Corresponding Distributions.*



A.3 Response Time Characteristic Curve

Figure 36: *Response Time Characteristic Curve of Two Items with Identical Time Intensity ($\lambda_k = 4$) and Differing Item Speed Sensitivity Parameters ϕ_k.*



A.4 Priors for Empirical Data Analysis

The identity matrix is notated as I_n with the size of n. $\sigma_{\theta_i,\zeta_i}$ is truncated to stay in range of $-\sqrt{\sigma_\theta^2\sigma_\zeta^2}$ and $\sqrt{\sigma_\theta^2\sigma_\zeta^2}$ (with $\sigma_\theta^2 = 1$ for model identification) to keep the person parameter covariance matrix positive definite. Priors for the hierarchical framework with a 2PL model for ability and a 2PLN model for speed:

$$\begin{aligned}
\Sigma_P &\sim InverseWishart(I_3, 4) \\
\sigma_{\theta_i,\zeta_i} &\sim N(0, 10000) \text{ truncated at } [-\sigma_\zeta, \sigma_\zeta] \\
\frac{1}{\sigma_\zeta^2} &\sim \Gamma(0.01, 0.01) \\
\frac{1}{\sigma_{\sigma_\epsilon}^2} &\sim \Gamma(0.01, 0.001) \\
\mu_{\sigma_\epsilon} &\sim N(0, 1000000) \\
\mu_b &\sim N(0, 1000000) \\
\mu_a &\sim N(1, 1000000) \\
\mu_\lambda &\sim N(1, 1000000)
\end{aligned} \tag{75}$$

Priors for the hierarchical framework with a 2PL model for ability and a 3PLN model for speed:

$$\begin{aligned}
\Sigma_P &\sim InverseWishart(I_4, 5) \\
\sigma_{\theta_i,\zeta_i} &\sim N(0, 10000) \\
\frac{1}{\sigma_{\sigma_\epsilon}^2} &\sim \Gamma(0.01, 0.001) \\
\mu_{\sigma_\epsilon} &\sim N(0, 1000000) \\
\mu_b &\sim N(0, 1000000) \\
\mu_a &\sim N(1, 1000000) \\
\mu_\lambda &\sim N(1, 1000000) \\
\mu_\phi &\sim N(1, 1000000)
\end{aligned} \tag{76}$$

The first model was identified by fixing the hyperpriors of the person ability and person speed distributions: The means of the person parameter distributions were fixed to 0 ($M_{\theta_i} = 0$ and $M_{\zeta_i} = 0$) and the variance of the ability was fixed to 1 ($Var_{\theta_i} = 1$). For the 3PLN model the variance of the person speed was also fixed to 1 ($Var_{\zeta_i} = 1$). All item parameters were estimated freely.

A.5 Empirical Model Fit

Table 14: *DIC for the Hierarchical Framework with the 2PLN Model and the 3PLN Model and the Corresponding Difference for all Math Booklets.*

Booklet	$DIC(3PLN)$	$DIC(2PLN)$	Δ_{DIC}
M01	251955	253042	1087
M02	213884	215179	1295
M03	231336	231690	354
M04	256032	256617	585
M05	257370	257703	333
M06ab	267682	268551	869

A.6 Multivariate Normal Distributions for Data Generation

Means of the multivariate normal distribution:

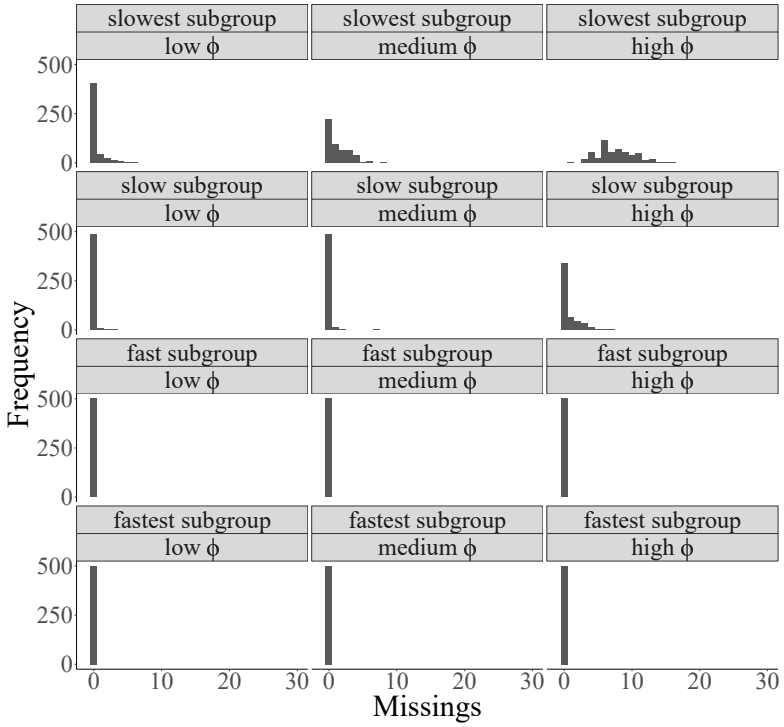
$$\mu_I = (\mu_a = 1.12, \mu_b = 0.54, \mu_\phi = 0.3, \mu_\lambda = 4.26) \tag{77}$$

