Using Response Times for Modeling Missing Responses in Large-Scale Assessments

Dissertation
zur Erlangung des akademischen Grades
Doctor Philosophiae (Dr. phil.)
am Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

vorgelegt von
Esther Ulitzsch

Berlin, 2019

Erstgutachterin: Prof. Dr. Steffi Pohl

Zweitgutachter: Prof. Dr. Michael Eid

Tag der Disputation: 28. Februar 2020

Summary

Examinees differ in how they interact with assessments. In low-stakes large-scale assessments (LSAs), missing responses pose an obvious example of such differences. Understanding the underlying mechanisms is paramount for making appropriate decisions on how to deal with missing responses in data analysis and drawing valid inferences on examinee competencies. Against this background, the present work aims at providing approaches for a nuanced modeling and understanding of test-taking behavior associated with the occurrence of missing responses in LSAs. These approaches are aimed at a) improving the treatment of missing responses in LSAs, b) supporting a better understanding of missingness mechanisms in particular and examinee test-taking behavior in general, and c) considering differences in test-taking behavior underlying missing responses when drawing inferences about examinee competencies. To that end, the present work leverages the additional information contained in response times and integrates research on modeling missing responses with research on modeling response times associated with observed responses. By documenting lengths of interactions, response times contain valuable information on how examinees interact with assessments and may as such critically contribute to understanding the processes underlying both observed and missing responses.

This work presents four modeling approaches that focus on different aspects and mechanisms of missing responses. The first two approaches focus on modeling not-reached items. The second two approaches aim at modeling omitted items.

The first approach employs the framework for the joint modeling of speed and ability by van der Linden (2007) for modeling the mechanism underlying not-reached items due to lack of working speed. On the basis of both theoretical considerations as well as a comprehensive simulation study, it is argued that by accounting for differences in speed this framework is well suited for modeling the mechanism underlying not-reached items due to lack thereof. In assessing empirical test-level response times, it is, however, also illustrated that some examinees quit the assessment before reaching the end of the test or being forced to stop working due to a time limit. Building on these results, the second approach of this work aims at disentangling and jointly modeling multiple mechanisms underlying not-reached items. Employing information on response times, not-reached items due to lack of speed are distinguished from not-reached items due to quitting. The former is modeled by considering examinee speed. Quitting behavior – defined as stopping to work before the time limit is reached while there are still unanswered items – is modeled as a survival process, with the item position at which examinees are most

likely to quit being governed by their test endurance, conceptualized as a third latent variable besides speed and ability.

The third approach presented in this work focuses on jointly modeling omission behavior and response behavior, thus providing a better understanding of how these two types of behavior differ. For doing so, the approach extends the framework for jointly modeling speed and ability by a model component for the omission process and introduces the concept of different speed levels examinees operate on when generating responses and omitting items. This approach supports a more nuanced understanding of both the missingness mechanism underlying omissions and examinee pacing behavior through assessment of whether examinees employ different pacing strategies when generating responses or omitting items

The fourth approach builds on previous theoretical work relating omitted responses to examinee disengagement and provides a model-based approach that allows for identifying and modeling examinee disengagement in terms of both omission and guessing behavior. Disengagement is identified at the item-by-examinee level by employing a mixture modeling approach that allows for different data-generating processes underlying item responses and omissions as well as different distributions of response times associated with engaged and disengaged behavior. Item-by-examinee mixing proportions themselves are modeled as a function of additional person and item parameters. This allows relating disengagement to ability and speed as well as identifying items that are likely to evoke disengaged test-taking behavior.

The approaches presented in this work are tested and illustrated by a) evaluating their statistical performance under conditions typically encountered in LSAs by means of comprehensive simulation studies, b) illustrating their advances over previously developed approaches, and c) applying them to real data from major LSAs, thereby illustrating their potential for understanding examinee test-taking behavior in general and missingness mechanisms in particular. The potential of the approaches developed in this work for deepening the understanding of results from LSAs is discussed and implications for the improvement of assessment procedures – ranging from construction and administration to analysis, interpretation and reporting – are derived. Limitations of the proposed approaches are discussed and suggestions for future research are provided.

Zusammenfassung

Personen unterscheiden sich in der Art und Weise wie sie mit Tests interagieren. In low-stakes Large-Scale Assessments (LSAs) stellen starke Variationen im Auftreten fehlender Antworten auf Individual- wie auf Gruppenebene sichtbare Manifestationen solcher Unterschieden dar. Ein umfassendes Verständnis der Mechanismen fehlender Werte ist sowohl für die Entscheidung, wie diese in der Datenanalyse gehandhabt werden sollen, als auch für das Ziehen valider Schlussfolgerungen aus den Testergebnissen von größter Bedeutung. Vor diesem Hintergrund werden in der vorliegenden Arbeit modellbasierte Ansätze zur Modellierung fehlender Antworten weiterentwickelt. Ziel hierbei ist es, a) Methoden für einen verbesserten Umgang mit fehlenden Antworten bereitzustellen, b) ein differenzierteres Verständnis von Testverhalten und insbesondere von zu fehlenden Werte führendem Verhalten zu ermöglichen und c) Unterschiede in diesem Verhalten bei der Analyse und Interpretation von LSA Daten zu berücksichtigen. Hierfür werden Antwortszeiten als zusätzliche Informationsquelle zu Testverhalten nutzbar gemacht und bestehende Forschungszweige zur Modellierung fehlender Werte mit Forschungszweigen zur Modellierung von Antwortszeiten im Kontext beobachteter Itemantworten verknüpft. Antwortszeiten dokumentieren, wie lange TestandInnen mit Items interagieren und können dadurch erheblich zum Verständnis derjenigen Prozesse beitragen, die Itemantworten – sowohl beobachteten als auch fehlenden – zugrunde liegen.

Es werden vier Ansätze vorgestellt, die je verschiedene Aspekte und Mechanismen modellieren. Die ersten beiden Ansätze behandeln die Modellierung nicht erreichter Items, die letzten beiden Ansätze die Modellierung ausgelassener Items.

Zunächst wird das Potential des Ansatzes zur gemeinsamen Modellierung von Fähigkeit und Bearbeitungsgeschwindigkeit von van der Linden (2007) für die Modellierung eines Nicht-Erreichen des Testendes aufgrund mangelnder Bearbeitungsgeschwindigkeit evaluiert. Dies geschieht sowohl auf Grundlage theoretischer Überlegungen als auch mittels einer umfassenden Simulationsstudie. In einer Untersuchung empirischer Bearbeitungszeiten auf Testebene wird jedoch auch aufgezeigt, dass einige TestandInnen die Testung abbrechen noch bevor sie das Ende des Tests erreichen oder an die Zeitgrenze stoßen. Der zweite vorgestellte Ansatz baut auf diesen Ergebnissen auf, indem er mehrere Mechanismen des Nicht-Erreichen des Testendes differenziert und gemeinsam modelliert. Informationen zu Antwortszeiten werden genutzt um zwischen nicht erreichten Items aus mangelnder Bearbeitungsgeschwindigkeit und Abbruch zu unterscheiden. Ersteres wird durch die Berücksichtigung der Bearbeitungsgeschwindigkeit modelliert. Letzteres wird als Verweildauerprozess

konzeptualisiert, bei dem die Itemposition, an der die Testung abgebrochen wurde, als Funktion der Testausdauer der TestandInnen modelliert wird.

Der dritte vorgestellte Ansatz konzentriert sich auf die gemeinsame Modellierung von Auslassungs- und Antwortverhalten. Zu diesem Zweck wird der Ansatz zur gemeinsamen Modellierung von Fähigkeit und Bearbeitungsgeschwindigkeit um die Auslassungsneigung der TestandInnen erweitert und das Konzept verschiedener Bearbeitungsgeschwindigkeiten, einhergehend mit Antwortverhalten einerseits und Auslassungsverhalten andererseits, eingeführt. Dadurch erlaubt der vorgestellte Ansatz ein differenzierteres Verständnis des Auslassungsmechanismus sowie die Untersuchung von verschiedenen Testverhalten und Bearbeitungsgeschwindigkeiten, die mit Antwort- und Auslassungsverhalten einhergehen.

Der vierte Ansatz baut auf theoretischen Arbeiten zu Auslassungen als Indikator mangelnder Anstrengungsbereitschaft auf. Es wird ein modellbasierter Ansatz für die Identifizierung und Modellierung mangelnder Anstrengungsbereitschaft entwickelt, wobei letztere durch sowohl Rate- als auch Auslassungsverhalten definiert ist. Anstrengungsbereitschaft identifiziert der entwickelte Ansatz mittels Mischverteilungsmodellen auf Item-mal-Personenebene. Dazu werden unterschiedliche datengenerierende Prozesse für beobachtete Itemantworten, Antwortszeiten und Auslassungen angenommen, die mit Rate- und Auslassungsverhalten einerseits und anstrengungsbereitem Testverhalten andererseits assoziiert sind. Die Mischungsgewichte auf Item-mal-Personenebene wiederum werden als Funktion zusätzlicher Personen- und Itemparameter modelliert. Auf diese Weise können Personen- und Itemcharakteristiken untersucht werden, die mit geringer Anstrengungsbereitschaft einhergehen; Anstrengungsbereitschaft, Fähigkeit und Bearbeitungsgeschwindigkeit können miteinander in Beziehung gesetzt oder Items identifiziert werden, die geringe Anstrengungsbereitschaft evozieren.

In der vorliegenden Arbeit werden alle Ansätze getestet und illustriert, indem a) ihre statistische Leistung unter für LSAs typischen Bedingungen mittels umfassender Simulationsstudien untersucht wird, b) die Vorteile der Modellierung der berücksichtigten Verhaltensweisen durch Vergleiche mit bisherigen Ansätzen herausgearbeitet werden sowie c) ihr Potential für das Verständnis von Testverhalten und insbesondere von Verhalten, welches zum Auftreten fehlender Werte führt, anhand empirischer Daten großer LSAs veranschaulicht wird. Die Potentiale der vorgestellten Ansätze zum Verständnis der Ergebnisse von LSAs werden abschließend diskutiert und Implikationen für die Analyse, Administrierung, Interpretation und Berichterstattung von LSAs abgeleitet. Limitationen der vorgestellten Ansätze werden problematisiert und Vorschläge für zukünftige Forschung formuliert.

Danksagung

An erster Stelle möchte ich mich bei *Steffi Pohl* für die herausragende Betreuung sowie die fachliche und persönliche Unterstützung und Förderung bedanken. Liebe Steffi, vielen Dank für alles – für das gemeinsame Knobeln und anregende Diskussionen, die vielen Möglichkeiten, die Du mir eröffnet hast, dafür, dass Du mir beigebracht hast den roten Faden und Fokus nicht aus dem Blick zu verlieren und nicht zuletzt für eine Promotionszeit, die ich mir schöner nicht hätte vorstellen können.

Besonderer Dank gilt auch *Matthias von Davier*, dessen großes Wissen und Begeisterung für Psychometrie mich immer wieder aufs Neue beeindruckt haben. Vielen Dank für die Einblicke, die Möglichkeiten zum Austausch und Diskussionen, die ich während meiner gesamten Promotionszeit, insbesondere aber zur Zeit meines Forschungsaufenthalts am National Board of Medical Examiners, als sehr bereichernd und motivierend empfunden habe.