Covariances of the multivariate normal distribution:

$$\Sigma_I = \begin{pmatrix} \sigma_a^2 = 0.45 & & & \\ \sigma_{b,a} = 0.05 & \sigma_b^2 = 1.00 & & \\ \sigma_{\phi,a} = 0.01 & \sigma_{\phi,b} = 0.03 & \sigma_\phi^2 = 0.01 & \\ \sigma_{\lambda,a} = -0.02 & \sigma_{\lambda,b} = 0.13 & \sigma_{\lambda,\phi} = 0.01 & \sigma_\lambda^2 = 0.25 \end{pmatrix} \tag{78}$$

A.7 Item Numbers Not Reached in Simulation

Figure 37: *Number of Not-Reached Items for the Low, Medium and High Speed Sensitivity Test Form, Across the Four Subgroups. Results Shown for a Randomly Selected Single Replication.*



A.8 Standard Deviations for Simulation Results Across Replications

Table 15: *Standard Deviations for Test Statistics per Test Form and per Speed Group, Across All Replications.*

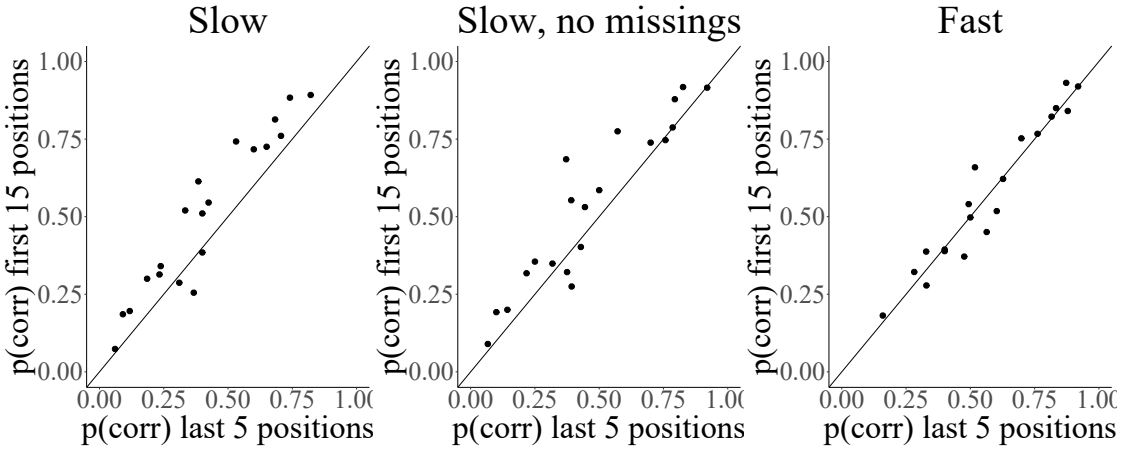
Test Form	ζ_i	$M(RT)$	$SD(RT)$	$M(mis)$	$SD(mis)$	$cor(\hat{\theta}, \theta)$	RMSE	$M(\theta_{diff})$
low ϕ	slowest	42.06	31.42	0.01	0.01	0.02	0.05	0.03
low ϕ	slow	33.39	24.36	0.00	0.01	0.01	0.05	0.02
low ϕ	fast	21.51	17.30	0.00	0.00	0.02	0.06	0.02
low ϕ	fastest	18.04	14.39	0.00	0.00	0.02	0.06	0.02
medium ϕ	slowest	46.88	33.10	0.02	0.02	0.02	0.07	0.05
medium ϕ	slow	33.97	25.37	0.00	0.01	0.02	0.06	0.02
medium ϕ	fast	19.86	16.64	0.00	0.00	0.01	0.05	0.02
medium ϕ	fastest	16.25	13.48	0.00	0.00	0.02	0.06	0.02
high ϕ	slowest	63.29	47.33	0.04	0.02	0.04	0.15	0.14
high ϕ	slow	41.06	29.78	0.01	0.01	0.02	0.06	0.03
high ϕ	fast	17.54	13.72	0.00	0.00	0.01	0.05	0.02
high ϕ	fastest	12.45	9.82	0.00	0.00	0.02	0.06	0.02

Note: Standard deviations across replications are depicted for mean cumulative response times M(RT) and the corresponding standard deviation SD(RT), mean proportion of missings M(mis), the corresponding standard deviation SD(mis), correlation between true and estimated ability $cor(\theta, \theta)$, root mean square error (RMSE) and average difference between true and estimated ability $M(\Delta_\theta)$.