Michael Eid danke ich dafür, meine Begeisterung für die Methodenforschung geweckt zu haben und für die Bereitschaft, diese Arbeit zu begutachten.

Benjamin Becker, *Johanna Hildebrandt*, *Emilija Meier-Faust* und *Daniel Schulze* gilt besonderer Dank für kritische Anmerkungen und Rückmeldung zur Rahmung der vorliegenden Arbeit, viel guten Zuspruch und entspannende Kaffee- und Kuchenpausen. Bei *Alexander Battaglia* möchte ich mich für die Erlaubnis bedanken, seinen Rechner für Bayesianische Schätzung zu zweckentfremden sowie für die Unterstützung bei Formulierungsfragen.

Randy Eichentopf, *Kevin Hoppe*, *Nina Pérez Delgado* und *Robert Reggentin* danke ich für Tabellenformatierung und sehr genaue Blicke auf das Literaturverzeichnis.

Die Simulationsstudien dieser Arbeit wären nicht durchführbar gewesen ohne die Bereitstellung von Rechenressourcen durch den *HPC Service der Freien Universität Berlin* und dessen unermüdliche Unterstützung bei technischen Hürden.

Tiefe Dankbarkeit empfinde ich gegenüber meinem Bruder *Vincent Ulitzsch*, der mich während der gesamten Phase der Entstehung dieser Arbeit in jeglicher Hinsicht unterstützt hat.

Herzlichster Dank gilt auch meinen Freunden für methodenferne Ablenkungen (besonders willkommen: Berge, Schwimmen und Essen).

Contents

Summary i

Zusammenfassung iii

Danksagung v

1 Introduction 1
- 1.1 Missing Responses in Large-Scale Assessments . . . 2
  - 1.1.1 Terminology . . . 2
  - 1.1.2 Prevalence of Missing Responses . . . 3
  - 1.1.3 Importance of Understanding Missingness Mechanisms . . . 3
  - 1.1.4 A Measurement Perspective on Missing Responses . . . 3
  - 1.1.5 An (Incomprehensive) Overview of Possible Missingness Mechanisms . . . 6
- 1.2 Dealing with Missing Responses Employing Information Retrievable from Paper-and-Pencil-Based Assessment . . . 7
  - 1.2.1 Classical Approaches . . . 7
  - 1.2.2 Model-Based Approaches . . . 8
- 1.3 Additional Insights Gained from Response Time Data . . . 12
- 1.4 Modeling Responses and Response Times Simultaneously but Separately 13
  - 1.4.1 Advantages of Jointly Modeling Responses and Response Times 15
- 1.5 Response Times and Missing Responses . . . 16
  - 1.5.1 Response Times and Not-Reached Items . . . 17
  - 1.5.2 Using Response Times for Coding Omissions . . . 18
- 1.6 Aims and Scope of the Present Work . . . 20

2 Using Response Times to Model Not-Reached Items due to Time Limits 23
- 2.1 Modeling Missing Values within IRT . . . 26
  - 2.1.1 Classical Approaches . . . 26
  - 2.1.2 Model-Based Approaches for Nonignorable Missing Responses 27
  - 2.1.3 The Impact of Response Time . . . 28
- 2.2 Response Times Informing the Missing Response Process . . . 28
- 2.3 Response Time Modeling within IRT . . . 30

2.3.1 Hierarchical Speed-Accuracy (SA) Model 30
2.4 Objectives 32
2.5 Method 33
2.5.1 Data Generation 34
2.5.2 Data Analysis 35
2.6 Results 36
2.6.1 Convergence and Efficiency 36
2.6.2 Performance of the Speed-Accuracy Model for Complete Data . 37
2.6.3 Performance of the Speed-Accuracy Model to Deal with Not-Reached Items 38
2.6.4 Illustrating the Shrinkage Effect due to Missing Values 40
2.6.5 Performance of the Manifest Missing Response Model for Incomplete Data 41
2.6.6 Additional Analyses 43
2.7 Empirical Data Analysis 45
2.8 Discussion 47
2.8.1 Implications for the Practice of Dealing with Not-Reached Items due to Time Limits 49

3 A Multi-Process Item Response Model for Not-Reached Items due to Time Limits and Quitting 51
3.1 Dealing with Not-Reached Items in Large-Scale Assessments 52
3.1.1 Model-Based Approaches for Nonignorable Missing Values . . . 53
3.1.2 Using Response Times to Model Not-Reached Items 55
3.2 Objective 56
3.3 Speed-Accuracy+Quitting Model 56
3.3.1 Identifying Quitting 57
3.3.2 Modeling Quitting 57
3.3.3 Second Level Models 59
3.4 Parameter Recovery 61
3.4.1 Data Generation 61
3.4.2 Estimation Procedure 62
3.4.3 Evaluation Criteria 64
3.4.4 Results 65
3.5 Empirical Example 69
3.5.1 Total Response Time Distributions 69
3.5.2 Investigating the Occurrence of Not-Reached Items 70
3.6 Discussion 71
3.6.1 Limitations and Future Research 73

4 Using Response Times for Joint Modeling of Response and Omission Behavior 75
4.1 Omitted Responses in Large-Scale Assessments 76
4.2 Approaches for Nonignorable Missing Responses 77
4.3 Information Obtained from Computer-Based Assessment 78
4.3.1 Timing Data and Omitted Items 79
4.4 Objective 81
4.5 Proposed Model 81
4.5.1 Modeling Response Behavior 83
4.5.2 Modeling Nonresponse Behavior 84
4.5.3 Second-Level Models 86
4.5.4 Prior Distributions 87
4.6 Parameter Recovery 88
4.6.1 Data Generation 89
4.6.2 Estimation Procedure 90
4.6.3 Evaluation Criteria 90
4.6.4 Results 91
4.7 Impact of Modeling Nonresponse Behavior on Ability Estimation . . . 98
4.8 Investigating Nonresponse Behavior with the Speed-Accuracy+ Omission Model 101
4.8.1 Estimation and Model Checking 101
4.8.2 Results 104
4.9 Discussion 106
4.9.1 Limitations and Future Research 110

5 A Hierarchical Latent Response Model for Inferences about Examinee Engagement in Terms of Guessing and Item-Level Nonresponse 112
5.1 Previous Approaches for Identifying and Handling Disengaged Behavior 113
5.1.1 Guessing and Perfunctory Answers 113
5.1.2 Omissions 116
5.2 Proposed Model 118
5.2.1 Engaged Behavior 119
5.2.2 Disengaged Behavior 120
5.2.3 Higher-Order Models 121
5.3 Prior Distributions 122
5.4 Parameter Recovery 123
5.4.1 Data Generation 123
5.4.2 Estimation Procedure 124
5.4.3 Results 124
5.5 Illustrating the Model 129

5.6 Empirical Example . . . 134
5.6.1 Estimation and Model Checking . . . 135
5.6.2 Results . . . 135
5.7 Discussion . . . 137
5.7.1 Limitations and Future Directions . . . 139

6 Discussion 142
6.1 Advantages of Using Response Times for Modeling Missing Responses 143
6.1.1 Modeling and Understanding Missingness Mechanisms . . . 143
6.1.2 Investigating Differences in Test-Taking Behavior . . . 144
6.1.3 Enhancing Ability and Item Parameter Estimation . . . 145
6.2 Limitations and Directions for Future Research . . . 145
6.2.1 Modeling Omissions and Not-Reached Items Jointly . . . 146
6.2.2 Identifying Subpopulations Differing in Test-Taking Behavior . . 147
6.2.3 Allowing for Varying Test-Taking Behavior Across the Test . . . 148
6.2.4 Model Validation . . . 148
6.2.5 Dealing with Operational Challenges of Large-Scale Assessments 149
6.2.6 Considering Additional Data on Test-Taking Behavior . . . 150
6.2.7 Modeling Missingness Mechanisms in Noncognitive Assessments 151
6.3 Recommendations for Model Application . . . 152
6.4 Implications . . . 153
6.4.1 Implications for Test Construction . . . 153
6.4.2 Implications for Test Administration . . . 154
6.4.3 Implications for Analysis of Large-Scale Assessment Data . . . 155
6.4.4 Implications for Reporting on Results of Large-Scale Assessments 156
6.5 Conclusion . . . 159

References 160

A Appendix to Chapter 2 177
A.1 JAGS Code and Prior Settings . . . 177
A.1.1 JAGS Code: Speed-Accuracy Model . . . 178
A.1.2 JAGS Code: Manifest Missing Data Model . . . 179
A.2 Differences in Speed Estimates . . . 180
A.3 Subsequent Analyses . . . 182

B Appendix to Chapter 3 183
B.1 JAGS Code . . . 183
B.2 Coverage . . . 185
B.3 Parameter Recovery . . . 187

C Appendix to Chapter 4 189
C.1 Stan Code . . . 189
C.2 Coverage . . . 192
C.3 Parameter Recovery . . . 195
C.4 Posterior Predictive Checks . . . 216

D Appendix to Chapter 5 218
D.1 Special Cases of the SA+E Framework . . . 218
D.1.1 Guesses Only . . . 218
D.1.2 Omissions Only . . . 218
D.2 Stan Code . . . 220
D.3 Parameter Recovery . . . 221
D.4 Posterior Predictive Checks . . . 224

Contributions 227

Erklärung 228

1
Introduction

Large-scale assessments (LSAs) aim at measuring examinee competencies (von Davier, Gonzalez, Kirsch, & Yamamoto, 2013). Their results are of increasing importance, driving discussions on educational systems and informing policy decisions (e.g, Addey, Sellar, Steiner-Khamsi, Lingard, & Verger, 2017; UNESCO Institute for Statistics, 2018). When comparing differences in performance, it is assumed that these solely go back to differences in competencies. However, this interpretation is in jeopardy when examinees not only differ in competency but also in the way they approach the assessment. As such, differences in test-taking behavior might pose an additional source of variation between examinees. For LSAs to allow valid inferences on competency differences, recent policy papers on the analysis, interpretation, and communication of LSA data thus called for "unpacking" examinee performance; that is, to identify and describe sources of differences in performance that go beyond differences in competencies (Singer & Braun, 2018). This work will follow their call by focusing on test-taking behavior associated with the occurrence of missing responses as an additional source of variation between examinees.

In LSAs it is rather common that examinees do not provide answers to all items administered, either due to omitting items or due to not reaching the end of the assessment. Understanding the underlying processes is important for at least two reasons: First, missing responses force researchers to explicate their beliefs on the nature of the underlying behavior, which becomes evident in their decision on how missing data are treated in analysis. For an adequate decision, a comprehensive understanding of test-taking behavior underlying the occurrence of missing responses is paramount. Second, there is large variation in the occurrence of missing responses across examinees and countries, indicating differences in how examinees approach the assessment. Understanding missingness mechanisms thus supports a better understanding of differences in test-taking behavior in general and allows considering these differences when drawing inferences on examinee competencies.

The present work aims at providing approaches for modeling test-taking behavior underlying missing responses, thereby a) providing tools for improving the treatment

of missing responses due to examinee behavior in LSAs, b) supporting a more nuanced understanding of missingness mechanisms in particular and examinee test-taking behavior in general, and c) considering differences in test-taking behavior underlying missing responses when drawing inferences on examinee competencies. To that end, this work leverages the additional information contained in response times (RTs). By documenting the length of interactions on both the item as well on the test level, RTs contain valuable information on essential features of processes underlying both observed and missing responses, namely on their duration. As such, RTs may critically contribute to understanding the processes operating when examinees interact with an assessment in general and the processes underlying the occurrence of missing responses in particular.

In what follows, the approaches introduced in this work will be further motivated and shortly summarized. First, the necessity of and the advantages coming with modeling mechanisms underlying missing responses are discussed. Second, an overview of previous approaches for dealing with missing responses based on information retrievable from paper-and-pencil-based assessment is provided. Third, the potential of considering RTs for understanding and modeling test-taking behavior in general and behavior underlying missing responses in particular is discussed and previous approaches utilizing RTs for modeling response behavior and handling missing responses are reviewed. Based on these considerations, the objectives of the present work are derived and its approaches will be shortly discussed. In the main body of this work, four approaches for modeling behavioral processes associated with missing responses are presented, with two approaches focusing on the occurrence of not-reached items and two on item omissions.

1.1 Missing Responses in Large-Scale Assessments

1.1.1 Terminology

This work focuses on missing observations on items that, although administered, have not been responded to. Two types of such missing responses are distinguished in LSA data: omitted and not-reached items (NRIs). Items are said to be omitted when the examinee has seen an item but, for whatever reason, has decided not to respond (Mislevy & Wu, 1996). In the case that an examinee failed to attempt a sequence of items presented at the end of a test, the resulting missing responses are referred to as not-reached items (Mislevy & Wu, 1996).

1.1.2 Prevalence of Missing Responses

Missing responses occur to a considerable degree in LSAs. At the same time, there is large variation in missingness rates across countries, time, studies, and domains. For instance, in 2012, omission rates within the Programme for the International Assessment of Adult Competencies (PIAAC) ranged from 2% for the numeracy domain in Korea to 25.9% for the literacy domain in Chile (OECD, 2013). Likewise, there is strong variation in whether or not examinees reach the end of the assessment: The Progress in International Reading Literacy Study on online informational reading (ePIRLS) 2016, for instance, reported that 2% of examinees in Singapore did not reach the end of the test, while in Georgia this number was as high as 29% (Foy, 2018). These large differences in the prevalence of missing responses indicate large differences in how examinees approach the assessment.

1.1.3 Importance of Understanding Missingness Mechanisms

Prior to analyzing LSA data, researchers need to decide how to deal with missing responses. A better comprehension of the behavior underlying the occurrence of missing responses may help to inform the proper treatment of missing data when estimating parameters of interest (Jakwerth, Stancavage, & Reed, 2003; Köhler, Pohl, & Carstensen, 2015a). In addition, by indicating differences in test-taking behavior, missing values due to examinee behavior support insights into how examinees approached the assessment and might as such be of great value for understanding examinee performance. Performance decline in Ireland from 2000 to 2009 in the Programme for International Student Assessment (PISA, OECD, 2017), for instance, has been attributed to a decline of test-taking motivation rather than a decline in competency, with an increase in item omissions considered being indicative thereof (Cosgrove, 2011). In the present work, it is therefore argued to treat missing responses as a valuable source of information supporting a more comprehensive understanding of differences in test-taking behavior, and, as such, differences in performance, rather than a mere nuisance need to be dealt with.

1.1.4 A Measurement Perspective on Missing Responses

For the last decades, item response theory (IRT) models have been the predominant measurement method in LSAs (von Davier et al., 2013). IRT is grounded on the assumption that the probability of a correct answer provided by examinee i to item j can be described as a function of the examinee's location on the latent construct to be measured and one or more parameters characterizing the particular item (Molenaar,

1995). Let $u_{ij} = 1$ and $u_{ij} = 0$ indicate a correct and incorrect response, respectively. In the most simple IRT model, the Rasch model, the probability of observing a correct response $p(u_{ij} = 1)$ is modeled as a function of examinee ability θ_i and item difficulty b_j:

$$p(u_{ij} = 1) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}. \tag{1.1}$$

In IRT as employed in LSAs, item and person parameters are usually modeled as fixed and random effects, respectively (Molenaar, 1995). Assuming a normal distribution for examinee ability θ yields the following likelihood function

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{j=1}^{K} p(u_{ij}|b_j, \theta_i)^{1-d_{ij}} g(\theta_i|\mu_\theta, \sigma_\theta), \tag{1.2}$$

where $g(\theta_i|\mu_\theta, \sigma_\theta)$ denotes the normal density of the latent trait in the population, with μ_θ and σ_θ giving the mean and standard deviation of the distribution. In the presence of missing responses, d_{ij} denotes whether or not a response of examinee i to item j was observed, with $d_{ij} = 0$ denoting an observed and $d_{ij} = 1$ a missing response. Thus, estimates of parameters of interest – in this case difficulty parameters as well as parameters of the distribution of ability – are estimated based on observed responses, while unobserved responses do not contribute to the likelihood function and are thus ignored.

Whether or not ignoring missing responses poses a threat to validity and unbiasedness of the conclusions drawn from LSAs depends on the missingness mechanisms. The important role of missingness mechanisms for the decision on how to deal with missing values in the analysis of incomplete data sets has first been fully acknowledged and formalized in the theory of Rubin (1976) by treating missingness as a probabilistic phenomenon and, on the basis of the distributional properties of missingness, deriving conditions under which missing data can and cannot be ignored in data analysis (Little & Rubin, 2002). Note that in this work, instead of Rubin's original notation, the notation and terminology introduced by Little and Rubin (2002) are employed and adapted to the context of IRT and the notation of the present work.

For a N subjects by K item response indicator matrix $\mathbf{U}$, a missingness indicator matrix $\mathbf{D}$ of the same dimensions contains information on which elements of $\mathbf{U}$ have been observed and which are missing. That is, each element of $\mathbf{D}$ is defined as

$$d_{ij} = \begin{cases} 0 & \text{if } u_{ij} \text{ is observed} \\ 1 & \text{if } u_{ij} \text{ is not observed.} \end{cases} \tag{1.3}$$

$\mathbf{D}$ is treated as a set of random variables whose distribution is referred to as missing data mechanism (Schafer & Graham, 2002). Based on the conditional distribution of $\mathbf{D}$ given $\mathbf{U}$, $f(\mathbf{D}|\mathbf{U}, \psi)$, where ψ denotes the unknown parameters of the distribution of $\mathbf{D}$, Rubin (1976) has derived conditions under which missing data can be ignored without threatening valid inferences. For doing so, Rubin (1976) has distinguished between the observed $\mathbf{U}_{obs}$ and missing components $\mathbf{U}_{mis}$ of $\mathbf{U}$.

MISSING COMPLETELY AT RANDOM Data are missing completely at random (MCAR) if, for all possible values of ψ, missingness does not depend on the data $\mathbf{U}$, missing or observed:

$$p(\mathbf{D}|\mathbf{U}, \psi) = p(\mathbf{D}|\psi) \quad \text{for all } \mathbf{U}, \psi. \tag{1.4}$$

Planned missingness occurring due to incomplete block designs (see Gonzalez & Rutkowski, 2010; Mislevy & Wu, 1996) poses a typical example for responses being MCAR. In assessments with incomplete block designs such as PISA (OECD, 2017), examinees are administered different item blocks consisting of a fraction of all items available in that assessment, with the blocks an examinee receives being assigned randomly.

MISSING AT RANDOM The missing data mechanism satisfies missing at random (MAR) if, for all possible values of ψ, missingness depends on observed values, however, not on missing ones. That is,

$$p(\mathbf{D}|\mathbf{U}, \psi) = p(\mathbf{D}|\mathbf{U}_{obs}, \psi) \quad \text{for all } \mathbf{U}_{mis}, \psi. \tag{1.5}$$

An example for responses being MAR in LSAs are missing responses occurring due to multistage adaptive testing (Mislevy & Wu, 1996). In PIAAC, for instance, examinees are routed to item blocks in an adaptive way based on their performance in previous blocks, such that more able examinees are more likely to receive a more difficult set of items (OECD, 2013).

MISSING NOT AT RANDOM The missing data mechanism is missing not at random (MNAR) if Equation 1.5 is violated, that is, if the conditional distribution of $\mathbf{D}$ given $\mathbf{U}_{obs}$ and ψ does depend on the missing values $\mathbf{U}_{mis}$. Hence, MNAR can be written as

$$p(\mathbf{D}|\mathbf{U}, \psi) \neq p(\mathbf{D}|\mathbf{U}_{obs}, \psi). \tag{1.6}$$

In assessment data, item omissions occurring because examinees could not solve the items they omitted represent an example of missing responses being MNAR since the probability of a missing response depends on the missing response itself (Mislevy & Wu, 1996).

DISTINCTNESS In addition, Rubin (1976) has introduced the condition of distinctness of ψ from the parameters of interest χ – e.g., item and person parameters. For likelihood-based inference, this implies that the joint parameter space of $(\chi; \psi)$ is the product of the parameter space of χ and the parameter space of ψ. For Bayesian inference, this implies that prior distributions for χ and ψ are independent (Rubin, 1976).

An intuitive example for a violation of distinctness in the context of LSAs are NRIs occurring due to lack of speed in the case that the level of ability examinees show on the assessment is related to the level of speed with which they generate responses.

IGNORABILITY The missing data mechanism is ignorable in the case that the missing data process is either MCAR or MAR and ψ is distinct from parameters of interest χ (Rubin, 1976). Under these conditions, $\mathbf{D}$ does not contain additional information on the parameters to be estimated above and beyond the observed data $\mathbf{U}_{obs}$. This, in turn, implies that missing data can be ignored in data analysis and that it is not necessary to incorporate the missingness mechanism into models for the observed data processes (Holman & Glas, 2005). If, however, ignorability does not hold and the missing data process is not considered, validity of both likelihood-based and Bayesian inference is jeopardized. Under such conditions in order to draw unbiased inferences about the parameters of interest, there is a need to establish a model for the processes that cause missing data (Rubin, 1976).

1.1.5 An (Incomprehensive) Overview of Possible Missingness Mechanisms

Based on theoretical considerations as well as empirical evidence, various mechanisms have been discussed in the literature as underlying item omissions. These range from lack of confidence in the correct answer (Jakwerth et al., 2003; Mislevy & Wu, 1996), to lack of willingness to engage with the assessment, respectively lack of motivation (Cosgrove, 2011; van Barneveld, Pharand, Ruberto, & Haggarty, 2013), fatigue, or refusal to participate (OECD, 2013). NRIs are often attributed to lack of speed and reaching the time limit (Mislevy & Wu, 1996; Tijmstra & Bolsinova, 2018). However, examinees have also been found to stop the assessment before reaching

the time limit or the end of the test due to, e.g., feeling overtaxed to solve the items administered or being unwilling to further respond (OECD, 2013).