B Appendix to Chapter 4

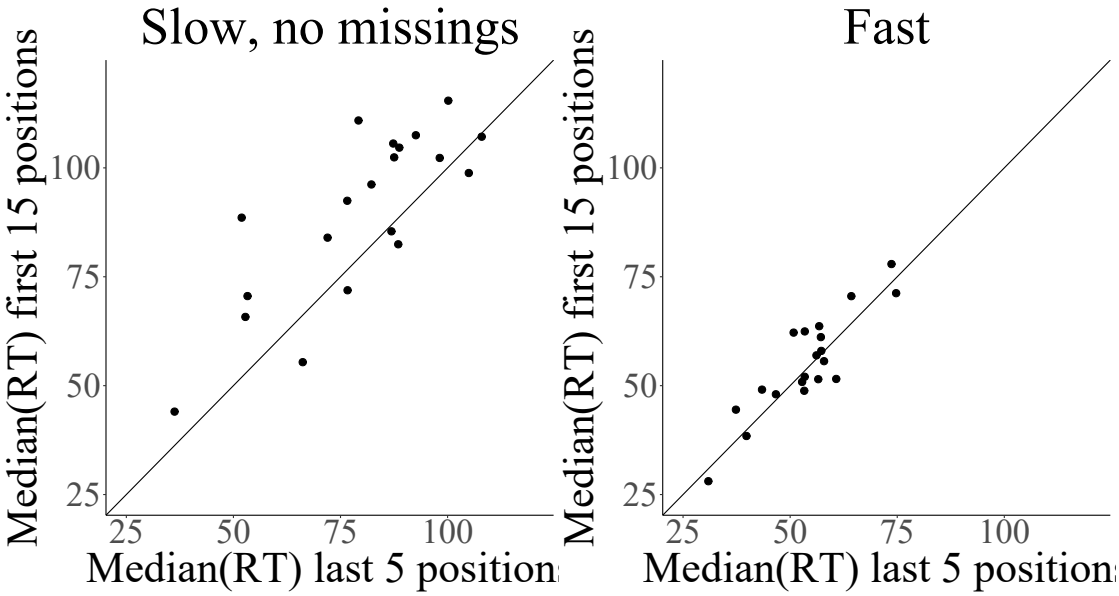
B.1 Speededness Analyses

Figure 38: *Percentage of Correct Answers on Items When the Item Occurred on Position 1-15 and 16-20.*



Note. Results displayed for all test-takers with overall test times greater than 30 minutes ("slow"), for all test-takers with overall test times greater than 30 minutes and no missings ("slow, no missings") and for test takers with overall test times less than 30 minutes ("fast").

Figure 39: *Median Response Times on Items When the Item Occurred on Position 1-15 and 16-20.*



Note. Results displayed for all test-takers with overall test times greater than 30 minutes and no missings ("slow, no missings") and for test takers with overall test times less than 30 minutes ("fast").

B.2 Illustrative Data Simulation

Table 16: *Summary Statistics of Simulated Test Scores for 100 Replications for Seven Different Test-Takers with Different Speed (ζ) and Ability Levels (θ).*

ζ	θ	M(range)	Max(range)
-0.78	-0.87	3.35	8
-0.76	-0.83	3.30	7
-0.59	-0.02	2.88	7
-0.83	0.10	3.89	9
-0.64	-0.74	2.87	7
-0.59	-0.95	2.85	10
-0.58	-0.57	3.01	6

Note: Mean range ($M(range)$) and maximum range ($Max(range)$) of differences between test forms.

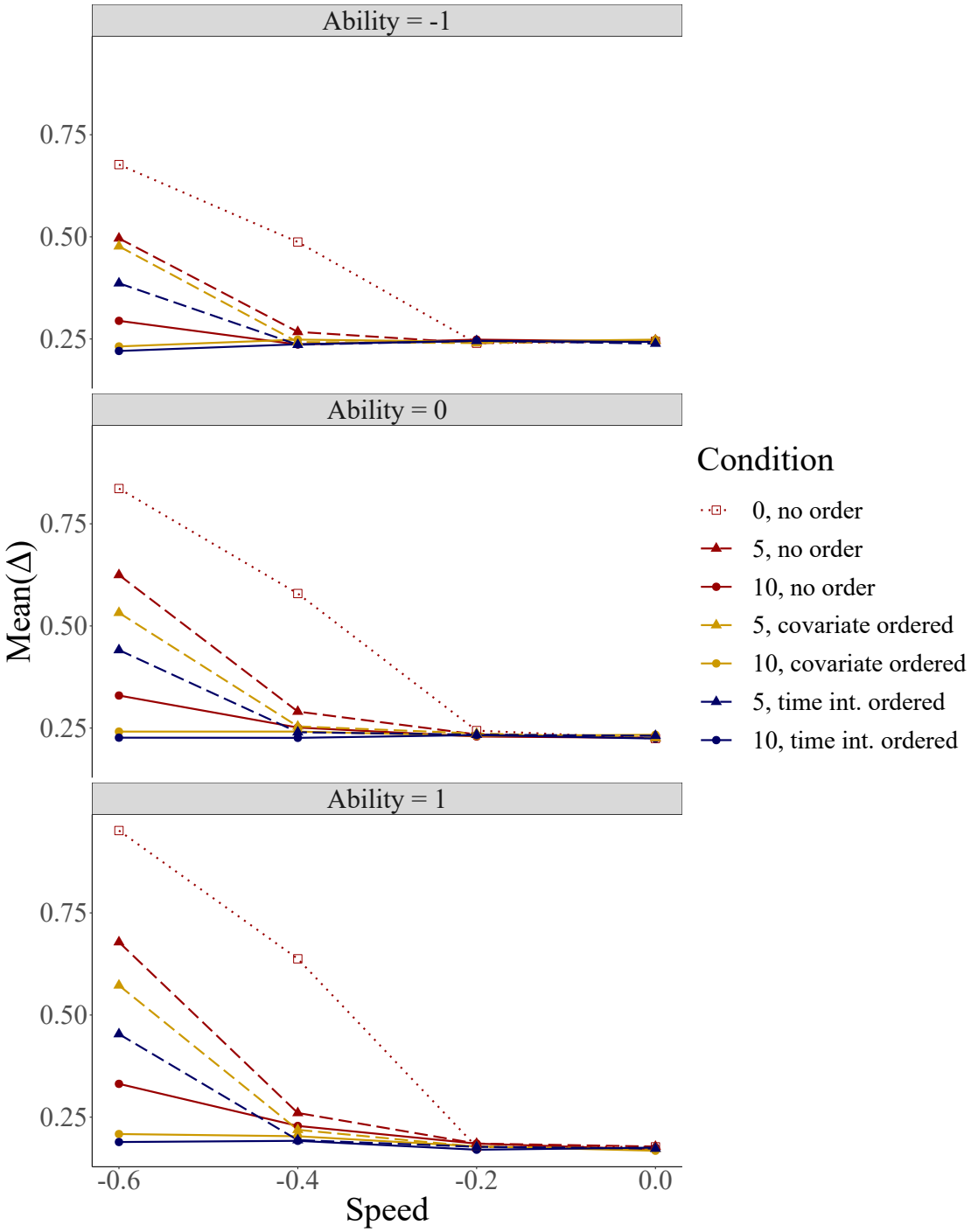
B.3 Simulation Study Results

Table 17: *Mean and Standard Deviation of Item and Person Parameters in the Hierarchical Estimated Model Using Organizational Psychology Exam Data.*

Parameter	M	SD
b	-0.78	0.45
λ	3.84	0.34
θ	0.00	0.33
ζ	0.00	0.22

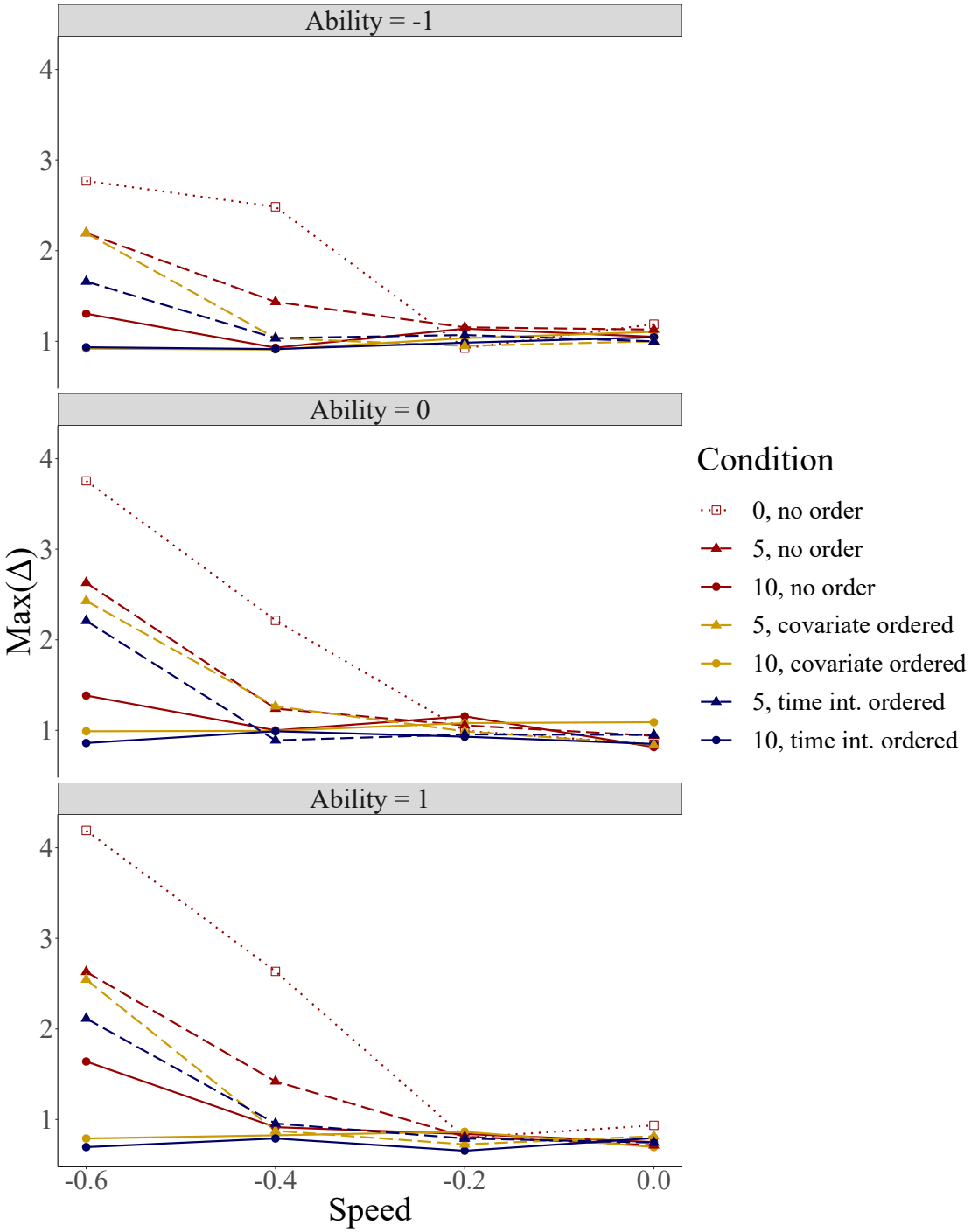
Note: Item Difficulty (b), Item Time Intensity (λ), Person Ability (θ), Person Speed (ζ).