Above and beyond, in low-stakes LSAs, omission behavior is often seen as an indicator of lack of test-taking motivation (Cosgrove, 2011; Sachse, Mahler, & Pohl, 2019; van Barneveld et al., 2013; Wise & Gao, 2017). Although coming with major implications for policy makers and society at large, test performance in LSAs comes with little or no consequences for examinees themselves. This renders it highly probable that (at least some) examinees might not be fully willing "to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 1). Instead, they "may opt to skip questions, guess randomly, mark patterns, fail to review their answers for accuracy before handing in their work, or quit answering assessment items entirely" (van Barneveld et al., 2013, p. 44). In the case that a portion of item omissions goes back to lack of test-taking motivation it is also likely for some observed item responses to stem from disengaged test-taking behavior such as guessing or perfunctory answering (Wise & Gao, 2017). If so, the data-generating process implied by models of IRT (see Equation 5.1) does not hold for all observed responses.

Empirically, the occurrence of both item omissions and NRIs has often been found to be related to ability (Köhler et al., 2015a; Pohl, Gräfe, & Rose, 2014; Robitzsch, 2014; Sachse et al., 2019; van den Wollenberg, 1979). This indicates that, whatever the specific processes underlying missing responses, examinees with different levels of ability differ in test-taking behavior yielding missing responses. Hence, item omissions and NRIs are highly likely to be nonignorable. Thus, for not jeopardizing validity of inferences drawn from LSAs, the underlying mechanisms need to be accounted for.

1.2 Dealing with Missing Responses Employing Information Retrievable from Paper-and-Pencil-Based Assessment

1.2.1 Classical Approaches

Operationally in LSAs, it is common to deal with missing values due to item omissions and NRIs by either ignoring missing responses or by scoring them as (partially) incorrect. Two-stage approaches can also be encountered. The Trends in International Mathematics and Science Study (TIMSS) and PIRLS, for instance, ignore item omissions when calibrating item parameters and subsequently treat missing responses as incorrect when estimating ability parameters (Foy, 2017, 2018).

ASSUMPTIONS AND LIMITATIONS Both scoring missing responses as (partially) incorrect as well as ignoring missing responses comes with strong assumptions concerning the data-generating processes underlying missing responses. Ignoring missing data implies ignorability of the underlying missingness mechanisms (Rose, von Davier, & Xu, 2010). That is, by ignoring missing responses it is assumed that missing responses are MAR given the observed responses and the considered background variables. Empirical evidence (e.g., Köhler et al., 2015a; Pohl et al., 2014; Rose et al., 2010) as well as substantial considerations (e.g., Mislevy & Wu, 1996) on examinee behavior during the test, however, suggest that omitted responses as well as NRIs are likely to be nonignorable and thus need to be accounted for. Not accounting for nonignorable missing responses can have a strong impact on the conclusions drawn from assessment data and has been shown to yield biased person and item parameter estimates (Pohl et al., 2014; Rose, 2013), distort country rankings (Köhler, Pohl, & Carstensen, 2017; Rose et al., 2010), as well the conclusions drawn from trend analyses (Sachse et al., 2019).

Approaches scoring missing responses as incorrect assume that in assessment data, there is no true response "hidden" by the missing value, but that instead missing responses represent a separate response category, as well as that missing responses indicate that the examinee did not know the answer (Rohwer, 2013). Scoring approaches, however, collide with assumptions of IRT. While IRT is grounded on the assumption that the probability of a correct response is always greater than zero and increases monotonically as a function of ability, scoring approaches implicitly assume that the probability of solving a missing item is zero or equals a fixed expression, regardless of the examinee's ability level (Lord, 1974; Rose, 2013).

1.2.2 Model-Based Approaches

To model mechanisms underlying nonignorable missing values when estimating the parameters of interest, various models have been introduced (see Heckman, 1977; Little, 1993, for early developments of such models). Within IRT, model-based approaches for nonignorable missing data have in common that they perceive missing data patterns to be a result of a person-specific tendency to (not) respond. This tendency is incorporated into the IRT model either via an additional latent (Debeer, Janssen, & Boeck, 2017; Glas & Pimentel, 2008; Holman & Glas, 2005; List, Köller, & Nagy, 2019; Moustaki & Knott, 2000) or manifest variable (Rose, von Davier, & Nagengast, 2017; Rose et al., 2010). In the following, when reviewing latent model-based approaches, it will be focused on modeling item omissions. When reviewing manifest model-based approaches, it will be focused on modeling NRIs. Note that

latent approaches have also been adjusted to the context of NRIs (Debeer et al., 2017; Glas & Pimentel, 2008; List et al., 2019) and there are manifest approaches for item omissions (Rose et al., 2010).

LATENT MODEL-BASED APPROACH FOR NONIGNORABLE ITEM OMISSIONS Latent model-based approaches for nonignorable item omissions (Debeer et al., 2017; Holman & Glas, 2005; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999) incorporate an additional latent variable, conceptualized as the examinees' omission propensity, into the model for ability estimation. This is commonly done within a multidimensional IRT framework as depicted in Figure 1.1.

For response indicators u_{ij} customary IRT models are employed, e.g., a Rasch model as given by Equation 5.1. Omission indicators d_{ij} constitute the measurement model for the examinees' latent omissions propensity, taking the value 1 if examinee i omitted item j and 0 if a response was observed. The measurement model for omission propensity can be parameterized employing either a two-parameter logistic model (2PL) or a Rasch model. However, since missing data is usually sparse, it is recommended to employ a Rasch model (Holman & Glas, 2005; Pohl et al., 2014). That is, the probability that examinee i omits item j is assumed to be a function of examinee omission propensity ξ_i and item omission difficulty a_j:

$$p(d_{ij} = 1) = \frac{\exp(a_j - \xi_i)}{1 - \exp(a_j - \xi_i)}. \tag{1.7}$$

Missing propensity ξ and ability θ are assumed to follow a multivariate normal distribution with mean vector

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\xi) \tag{1.8}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\xi} \\ \sigma_{\theta\xi} & \sigma_\xi^2 \end{pmatrix}. \tag{1.9}$$

Doing so yields the following likelihood

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{j=1}^{K} p(u_{ij}|b_j, \theta_i)^{(1-d_{ij})} f(d_{ij}|a_j, \xi_i) g(\theta_i, \xi_i|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}), \tag{1.10}$$

with $g(\theta_i, \xi_i|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ denoting the bivariate normal density of the person parameters. Modeling the omission mechanism in terms of an additional latent variable (omission propensity) and assuming a joint distribution of omission propensity and ability corresponds to the assumption that item omissions are MAR given examinee

omission propensity. As such, in the model, the correlation between θ and ξ indicates the degree of ignorablility in the data, with higher deviations of the correlation from zero indicating a higher degree of nonignorability in the sense that the parameters governing the distribution of observed responses are not distinct from the parameters governing the distribution of item omissions (Holman & Glas, 2005; Pohl & Becker, 2019).

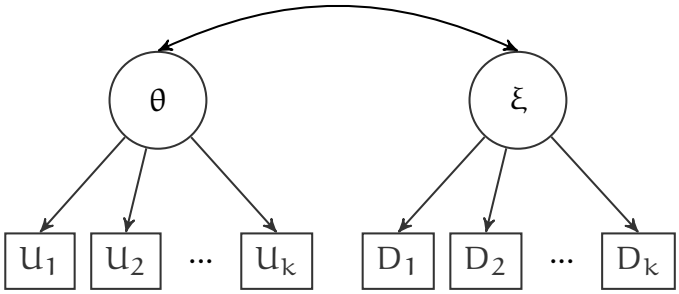

Figure 1.1. Conceptual path diagram for the latent model-based approach for nonignorable item omissions. $\mathbf{U}_j$ and $\mathbf{D}_j$ represent the response and omissions indicators for item j. θ and ξ represent latent ability and omission propensity.

The model has been subject to further developments allowing for the relationship between ability and omission propensity to differ from multivariate normal, omission propensity to be item-type specific, or for including covariates explaining differences in omission propensity (Glas, Pimentel, & Lamers, 2015; Köhler, Pohl, & Carstensen, 2015b).

MANIFEST MODEL-BASED APPROACH FOR NONIGNORABLE NOT-REACHED ITEMS Due to the monotone missingness pattern resulting from NRIs, where all items preceding the first NRI are reached and all items subsequent to the first NRI are missing, all information on the missingness pattern following from NRIs is contained in the person-level number (or proportion) of NRIs (Rose, 2013; Rose et al., 2017). Therefore, for modeling NRIs, Rose et al. (2010) suggested to consider information on the person-level proportion of NRIs $\bar{d}_i$ as a manifest variable in the background model. This can be achieved by either regressing ability θ_i on $\bar{d}_i$ or by applying multi-group IRT models where stratification on $\bar{d}_i$ serves as a grouping variable (Rose et al., 2010). The approach is illustrated in Figure 1.2. The model is currently implemented in the PISA analysis framework for handling NRIs (OECD, 2017).

ASSUMPTIONS AND LIMITATIONS Models aiming at modeling the missingness mechanisms underlying item omissions and NRIs rely on assumptions concerning the nature of these processes, some of which might not be met in LSA data. Köhler et al. (2015b), Pohl and Becker (2019) as well as Robitzsch (2014) addressed and

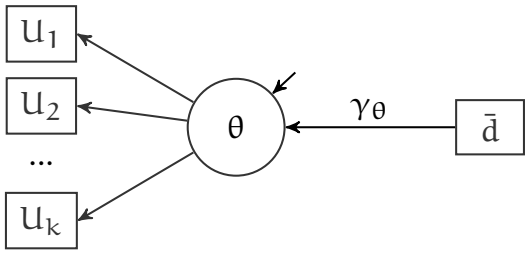

Figure 1.2. Conceptual path diagram for the manifest model-based approach for nonignorable not-reached items by Rose, von Davier, and Xu (2010). $\mathbf{U}_j$ represents the response indicators for item j. θ denotes latent ability. $\bar{\mathbf{d}}$ gives the proportion of not-reached items.

evaluated some of the major underlying assumptions for model-based approaches for nonignorable item omissions. Köhler et al. (2015b) suggested model adjustments of latent model-based approaches for item omissions that a) allow for omission propensity to differ for different item types as well as b) incorporate a distributional assumption for the joint distribution of ability and omission propensity more flexible than a multivariate normal. In addition, the authors showed that the assumption of a unidimensional omission propensity as well the assumption of a multivariate normal distribution of ability and omission propensity oftentimes are not met in real LSA data. Pohl and Becker (2019) and Robitzsch (2014) addressed the assumption concerning the kind of nonignorability considered by model-based approaches for omitted items: By modeling the joint distribution of ability and omission propensity, model-based approaches for item omissions account for missingness mechanisms that are nonignorable in the sense that the parameters governing the distribution of observed responses are not distinct from the parameters governing the distribution of item omissions. As such, they assume item omissions to be MAR given the parameters of the omission model. Pohl and Becker (2019) further investigated the consequences of violations of this assumption for the performance of model-based approaches for item omissions and showed that model-based approaches in their present form can not properly deal with missingness mechanisms that directly depend on the unobserved response as would, for instance, be the case when examinees omit items they otherwise would have responded to incorrectly. Robitzsch (2014) has provided means for conducting sensitivity analyses on how violations of the kind of nonignorability assumed in model-based approaches affects conclusions on examinee ability.

A further limitation of model-based approaches for NRIs and omitted items lies in the scope of conclusions that can be gained on the basis of these models. Although model-based approaches for item omissions and NRIs allow for conclusions

on examinee and item characteristics associated with the occurrence of missing responses (Köhler et al., 2015a), they do not allow for further insights into the nature of test-taking behavior underlying the occurrence of NRIs and omitted items.

This work aims at filling this gap by providing models that allow for a more nuanced understanding of the test-taking behavior underlying NRIs and omitted items by considering additional information on test-taking behavior that goes beyond responses, information on missing values, and covariates. When doing so, this work focuses on leveraging the additional information contained in RTs retrievable from computerized testing. By documenting the duration of interactions on the item as well as on the test level, these pose a valuable source of information on the processes underlying both observed and missing responses.

1.3 Additional Insights Gained from Response Time Data

Using RTs for inferring the nature of cognitive processes has a long tradition in psychology and is a key element for drawing inferences about cognitive and behavioral processes in a variety of paradigms and theoretical frameworks (see De Boeck & Jeon, 2019; Kyllonen & Zu, 2016; Schnipke & Scrams, 2002, for overviews). Frameworks utilizing RTs for identifying and modeling cognitive or behavioral differences are built on the rationale that differences in RTs are indicative of qualitative or quantitative differences in cognitive processes that differ in the time required for their execution. Hence, RTs pose "natural and evident kinds of data to investigate processes" (De Boeck & Jeon, 2019, p. 1). With LSAs moving to computerized testing, it becomes common practice to register RTs for all item responses, resulting in parallel data: For every item examinees have interacted with, information on both the response given and the time examinees interacted with that particular item becomes available. This has stimulated a rapidly growing body of research on how to integrate item-by-examinee level RTs with IRT models for a more comprehensive understanding for the processes underlying observed responses. Such approaches either integrate RTs with response models (Molenaar & De Boeck, 2018; Roskam, 1987; Verhelst, Verstralen, & Jansen, 1997), incorporate responses in models for RTs (Lee, 2007; Thissen, 1983) or model RTs and responses simultaneously but separately (van der Linden, 2007). Exhaustive overviews of psychometric models considering RTs can be found in De Boeck and Jeon (2019), Lee and Chen (2011), Schnipke and Scrams (2002), and van der Linden (2007). Considering the additional information contained in RTs when modeling response processes has repeatedly been proven to be of immense value from both a measurement as well as a substantive perspective, strengthening estimation of both person and item parameters (e.g., Ranger & Kuhn, 2012b; van

der Linden, Klein Entink, & Fox, 2010) as well as allowing for inferences on how examinees allocate their time during the assessment (e.g., Fox & Marianti, 2016) or detecting differences in response processes (e.g., Molenaar, Oberski, Vermunt, & De Boeck, 2016; Partchev & De Boeck, 2012; Wang & Xu, 2015).

1.4 Modeling Responses and Response Times Simultaneously but Separately

For building on current practices for modeling observed responses jointly with the associated RTs, this work builds on the hierarchical speed-accuracy (SA) model by van der Linden (2007) and further developments thereof. Van der Linden (2007) proposed to integrate RTs with the modeling of item responses within a hierarchical framework, allowing for modeling RTs and responses simultaneously but separately (Figure 1.3). In the SA model, responses and RTs comprise indicators for first-level measurement models of ability and speed, respectively. On the second level, a joint distribution is assumed for the first-level person and item parameters.

Van der Linden has presented the SA model with a three-parameter logistic model (3PL) for response indicators:

$$p(u_{ij} = 1) = x_j + (1 - x_j)\frac{\exp(\nu_j\theta_i - b_j)}{1 + \exp(\nu_j\theta_i - b_j)}, \tag{1.11}$$

with θ_i denoting examinee i's ability and ν_j, b_j, and x_j denoting item i's discrimination, difficulty, and pseudo-guessing parameter, respectively.

For RTs t_{ij}, denoting the time examinee i required to generate an answer to item j, van der Linden (2007) suggested a lognormal model with separate person and item parameters. In this model, logarithmized RTs are assumed to follow a normal distribution governed by examinee speed τ_i and item time intensity β_j:

$$\ln(t_{ij}) \sim \mathcal{N}(\beta_j - \tau_i, \alpha_j^{-2}). \tag{1.12}$$

α_j represents the inverse of the RTs' standard deviation and can be interpreted as a time discrimination parameter. That is, the larger α_j, the larger the proportion of the RT variance that stems from differences in speed across examinees.

On the second level, joint multivariate normal distributions of person and item parameters are specified. The joint distribution of person parameters is assumed to be multivariate normal with mean vector

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\tau) \tag{1.13}$$

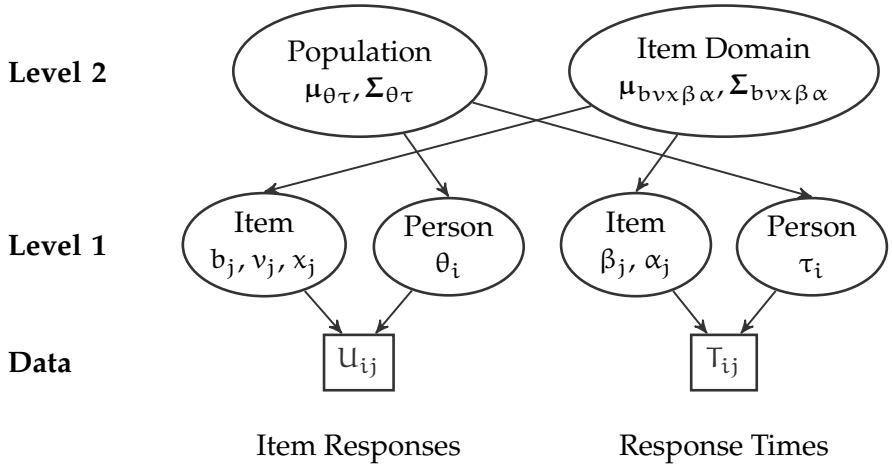

Figure 1.3. Hierarchical framework for the joint modeling of speed and accuracy by van der Linden (2007).

and covariance matrix

$$\Sigma_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{1.14}$$

Analogously, the joint distribution of item parameters is assumed to be multivariate normal with mean vector

$$\mu_{\mathcal{I}} = (\mu_b, \mu_\nu, \mu_x, \mu_\beta, \mu_\alpha) \tag{1.15}$$

and covariance matrix[1]

$$\Sigma_{\mathcal{I}} = \begin{pmatrix} \sigma_b^2 & \sigma_{b\nu} & \sigma_{bx} & \sigma_{b\beta} & \sigma_{b\alpha} \\ \sigma_{b\nu} & \sigma_\nu^2 & \sigma_{\nu x} & \sigma_{\nu\beta} & \sigma_{\nu\alpha} \\ \sigma_{bx} & \sigma_{\nu x} & \sigma_x^2 & \sigma_{x\beta} & \sigma_{x\alpha} \\ \sigma_{b\beta} & \sigma_{\nu\beta} & \sigma_{x\beta} & \sigma_\beta^2 & \sigma_{\beta\alpha} \\ \sigma_{b\alpha} & \sigma_{\nu\alpha} & \sigma_{x\alpha} & \sigma_{\beta\alpha} & \sigma_\alpha^2 \end{pmatrix}. \tag{1.16}$$

This yields the following likelihood

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{j=1}^{K} p(u_{ij}|b_j, \nu_j, x_j, \theta_i) f(t_{ij}|\beta_j, \tau_i, \alpha_j) g(\theta_i, \tau_i|\mu_{\mathcal{P}}, \Sigma_{\mathcal{P}}) h(b_j, \nu_j, x_j, \beta_j, \alpha_j|\mu_{\mathcal{I}}, \Sigma_{\mathcal{I}}). \tag{1.17}$$

The first two terms incorporate the assumption of conditional independence of response and RT indicators given the second-order variables of the model. Multivariate normal densities of the person and item parameters are denoted by $g(\theta_i, \tau_i|\mu_{\mathcal{P}}, \Sigma_{\mathcal{P}})$ and $h(b_j, \nu_j, x_j, \beta_j, \alpha_j|\mu_{\mathcal{I}}, \Sigma_{\mathcal{I}})$, respectively. This also illustrates that the hierarchical

[1]Note that van der Linden (2007) and Klein Entink, Fox, and van der Linden (2009) have suggested to consider transformations of ν, x, and α in the joint distribution of item parameters in order to improve the ranges and account for the skewness of typical empirical distributions of these parameters.

framework models the relation between speed and ability for a population of examinees as well as the relations between item parameters for the response and RT models for a population of items separately from the impact of these parameters on the responses and RTs on the level of individual examinees or items.

1.4.1 Advantages of Jointly Modeling Responses and Response Times

Considering speed when estimating ability has been shown to come with various advantages. From a measurement perspective, modeling the additional information contained in RTs has been shown to strengthen the measurement of ability in the sense that the SA model yields less biased and more reliable ability estimated as compared to methods that only consider responses (van der Linden et al., 2010). From a substantive perspective, the model allows assessing a) how much time examinees require to generate a response to a given item, b) how this relates to item difficulty as well as c) how, on an interindividual level, speed relates to ability.

In addition, Pohl and von Davier (2018) have argued that considering speed and ability jointly allows to disentangle different aspects contributing to examinee performance: When interacting with a test, examinees choose a certain level of speed based on such factors as their understanding of the test instructions, perception of the time limit, and style of work (van der Linden, 2011a). In the literature, the chosen level of speed is referred to as the examinee's effective speed during the test. This, in turn, determines the level of accuracy (or effective ability) examinees can show under the chosen speed level. Conversely, this implies that if examinees would want to put a stronger focus on accuracy, they could do so at the expense of working at a slower pace, resulting in higher effective ability and lower effective speed (van der Linden, 2009). This intraindividual relationship between effective speed and ability is referred to as speed-accuracy trade-off (see van der Linden, 2007, for an introduction). For the interpretation of results obtained from LSAs, examinees choosing to operate on different speed levels associated with different levels of effective ability could be problematic. Van der Linden (2007) has noted that a) test scores can only be assumed to reflect the rank order of the examinees' abilities when examinees choose the same level of effective speed and that b) otherwise examinees' test scores are "confounded with their decision on speed" (p. 21). Pohl and von Davier (2018) have argued that modeling and reporting on the displayed level of effective speed jointly with the displayed level of effective ability allows to explicate the chosen speed level determining the effective ability examinees operated on when generating responses and thus allows to disentangle different aspects involved in examinee performance. Since the intraindividual relationship between speed and accuracy is

unknown, estimating the effective ability for a certain level of chosen speed is not possible with assessment designs as currently implemented in LSAs.[2]

A further advantage of the SA model lies in its flexibility. By choosing different "plug-ins" for the component models (that is, the first-level models for responses and RTs and/or higher order models), the model can be adjusted to incorporate different assumptions on how examinees interact with the assessment. As such, the SA model has been subject to various extensions and modifications. Examples for these are models that allow for varying speed and accuracy throughout the test (Fox & Marianti, 2016; Molenaar et al., 2016), models that assume distributions for RTs different than lognormal (Klein Entink, van der Linden, & Fox, 2009), or aim at detecting and modeling differences in response processes, e.g., aberrant response behavior (van der Linden & Guo, 2008) such as rapid guessing behavior (Wang & Xu, 2015) or item preknowledge (Lee, 2018; Wang, Xu, Shang, & Kuncel, 2018).