Figure 40: *Mean Difference in Ability Estimation Between Test Forms.*



Note. Test forms share the exact same items but with either random orderings ("0, no order") or five or ten items were fixed at the end of the test forms, either with random ordering ("no order"), ordered by a covariate of time intensity ("covariate ordered") or ordered by time intensity ("time int. ordered").

Figure 41: *Maximum Difference in Ability Estimation Between Test Forms.*



Note. Test forms share the exact same items but with either random orderings ("0, no order") or five or ten items were fixed at the end of the test forms, either with random ordering ("no order"), ordered by a covariate of time intensity ("covariate ordered") or ordered by time intensity ("time int. ordered").

C Appendix to Chapter 5

C.1 Likelihood Functions

Writing down models' likelihoods can facilitate their translation into Stan code. We, therefore, here provide likelihood functions for the models discussed in the tutorial.

C.1.1 The Hierarchical Framework by van der Linden (2007)

The likelihood function of van der Linden's (2007) hierarchical framework can be written as

$$\mathcal{L} = \prod_{i=1}^{I}\prod_{k=1}^{K} p(y_{ik}|a_k, b_k, \theta_i)p(RT_{ik}|\lambda_k, \phi_k, \zeta_i, \sigma_{\epsilon_k}) \times p(\theta_i, \zeta_i|\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I})p(\ln a_k, b_k, \ln \phi_k, \lambda_k|\boldsymbol{\mu_K}, \boldsymbol{\Sigma_K}). \quad (79)$$

Here, the first two components give the probabilities of the observed responses and response times under the response and response time model, respectively, incorporating the assumption of conditional independence among observed indicators. The second two components give the probability of person, respectively item parameters given the respective multivariate normal distributions. Each of the components can be translated into a log probability increment statement `target+=` in Stan. For instance, the statement

```
target += bernoulli_logit_lpmf(y | discrimination[kk] .*
(ability[ii] - difficulty[kk]));
```

mirrors the likelihood's first component and adds to the log density the log probability of observing the given response vector given the response model. The likelihood functions of the discussed model extensions can easily be obtained from Equation 1.

C.1.2 Modeling the Difficulty-Distance Hypothesis for Non-Cognitive Data as in Ferrando and Lorenzo-Seva (2007)

The likelihood function for the model extension incorporating the distance-difficulty hypothesis when modeling data from non-cognitive assessments by Ferrando and Lorenzo-Seva (2007) considers examinee ability and item difficulty in the component model for response times. The speed sensitivity parameter is dropped.

$$\mathcal{L} = \prod_{i=1}^{I}\prod_{k=1}^{K} p(y_{ik}|a_k, b_k, \theta_i)p(RT_{ik}|\lambda_k, \zeta_i, \beta, a_k, b_k, \theta_i, \sigma_{\epsilon_k}) \\ \times p(\theta_i, \zeta_i|\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I})p(\ln a_k, b_k, \lambda_k|\boldsymbol{\mu_K}, \boldsymbol{\Sigma_K}) \quad (80)$$