1.5 Response Times and Missing Responses

Compared to the extensive body of research on utilizing RTs for understanding and modeling the processes underlying observed responses, approaches for utilizing RTs for handling missing responses are relatively sparse. The additional information on examinee behavior contained in RTs, however, might come especially valuable when investigating the occurrence of missing responses and time allocation strategies associated therewith. With respect to NRIs, cumulated RTs on the test level allow for assessing whether examinees a) allocated their time unfavorably and reached the time limit before reaching the end of the test or b) have quit the assessment before reaching the time limit or the end of the test. Previous work has acknowledged the role of speed for the occurrence of NRIs, however, only on a conceptual level. In addition, little attention has been paid to quitting as a potential mechanism underlying NRIs as well as to the potential of RTs for disentangling different mechanisms underlying NRIs.

When items are omitted, RTs allow for inferences on the underlying behavioral processes by, e.g., assessing whether examinees a) quickly skipped responses with RTs being far below under what was to be expected for examinees taking time to process and evaluate the item or b) whether RTs are similar to typical RTs associated with observed responses for the same item, suggesting that examinees seriously considered the item but, for whatever reason, decided to omit it (Weeks, von Davier, & Yamamoto, 2016). So far, however, only few approaches exist that leverage the

[2]See Goldhammer (2015) for suggestions on how to estimate intraindividual speed-accuracy trade-offs with experimental designs.

information contained in RTs for dealing with item omissions, the majority of which are rather heuristic. In what follows, previous work on the use of RTs for handling NRIs and item omissions will be reviewed and discussed.

1.5.1 Response Times and Not-Reached Items

In tests with time limits, examinees' decisions on speed determines the probability of examinees running out of time and showing NRIs (van der Linden, 2011a). As delineated above, RTs support inferences on the level of speed with which examinees generated responses, and, as such, contain valuable information on the mechanisms underlying NRIs. The relationship between speed and the occurrence of NRIs has been noted in previous research. Tijmstra and Bolsinova (2018) have argued to conceptualize the occurrence of NRIs as an indicator that examinees did not comply with test instructions. According to this line of argumentation, by setting time limits, test administrators aim at measuring the level of ability displayed at an optimal speed level that makes use of all the time available for completing the test. Examinees operating at a speed level insufficient to complete the test in time are thus not complying with test instructions and, by taking more time on the items they generate responses to, show a higher degree of effective ability than under the optimal speed level. The authors concluded that currently no methods for dealing with NRIs are available that allow for inferring on the level of ability that would have been observed under the optimal speed level.

Pohl and von Davier (2018), however, have pointed out that Tijmstra and Bolsinova's line of argumentation comes with rather strong assumptions. First, they neglected that examinees often tend to operate on speed levels higher than optimal. This is evident in PISA, where the majority of examinees spent far less time on the assessment than they had available (OECD, 2017). Second, Tijmstra and Bolsinova (2018) assumed NRIs to result from a lack of speed. However, LSAs such as PIAAC also report that examinees did not reach the end of the assessment due to quitting (OECD, 2013). Third, the objectives of major LSAs do not necessarily entail aiming at estimating ability at an optimal speed level. PIAAC, for instance, aims at measuring skills that are "necessary for fully integrating and participating in the labor market, education and training, and social and civic life" (OECD, 2013, p. 16). One could argue that choosing an appropriate speed level is an important aspect of this skill set and that, as such, differences in how examinees approach the assessment mirror important aspects of real-life behavior (see Pohl & von Davier, 2018).

1.5.2 Using Response Times for Coding Omissions

Omissions usually come with RTs considerably shorter and less variable than RTs associated with observed responses (Weeks et al., 2016), indicating different underlying processes that differ in their execution time. Current approaches leveraging the separation of RT distributions associated with omitted and observed responses predominantly use RTs for coding omissions prior to scaling. For doing so, they aim at identifying RT thresholds that separate RT distributions associated with omitted and observed responses. Such approaches either employ RTs for scoring omissions or for scoring both omissions and observed responses.

RESPONSE-TIME-BASED SCORING METHODS FOR OMISSIONS RT-based scoring methods for omissions assume omissions associated with RTs below a certain threshold to stem from processes different from those operating when examinees generate responses. Omissions associated with RTs exceeding the threshold are assumed to stem from processes similar to those underlying (incorrect) observed responses, since the examinee engaged sufficiently long with the item to generate a valid response, however decided not to. Accordingly, in such approaches, the former type of item omissions is treated as missing and the latter as wrong in all further analyses. Different methods exist for identifying the thresholds. PIAAC first introduced RT-based scoring methods for dealing with item omissions by scoring item omissions associated with RTs below and above five seconds as missing and incorrect, respectively (Yamamoto, Khorramdel, & von Davier, 2013). This rule has, however, been criticized to be rather arbitrary (Weeks et al., 2016). Recently, coding methods have been extended to allow for setting more empirical-based thresholds (Frey, Spoden, Goldhammer, & Wenzel, 2018; Weeks et al., 2016). Such methods usually report thresholds that are considerably higher than five seconds and vary across items.

RESPONSE-TIME-BASED SCORING METHODS FOR OMISSIONS AND RESPONSES In recently developed RT-based scoring methods, RTs are employed for scoring omissions and responses simultaneously, with the underlying rationale being that omissions and responses below a certain threshold can be assumed to stem from disengaged test-taking behavior. In their RT-based scoring framework for identifying disengaged test-taking behavior, Wise and Gao (2017) have defined disengaged test-taking behavior as "quickly proceeding through the test without applying [..] knowledge, skills, and abilities" (p. 348). To achieve that end, examinees are assumed to either a) rapidly guess on items with a multiple-choice format, b) provide perfunctory answers to items with an open-response format, or c) rapidly omit items and provide no response at all. This, in turn, implies a) that disengaged responses differ from

engaged responses in that they are not generated according to the level of examinee ability and, as such, show different measurement properties and b) that disengaged omissions and disengaged responses stem from the same type of test-taking behavior. In addition, in the approach by Wise and Gao (2017) disengaged behavior is assumed to be less time consuming than engaged behavior, resulting in shorter RTs. Wise and Gao (2017) have incorporated these assumptions in RT-based scoring methods by scoring omissions and responses associated with RTs below a certain threshold as missing. Omissions above the threshold are scored as incorrect and observed responses correspond to the observed value. Wise and Gao (2017) thereby have brought together research on employing RTs for scoring responses as either disengaged guesses or engaged responses with research employing RTs for scoring item omissions.

ASSUMPTIONS AND LIMITATIONS Although RT-based scoring approaches demonstrate the use of RTs for dealing with item omissions, they come with strong assumptions concerning the data-generating processes underlying item omissions and the associated RTs. First, RT distributions for omissions assumed to stem from processes different and similar to those operating when examinees generate (incorrect) responses are not allowed to overlap. The same is true for the RT distributions associated with engaged and disengaged responses in the framework by Wise and Gao (2017). This is a rather strong assumption and, as pointed out by Wise (2017), will inevitably result in misclassifications whenever RT distributions overlap. Second, RT-based scoring approaches for item omissions come with strong assumptions concerning the processes underlying item omissions associated with short and long RTs. While rapid item omissions are assumed to be ignorable, the probability to solve an omitted item which has been engaged with for some time is assumed to be zero (see Lord, 1983; Rose, 2013). These assumptions have been thoroughly discussed in Section 1.2.1. Likewise, in the approach for scoring both omissions and responses suggested by Wise and Gao (2017), ignoring item omissions and responses assumed to stem from disengaged behavior comes with the assumption that ability is unrelated to the processes underlying disengaged omissions and responses. Empirical evidence, however, strongly suggests that this assumption is violated, with ability and disengaged test-taking behavior frequently being found to be related (e.g., Boe, May, & Boruch, 2002; Braun, Kirsch, & Yamamoto, 2011; Goldhammer, Martens, Christoph, & Lüdtke, 2016; Wise, Pastor, & Kong, 2009).

1.6 Aims and Scope of the Present Work

In LSAs, missing responses due to item omissions and NRIs occur to a considerable degree. Understanding their occurrence is of utmost importance. First, researchers need to take some decision on how to handle missing responses in data analysis. For an adequate decision, a comprehensive understanding of the missingness mechanisms is key. Second, differences in missingness rates on both the individual and the country level indicate substantial differences in how examinees interact with assessments. Understanding these differences supports insights into, and allows considering differences in test-taking behavior when drawing inferences on examinee competencies.

Different approaches exist for handling missing responses, entailing different assumptions on the underlying mechanisms. Traditionally, approaches for handling missing responses have relied on information retrievable from paper-and-pencil-based assessment, i.e., responses, information on missingness, and covariates. With the rise of computerized testing and the opportunity to log additional information on how examinees interacted with the assessment, some approaches emerged that aim at leveraging the information contained in RTs for handling missing responses. Yet, these approaches are somewhat heuristic and rely on rather strict assumptions concerning the underlying missingness mechanisms.

Tackling the limitations of previous approaches for missing responses in LSAs, this work aims at utilizing the additional information contained in RTs for the objective of providing model-based approaches that support modeling as well as a nuanced understanding of the test-taking behavior underlying the occurrence of missing responses in LSAs. This is achieved by integrating research on the modeling of missing responses and on the modeling of RTs associated with observed responses. When doing so, separate frameworks are presented for modeling mechanisms underlying NRIs (Chapter 2 and 3) and omitted items (Chapter 4 and 5).

Chapter 2 delineates and assesses the potential of modeling speed jointly with ability as done in van der Linden's SA framework for modeling the mechanism underlying NRIs due to lack of speed. It is argued that by accounting for differences in speed, the SA model is well suited for modeling the mechanism underlying NRIs due to lack thereof. On the basis of theoretical considerations as well as a simulation study, it is shown that, compared to the manifest model-based approach for modeling NRIs by Rose et al. (2010), the SA model provides a closer description of the missingness mechanism since a) the SA model includes a direct measure of the mechanism underlying NRIs due to lack of speed, as compared to the number of NRIs considered in the manifest model-based approach as a rough proxy measure, b)

the SA model considers differences in working speed also for examinees who reached the end of the test and thus did not show NRIs, and c) the SA model can also deal with varying enforcement of time limits – given, that NRIs are the result of lack of speed. In an application of the approach to empirical data, Chapter 2 also illustrates the limitations of the SA model for modeling NRIs: Assessing distributions of total time spent on the assessment jointly with information on the number of reached items, it is shown that in the data set at hand some examinees with NRIs displayed test-level RTs close to the time limit, while the vast majority of examinees with NRIs exhibited test-level RTs far below the time limit due to quitting the assessment before reaching the time limit or the end of the test. This illustrates that within the same assessment two different mechanisms – lack of speed and quitting – can underlie the occurrence of NRIs.