C.1.3 Modeling Conditional Dependence of Response Times and Accuracy as in Bolsinova et al. (2017)

The likelihood function for the model extension accommodating conditional dependence by Bolsinova et al. (2017) considers observed response times, examinee speed, item time intensity, and the response time residual variance in the component model for item responses. The speed sensitivity parameter is dropped.

$$\mathcal{L} = \prod_{i=1}^{I}\prod_{k=1}^{K} p(y_{ik}|a_{0k}, a_{1k}, b_{0k}, b_{1k}, \theta_i, RT_{ik}, \lambda_k, \zeta_i, \sigma_{\epsilon_k})p(RT_{ik}|\lambda_k, \zeta_i, \sigma_{\epsilon_k}) \\ \times p(\theta_i, \zeta_i|\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I})p(\ln a_k, b_k, \lambda_k|\boldsymbol{\mu_K}, \boldsymbol{\Sigma_K}) \quad (81)$$

C.1.4 Modeling Qualitative Differences in Response Behavior as in Ulitzsch et al. (2020)

The mixture model accommodating rapid guessing behavior by Ulitzsch et al. (2020) assumes the response and response time models of van der Linden's (2007) hierarchical framework to hold when examinees are engaged, while responses and response times stemming from rapid guessing behavior are assumed to be unreflective of examinees' ability and speed. The likelihood function is that of a mixture model, i.e.,

$$\mathcal{L} = \prod_{i=1}^{I} \prod_{k=1}^{K} (p(\Delta_{ik} = 1|\psi_i, \iota_k)p(y_{ik}|a_k, b_k, \theta_i)p(RT_{ik}|\mu_d, \lambda_k^*, \phi_k, \zeta_i, \sigma_{\epsilon_k}) +$$
$$(1 - p(\Delta_{ik} = 1|\psi_i, \iota_k))p(y_{ik}|c)p(RT_{ik}|\mu_d, \sigma_d))$$
$$\times p(\theta_i, \zeta_i, \psi_i|\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I})p(\ln a_k, b_k, \ln \phi_k, \lambda_k, \iota_k|\boldsymbol{\mu_K}, \boldsymbol{\Sigma_K}), \quad (82)$$

with the first row representing the component model for engaged and the second the component model for rapid guessing behavior. Engagement and rapid guessing probabilities $p(\Delta_{ik} = 1)$ and $1 - p(\Delta_{ik} = 1)$ give the mixing proportions of this model.

C.2 Resources on Bayesian Modelling with Stan

C.2.1 General Introduction to Bayesian Modeling (Textbooks)

Congdon, P.D. (2021). *Bayesian Hierarchical Models: With Applications Using R (2nd ed.)*. CRC Press.

Donovan, T., & Mickey, R. (2019). *Bayesian Statistics for Beginners*. Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, Aki, & Rubin, D. B. (2014). *Bayesian data analysis (3rd ed.)*. CRC Press.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd ed.)*. Academic Press/Elsevier.

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. Sage.

McElreath, R. (2020). *Statistical Rethinking – A Bayesian Course with Examples in R and Stan (2nd ed.)*. CRC Press.

C.2.2 General Introduction to Hamiltonian MCMC

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*. https://doi.org/10.48550/arXiv.1701.02434

C.2.3 Resources for Stan, rstan, PPC and Model Evaluation

An overview over rstan, Stan and the respective documentation can be found at https://mc-stan.org, which includes the current Stan User's Guide (https://mc-stan.org/docs/stan-users-guide/index.html) and Stan Language Reference Manual (https://mc-stan.org/docs/reference-manual/index.html). Stan User's Guide also includes information about posterior predictive checks (https://mc-stan.org/docs/stan-users-guide/ppcs.html).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Gabry, J., & Mahr, T. (2017). *bayesplot: Plotting for Bayesian models* [R package version 1.9.0]. http://mc-stan.org/

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society, Series A*, 182, 389–402.

https://doi.org/10.1111/rssa.12378

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics, 40(5)*, 530–543. https://doi.org/10.3102/1076998615606113

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists, *The Quantitative Methods for Psychology, 12(3)*, 175–200. https://doi.org/10.20982/tqmp.12.3.p175

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27 (5)*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* [R package version 2.5.1]. https://mc-stan.org/loo

C.2.4 IRT Modeling with Stan

Ames, A.J., & Au, C.H. (2018). Using Stan for Item Response Theory Models. *Measurement: Interdisciplinary Research and Perspectives, 16*, 129–134. https://doi.org/10.1080/15366367.2018.1437304

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software, 100*, 1–54. https://doi.org/10.18637/jss.v100.i05

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, 78(3), 384–408. https://doi.org/10.1177/00131644176936

D Appendix to Chapter 6

D.1 Constraint Formulation for the Minimal Example

In the minimal example, the combination of constraints and objective results in an MILP model that can be mathematically formulated as follows. Let the items in the item pool have a unique index $k = 1, \ldots, j$, in this example $j = 30$. Let F be the total number of test forms to be assembled, here $F = 1$. The MILP model has two parts: (1) the objective function,

$$max\ \mathbf{c}^T\mathbf{x}, \tag{83}$$

and (2) a set of constraints,

$$\mathbf{Ax} \leq \mathbf{d}, \tag{84}$$

where $\mathbf{x}$ is the vector of variables that MILP needs to solve for. $\mathbf{x}$ contains binary decision variables x_{kf}, $k = 1, 2, \ldots, j$, and $f = 1, 2, \ldots, F$ for every item $\times$ test from combination, and one real-valued variable z. Hence, $\mathbf{x}$ is a vector of length $j \times F + 1$. The binary decision variables are defined as:

$$x_{kf} = \begin{cases} 1 & \text{if item } k \text{ is assigned to form } f, \\ 0 & \text{otherwise.} \end{cases} \tag{85}$$

Further, in the objective function (Equation (83)), $\mathbf{c}$ is a numeric vector of $j \times F + 1$ known coefficients for the objective function. In the set of constraints (Equation (84)) $\mathbf{A}$ is a known coefficient matrix with $j \times F + 1$ columns and with one row for each constraint and $\mathbf{d}$ is a vector with the corresponding right-hand values of the constraints.

The code in Figure 27 simultaneously creates coefficients for $\mathbf{c}$ as well as coefficients for one row in $\mathbf{A}$ and the corresponding value in $\mathbf{d}$. More specifically, the coefficients in $\mathbf{c}$ for all binary decision variables are set to zero, whereas the coefficient for the real-valued variable z is set to one 1. Thus, Equation (83) simplifies to:

$$max\ z. \tag{86}$$

In addition, the following constraint is added as one row in $\mathbf{A}$ and the corresponding value in $\mathbf{d}$:

$$\sum_{k=1}^{j} s_k \times x_{kf} - z \geq 0, \quad \text{for } f = 1, \ldots, F. \tag{87}$$

In Equation (87) s_k denotes the IIF value of item k at a medium ability level. Hence, the coefficients in $\mathbf{A}$ for the binary decision variables x_{kf} are set to s_k, whereas the coefficient for z is set to 1. Finally, the value in $\mathbf{d}$ corresponding to the row in $\mathbf{A}$ is set to 0. The combination of Equations (86) and (87) makes sure that the TIF of the test forms is maximized.

In addition, the following constraints are also enforced:

$$\sum_{k=1}^{j} x_{kf} = 10, \quad \text{for } f = 1, \ldots, F, \tag{88}$$

and

$$\begin{aligned} &\sum_{k=1}^{j} t_k \times x_{kf} \leq (8 \times 60) + 5, \quad \text{and} \\ &\sum_{k=1}^{j} t_k \times x_{kf} \geq (8 \times 60) - 5, \quad \text{for } f = 1, \ldots, F. \end{aligned} \tag{89}$$

In Equation (89) t_k denotes the average response time for item k in seconds. Hence, Equation (88) constrains the length of the test forms to be equal to ten and Equation (89) constrains the sum of the expected response times to be within five seconds of eight minutes. The left-hand sides and the right-hand sides of Equations (87) to (89) again correspond to the rows in $\mathbf{A}$ and $\mathbf{d}$, respectively. Note that the coefficient for z in these rows of $\mathbf{A}$ are set to zero.

D.2 Pilot Study Item Pool Illustration

Table 18: *First Five Items of the Simulated Pilot Study Item Pool.*

Item	diffCategory	Format	Domain	Time	Exclusions
1	2	cmc	listening	44.54	
2	4	cmc	listening	44.81	
3	4	mc	writing	32.36	76
4	2	mc	listening	48.03	
5	2	mc	writing	42.06	9

D.3 Large-Scale Assessment Item Pool Illustration

Table 19: *First 10 Items of the Simulated LSA Assessment Item Pool.*

Testlet	Item	Level	Format	Frequency	Infit	Time	Anchor
TRA5308	TRA5308a	IV	multiple choice	0.19	1.22	54.00	0
TRA5308	TRA5308b	IV	multiple choice	0.24	1.01	66.00	0
TRA5308	TRA5308c	II	multiple choice	0.42	1.22	89.00	0
TRA5308	TRA5308d	III	multiple choice	0.41	1.21	92.00	0
TRB6832	TRB6832a	III	open answer	0.51	1.21	85.00	0
TRB6832	TRB6832b	III	open answer	0.20	1.08	61.00	0
TRB6832	TRB6832c	IV	open answer	0.33	1.25	84.00	0
TRB6832	TRB6832d	II	open answer	0.49	1.05	109.00	0
TRC9792	TRC9792a	I	cmc	0.70	1.10	94.00	0
TRC9792	TRC9792b	I	cmc	0.61	1.02	110.00	0