Chapter 3 further builds on these results and extends the SA framework by additionally including a model for quitting behavior. The framework thus allows for disentangling and jointly modeling multiple mechanisms - lack of speed and quitting - underlying the occurrence of NRIs. In the proposed framework, quitting behavior is defined as stopping to work before the time limit is reached while there are still unanswered items. Based on the conclusions of Chapter 2, the SA framework is employed to model NRIs due to lack of speed. To model the quitting process, the framework utilizes the information contained in the number of reached items up to the point where the assessment has been quit. These are employed as indicators for a newly introduced latent variable, test endurance, that governs the item position at which examinees are most likely to quit the assessment and is modeled jointly with speed and ability. The framework assumes that every examinee will quit the assessment at some point, however, that quitting behavior is censored due to the fact that some examinees either reach the time limit or the end of the test before quitting. The censoring of quitting behavior is considered by drawing on survival modeling techniques. Considering test endurance jointly with speed and ability allows for disentangling, modeling, and assessing multiple mechanisms underlying NRIs simultaneously.

Chapter 4 provides a model-based framework that allows for assessing and modeling omission behavior and response behavior jointly while also allowing for a better understanding of time allocation strategies coming with the two types of behavior. This is achieved by extending the SA model by two additional latent variables: a) a latent omission propensity which accounts for the examinees' tendency to omit items as modeled in frameworks for modeling nonignorable item omissions (see Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999) and b) an additional speed factor based on RTs associated with item omissions, determining the speed

with which examinees omit items. The model provides, among other things, the possibility to a) model the missingness mechanism underlying omitted items when estimating ability, b) model timing data in the presence of missing responses, and c) get a better understanding of both the missingness mechanism underlying omissions and examinee pacing behavior through assessment of whether examinees employ different pacing strategies when generating responses or omitting items.

Chapter 5 builds on previous theoretical work relating item omissions to examinee disengagement and provides a model-based approach that allows for identifying and modeling examinee disengagement in terms of both omission and guessing behavior. It is thus built on the assumption that omissions stem from data-generating processes similar to those underlying disengaged guessing behavior and that both item omissions as well as disengaged guesses qualitatively differ from the processes operating when examinees provide responses by engagedly interacting with the assessment. The framework thereby brings together the approach presented in Chapter 4 with a) model-based approaches for identifying examinee disengagement in terms of rapid guesses and b) theoretical considerations underlying RT-based scoring methods that simultaneously score item omissions and observed responses. In the model, disengagement is identified on the item-by-examinee level by employing a mixture modeling approach that allows for different data-generating processes underlying item responses, omissions, and RTs associated with engaged and disengaged behavior. Item-by-examinee mixing proportions are modeled with a latent response framework as a function of examinee engagement and item engagement difficulty. This allows relating disengagement to ability and speed as well as identifying items that are likely to evoke disengaged test-taking behavior.

The approaches presented in Chapters 2 to 5 are tested and illustrated by a) evaluating their statistical performance under conditions typically encountered in LSAs by means of comprehensive simulation studies, b) illustrating their advances over previously developed approaches, and c) applying them to data from major LSAs, thereby illustrating their potential for understanding examinee test-taking behavior in general and behavior underlying the occurrence of missing responses in particular.

Finally, Chapter 6 discusses the potential of the developed frameworks for understanding results obtained from LSAs and derives practical implications for the analysis, interpretation and reporting of LSA data. Limitations of the proposed approaches are addressed and suggestions for future research are provided.

2

Using Response Times to Model Not-Reached Items due to Time Limits

This chapter is published as Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response time models to account for not-reached items. *Psychometrika*, *84*(3), 892–920. doi:10.1007/s11336-019-09669-2[1]

Missing values at the end of a test typically are the result of test takers running out of time and can as such be understood by studying test takers' working speed. As testing moves to computer-based assessment, response times become available allowing to simultaneously model speed and ability. Integrating research on response time modeling with research on modeling missing responses, we propose using response times to model missing values due to time limits. We identify similarities between approaches used to account for not-reached items (Rose et al., 2010) and the speed-accuracy (SA) model for joint modeling of effective speed and effective ability as proposed by van der Linden (2007). In a simulation, we show a) that the SA model can recover parameters in the presence of missing values due to time limits and b) that the response time model, using item-level timing information rather than a count of not-reached items, results in person parameter estimates that differ from missing data IRT models applied to not-reached items. We propose using the SA model to model the missing data process and to use both, ability and speed, to describe the performance of test takers. We illustrate the application of the model in an empirical analysis.

[1]The indices for persons and items have been adapted to the notation of the present work.

Large-scale assessments (LSAs) such as the Programme for International Student Assessment (PISA), the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the National Educational Panel Study (NEPS) aim at accurately measuring competencies such as reading comprehension or mathematical literacy. Competencies in these studies are assessed by tests containing a number of tasks that have to be completed in a certain time. Data collected in LSAs usually show a large proportion of missing responses due to the low-stakes nature of the assessment. Missing responses may be due to incomplete block assessment designs (planned missingness), due to item-level nonresponse (omitted responses), or items that were not reached (for example due to time limits). The amount of unplanned missing responses in LSAs is not negligible. In PISA 2006, for example, across all countries and all three domains (mathematics, reading, and science), an average of 10% of the items were omitted and 4% were not reached (OECD, 2009, p. 219–220). Even more important for country rankings, the amount of missing values largely varies across countries (from 1% in the Netherlands to 16% in Kyrgyzstan for omitted items and from 0.3% in Azerbaijan to 13% in Colombia for not-reached items).

This relatively large amount of missing responses needs to be dealt with in the psychometric analysis of test data. While not administered items can usually be considered as missing completely at random (MCAR) or missing at random (MAR), omitted and not-reached items (NRIs) are usually nonignorable and may lead to biased estimates of item and person parameters (see, e.g., Lord, 1983; Mislevy & Wu, 1996; Pohl et al., 2014). If not appropriately accounted for, estimates of group statistics can be biased by missing values as well, resulting in, for example, a different country ranking or biased regression coefficients when predicting test performance from explanatory variables (Köhler et al., 2017; Rose et al., 2010). In order to avoid biased item and person parameter estimates, the missing responses need to be appropriately dealt with.

Most of the models for missing values rely on information that is available in paper-and-pencil (P&P) tests, such as item responses, item nonresponses, and covariates. However, as testing moves to computer-based assessment (CBA), more information, in particular process and timing data, becomes available. In this article, we bring together research on missing values with research on process data, specifically response times (RTs), and aim to use RT information to model missing values and to describe the performance of test takers. Note that in the following we will focus on missing values due to not reaching the end of the test because of time limits; the proposed approaches are not necessarily suited for other types of missing values.

Note that the target ability that we aim at estimating is effective ability (and in addition effective speed) as defined by van der Linden (2007). Effective ability is the ability observed at the chosen (effective) speed level. Test takers may and usually do differ in the speed they chose for a given test. Although in substantive and methodological research, this is hardly ever discussed (see Kuhn & Ranger, 2015; Pohl & von Davier, 2018; Tijmstra & Bolsinova, 2018; van der Linden, 2007, for a few exceptions), this is what is done in almost all competence assessments in large-scale studies. In these assessments, test takers differ in their speed (even if RTs are not recorded) and no adjustment for speed is done.[2] While Tijmstra and Bolsinova (2018) suggest aiming at estimating optimal ability, that is, the ability observed when the test taker uses exactly the time given for answering all test items (i.e., optimal speed), Pohl and von Davier (2018) point out that optimal ability can only be estimated in very specific experimental settings that are hardly feasible in LSAs. They instead suggest estimating effective speed and effective ability as introduced by van der Linden (2007). By doing this, they explicate what has implicitly been modeled in many studies and with many modeling approaches before. It is also in line with the speed-accuracy model of van der Linden (2007). Pohl and von Davier (2018) argue that this approach a) allows to disentangle the different aspects of performance (a medium performance may be observed for persons with high ability and high speed as well as for persons with lower ability and lower speed), b) allows to estimate the same target ability for all groups of test takers, and c) better reflects real-life performance, as persons also need to choose their speed when solving real-world problems outside of testing situations. In this paper we will explicitly follow this approach, i.e., we will focus on effective speed and effective ability making no claims about optimal levels of these.

In the following, we will first introduce research on missing values. Then we will give an overview of research on models for RTs with a focus on the speed-accuracy model of van der Linden (2007). Finally, we will bring these two research lines together and show how RTs may be used for modeling missing data due to time limits. We will also discuss how RTs may be used for describing the performance of test takers in the presence of missing values.

[2] This is different in the study by Goldhammer (2015), who imposed item-level time limits to reduce the heterogeneity in RTs across persons. Note, however, that this only reduces heterogeneity in RTs across persons, but does not get rid of it. Furthermore, item-level time limits may result in guessing and item omission (Kuhn & Ranger, 2015; Pohl & von Davier, 2018).

2.1 Modeling Missing Values within IRT

2.1.1 Classical Approaches

There are different approaches to dealing with missing values (for an overview see, e.g., de Ayala, Plake, & Impara, 2001; Rose et al., 2010): 1) Missing responses may be ignored and, thus, treated as if they were not administered. This approach assumes that missing responses are MAR, given the observed responses on the items in the test (and other covariates in the background model). This approach is applied to missing values due to NRIs in NAEP (National Center for Education Statistics, 2009, May 13). 2) Missing responses may also be scored as incorrect responses, assuming that the subject did not know the answer. This is a deterministic scoring approach ignoring the fact that any respondent has a positive (even if low) probability to solve any item, given its trait level. Lord (1974) showed that the incorrect scoring method results in biased parameter estimates and proposed 3) to score missing responses as fractionally correct, for example, by scoring them according to the probability of guessing correctly. Fractional correct scoring is used for omitted multiple choice items in NAEP (National Center for Education Statistics, 2009, May 13). 4) In some educational studies (e.g., PISA until 2012, TIMSS, and PIRLS), a two-stage procedure for treating missing responses is used (see, e.g., OECD, 2009). For the estimation of item parameters, missing responses are ignored. The estimated item parameters are then used as fixed parameters for the estimation of person parameters where missing responses are scored as incorrect.[3]

Each of these approaches involves certain assumptions regarding the occurrence of missing responses. These assumptions do not necessarily hold in LSAs (see, e.g., de Ayala et al., 2001; Pohl et al., 2014; Rose et al., 2010). The approaches scoring missing values as incorrect do violate assumptions of IRT models (Lord, 1974; Rose, 2013). Also ignorability (i.e., MAR) of the missing values due to omitted items and NRIs usually does not hold. Different studies (e.g., Glas et al., 2015; Holman & Glas, 2005; Köhler et al., 2015a; Pohl et al., 2014; Rose et al., 2010) demonstrated that missing responses due to omission and test time limits often depend on the ability of the person and are thus nonignorable.