D.4 High-Stakes Assessment Item Pool Illustration

Table 20: *First Five Items of the Simulated High-Stakes Assessment Item Pool.*

Item	a	b	c	Category
1	0.54	−0.09	0.17	6
2	0.71	−1.07	0.24	1
3	0.84	−1.11	0.17	2
4	1.38	−0.71	0.21	3
5	1.26	−0.44	0.12	4

D.5 Item Category Distribution in the High-Stakes Assessment Item Pool

Table 21: *Item Category Distribution in the Item Pool and Test Specification.*

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6
Item Pool	23	26	22	29	29	36
HST	9	9	7	9	9	11
MST: Stage 1	4	4	3	4	4	5
MST: Stage 2	3	3	2	3	3	4

D.6 R Code for Multi-Stage Module Assembly

Figure 42: *Test Assembly Constraints for Multi-Stage Test Stage 1.*

```
maximinTIF1 <- combineConstraints(lapply(1:3, function(index)
  {
    maximinObjective(
      nForms = 1,
      itemValues = IIFs[,index],
      allowedDeviation = 0.5,
      itemIDs = items_diao2$item
    )
  })
)

contentConstraints1 <- itemCategoryConstraint(
  nForms = 1,
  itemCategories = items_diao2$category,
  operator = ">=",
  targetValues = c(4, 4, 3, 4, 4, 5),
  itemIDs = items_diao2$item)

noOverlap1 <- itemUsageConstraint(
  nForms = 1,
  itemIDs = items_diao2$item)

testLength1 <- itemsPerFormConstraint(
  nForms = 1,
  operator = "=",
  targetValue = 30,
  itemIDs = items_diao2$item)
```

Figure 43: *Test Assembly Constraints for Multi-Stage Test Stage 2.*

```
maximinTIF2 <- combineConstraints(lapply(1:3, function(index)
  {
    maximinObjective(
      nForms = 3,
      itemValues = IIFs_stage2[, index],
      allowedDeviation = 0.2,
      whichForms = index,
      itemIDs = items_diao2_stage2$item)
  })
)

contentConstraints2 <- itemCategoryConstraint(
  nForms = 3,
  itemCategories = items_diao2_stage2$Category,
  operator = ">=",
  targetValues = c(3, 3, 2, 3, 3, 4),
  itemIDs = items_diao2_stage2$item)

noOverlap2 <- itemUsageConstraint(
  nForms = 3,
  itemIDs = items_diao2_stage2$item)

testLength2 <- itemsPerFormConstraint(
  nForms = 3,
  operator = "=",
  targetValue = 20,
  itemIDs = items_diao2_stage2$item)
```

Contributions

The work on Chapter 4 was supported by the German Academic Exchange Service (DAAD) via funding a research stay (Kurzstipendium für Doktoranden) at ETS Global in Amsterdam, Netherlands.

Chapter 2

The core ideas of the manuscript were developed by BB. The design of the simulation study was developed by BB and DD. The analysis strategy for the empirical example was developed by BB. BB programmed the simulation study and analyzed the data. DD, SW, and FG supervised the work and gave feedback. BB wrote the first draft of the paper. DD, SW, and FG provided feedback on the writing and contributed to the final manuscript.

Chapter 3

The core ideas of the manuscript were developed by BB and DD. Software implementation of the proposed method was conducted by BB and DD. The illustrative examples were designed and programmed by BB. The analysis strategy for the empirical data analyses was developed and implemented by BB and SW. DD and FG supervised the work and gave feedback. BB wrote the first draft of the paper. DD, SW, and FG provided feedback on the writing and contributed to the final manuscript.

Chapter 4

The core ideas of the manuscript were developed by BB. The design of the simulation study and the analysis strategy were developed by BB. PvR and DM organized access to the empirical data sets. BB programmed the simulation study and conducted the empirical data analyses. PvR and DM supervised the work and provided feedback. BB wrote the first draft of the paper. PvR, DM, and DD provided feedback on the writing and contributed to the final manuscript.

Chapter 5

The core ideas of the manuscript were developed by BB, CK, and EU. BB, CK, and EU developed the illustrative examples; CK and EU programmed the illustrative examples. BB, CK, and EU wrote the first draft of the manuscript. BB, CK, and EU provided feedback on each other's writing and contributed to the final manuscript.

Chapter 6

The core ideas of the manuscript were developed by BB and DD. BB and DD programmed and documented the R package described in the manuscript. The illustrative examples were designed and programmed by BB, KAS, SW, and DD. BB wrote the first draft of the paper with KAS, SW, and DD writing first drafts for the illustrative examples. BB, DD, SW, and KAS provided feedback on the writing and contributed to the final manuscript.

Erklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbständig verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht verwendet. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, 31. März, 2023

Benjamin Becker