[3]This uses item parameters estimated with missing data ignored on data in which missing responses are coded as wrong. Hence, the item parameters do not fit the observed rates of wrong responses. This procedure was abandoned in PISA 2015 (OECD, 2017).

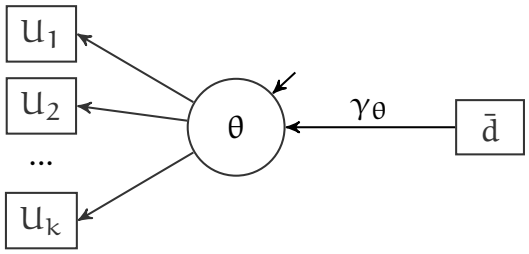

Figure 2.1. The manifest missing approach for not-reached items.

2.1.2 Model-Based Approaches for Nonignorable Missing Responses

Recently, model-based approaches for dealing with nonignorable missing data in IRT models have been developed. As these models may account for nonignorable missing data, which is most likely the missing data process present in cognitive test data, we focus on these types of models. In these approaches, the tendency to omit or not reach items is included in the model and accounted for in the estimation of the item and person parameters. The missing response tendency can be either included 1) via a latent missing propensity that is accounted for in a multidimensional IRT model (Glas et al., 2015; Holman & Glas, 2005; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999) or 2) by defining a manifest missing data indicator that is accounted for in a latent regression or multiple-group IRT model (manifest missing approach for not-reached items, mNRI, see, e.g., Rose et al., 2010). Rose (2013) showed that the mNRI model performs well and described the approach as sufficient for NRIs. In the following we will therefore focus on this approach. In the mNRI approach, there is a unidimensional IRT measurement model for the responses $\mathbf{U}_j$ to item j. Missing responses due to omissions and NRIs on response variables $\mathbf{U}_j$ of item j are treated as missing values in the measurement model of $\mathbf{U}_j$. A missing propensity is computed for each person as the relative number of NRIs $\bar{d}$. This missing propensity is included in the measurement model as an explanatory variable via latent regression, or alternatively, as a grouping variable used in a multiple-group IRT model (see Figure 2.1). As such, in the estimation of the item parameters and ability scores of the cognitive measurement model, the relative number of missing responses is controlled for. There is no restriction on the kind of measurement model for the latent ability.

Recently, researchers (Glas et al., 2015; Köhler et al., 2015a, 2015b; Moustaki & Knott, 2000) acknowledged that the missing response process may be even more complex and tried to more accurately model that mechanism by including further covariates explaining the missing data mechanism. Köhler et al. (2015a) found that additionally to ability, metacognitive competencies, reading speed, demographic variables (such as immigration background and schooltype) and interactions of

these variables are relevant predictors for the missing propensity. Glas et al. (2015) proposed a model-based approach for dealing with missing values that extends the latent variable approaches for missing values to incorporate further person characteristics.

2.1.3 The Impact of Response Time

Although the approaches accounting for nonignorable missing responses employ sophisticated modeling and show promising results, they rest on a number of implicit assumptions that may not necessarily hold in practice. The models and the simulation studies carried out in support of the models do not directly consider time on task for solving an item (e.g., Culbertson, April 2011; de Ayala et al., 2001; Finch, 2008; Holman & Glas, 2005; Köhler et al., 2017; Pohl et al., 2014; Rose et al., 2010). However, test takers who do not respond to all items have more time available to solve the items they choose to attempt, compared to test takers who attempt all items. In particular, test takers that do not reach a large number of items have often spent a much longer time on the (few) items they attempted. Studies using real data (e.g., Goldhammer & Kröhne, 2014; Semmes, Davison, & Close, 2011) show that the time available to solve an item does affect the probability of a correct response. This aspect has been, to our knowledge, neglected so far in research on missing response modeling.

2.2 Response Times Informing the Missing Response Process

Previous approaches for dealing with missing values rely solely on data of responses to test items (and on person characteristics). With the shift from P&P assessment to CBA, more information on how the test takers interact with items on a test becomes available. Specifically, CBAs typically collect data on the time test takers use to respond to each item. This information may be valuable for evaluating and modeling missing values in cognitive tests. RTs may provide information about the time allocation strategies of test takers. Moreover, process data do not only provide information on how much time it took to respond to an item, but also on how much time a person has spent on an item even if the person finally chooses not to respond (nonresponse time). The total time spent on an item (the time point when the item is first displayed to the time point when the respondent moves to the next item), no matter whether a response was produced or not, may help determine whether test takers indeed engaged in solving the item. Very short total time on an item may indicate that the test taker did not make an effort to solve the item, while longer total

times make it appear more likely that a person did attempt to solve the item (Weeks et al., 2016).

Operationally, LSAs such as PISA and TIMSS have not used RTs simply due to the fact that P&P assessments administered to groups of students within schools do not allow for accurate measures. The OECD Programme for International Assessment of Adult Competencies (PIAAC) was one of the first international assessment programme that was fully computer-based, in the sense that all test takers except those without sufficient computer experience were tested using laptop computers. The database for this assessment includes timing data and is publicly available. The analysis of RTs and missingness originated in this assessment with a more heuristic rule, based on the observation that missing data appeared to be associated with, on average, much shorter time measures (e.g., Yamamoto et al., 2013). That means that missing data with a nonresponse time of below a certain threshold was considered to be based on insufficient engagement with a task, while nonresponse that was associated with times usually observed together with incorrect or correct responses was considered as an indicator of a lack of skills. In other words, rapid skipping to the next item was not penalized, and considered a missing response, while respondents who did skip, but after a longer time had passed, would be assigned an incorrect response. This is consistent with findings about rapid guessing in assessments where test takers may feel forced to respond (Wise & DeMars, 2005) based on the high-stakes nature of the assessment. Using data from a large-scale NAEP computer-based field study, Lee and Jia (2014) found that rapid responses have no statistical association with the ability estimated based on responses given when sufficient time has passed. In PIAAC, researchers acknowledge the potential of RT for scoring missing values, however, their approach so far is based on heuristic rules. A sophisticated model that incorporates RT for modeling missing values would strengthen the existing approach.

In PIAAC, the threshold of considering an omitted item as a missing rather than an incorrect response is currently set to 5s. As research by Weeks et al. (2016) suggests, this rule may establish a lower bound, but item-dependent variability appears to exist. The authors found that RTs vary by item and argue that item-specific thresholds should be chosen. They furthermore showed that a 5s threshold may be too low and that – depending on the expected probability level of a response – median RT thresholds vary from 7 (expected probability of .50) to 41s (expected probability of .90).

IRT models have been proposed that utilize RTs as an additional source of information. In these approaches, RTs have mainly been used to model differential speededness of the test. Within this line of research, models have been developed

that incorporate RTs in the scaling model to account for differential speed of persons. While these models are quite elaborate, they have neither considered missing values in item responses, nor have they been used to model missing values in item responses. In the following, we will review these models and derive how they may be used for accounting for missing values.

2.3 Response Time Modeling within IRT

There are different kinds of models that incorporate RTs in the scaling of response data. These either incorporate RTs into the response model, incorporate responses into an RT model, or simultaneously model RTs and responses. An overview of these models is given in Schnipke and Scrams (2002) or Lee and Chen (2011). Note that none of the RT models explicitly deals with missing values. For our work, we focus on the third class of models, in which RTs and responses are simultaneously modeled. The simultaneous modeling allows depicting the different aspects of testing (ability and speed) separately, but in a combined model. We specifically focus on the speed-accuracy (SA) model proposed by van der Linden (2007). In the following, we describe this approach and discuss its potential utility to model missing values due to time limits in cognitive tests.

2.3.1 Hierarchical Speed-Accuracy (SA) Model

Van der Linden (2007) notes that, even when accounting for random error, test scores do not automatically reflect the rank order of the test takers' abilities. They do so only when test takers operate at the same speed. Otherwise test takers' scores are "confounded with their decision on speed" (p. 21). The approach of van der Linden clearly distinguishes between ability and speed and recognizes that different persons may choose different speed levels when working on a test. In the model, effective abilities at the chosen (effective) speed of the test takers are estimated. Van der Linden postulates different characteristics as the basis for his model. First, he proposes that RTs on test items should be treated as realizations of random variables. Second, van der Linden notes that the probability distribution of RTs is different from the distribution of the response variable, but related. Third, RT and speed are not equivalent. Instead, similar to the definition of speed in the natural sciences, speed is defined as the rate of change of some measure with respect to time. As a consequence, RT models with speed as a person parameter should also have an item parameter to quantify varying levels of time intensity. Fourth, speed and ability may be related. Van der Linden (2007) proposed a hierarchical model that incorporates a structure for simultaneous modeling item responses and RTs. He assumes that

response and RT distributions are determined by distinct parameters. At the lower level, the measurement models for item responses and RTs are specified, while at the higher level, the joint distribution of person parameters and the joint distribution of item parameters are specified. For the item responses $\mathbf{U}_j$ at the lower level, van der Linden assumes a 3PL model and models the probability of success of an item as

$$p_j(\theta_i) = c_j + (1 - c_j)\psi[a_j(\theta_i - b_i)] \tag{2.1}$$

with $\psi(\cdot)$ being the logistic function, θ_i being the ability parameter of test taker i, and a_j, b_j, and c_j being the discrimination, difficulty, and guessing parameters for item j, respectively.

For the RT T_{ij} for each item j and person i the model postulates a lognormal distribution:

$$\ln T_{ij} = \beta_j - \tau_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \alpha_j^{-2}), \tag{2.2}$$

where τ_i denotes the speed at which person i operates, β_j denotes the time intensity of the item, and α_j denotes the reciprocal of the standard deviation of the RTs on item j and can be interpreted as a time discrimination parameter. The appropriateness of a lognormal model for RTs has been investigated by van der Linden (2006) and Schnipke and Scrams (1997).[4]

At the higher level, van der Linden postulates parametric distributions for the item as well as for the person parameters of the two lower level models. For the person parameters, he assumes a multivariate normal distribution of the ability and speed variables. For the item parameters, he assumes a multivariate normal distribution for all item parameters in the response model and the RT model (i.e., for a_j, b_j, c_j, α_j, and β_j). The model is depicted in Figure 2.2.

Conditionally on ability and speed, the model assumes independence between responses to different items, independence between RTs on different items, and independence between responses and RTs on the same items. Thus, it is assumed that persons operate at constant ability and speed across the test. There may, however, be a dependency of accuracy and speed across persons. This is implemented at the higher level by means of a joint distribution for these random effects allowing for a correlation between speed and ability.

The model of van der Linden has been the basis for further model developments (e.g., Bolsinova, de Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2016; Fox & Marianti, 2016; Meng, Tao, & Chang, 2015; Molenaar et al., 2016; Molenaar, Tuerlinckx, & van der Maas, 2015; Ranger & Kuhn, 2012a; Ranger & Ortner, 2012;

[4]See Equation 4 in their paper.

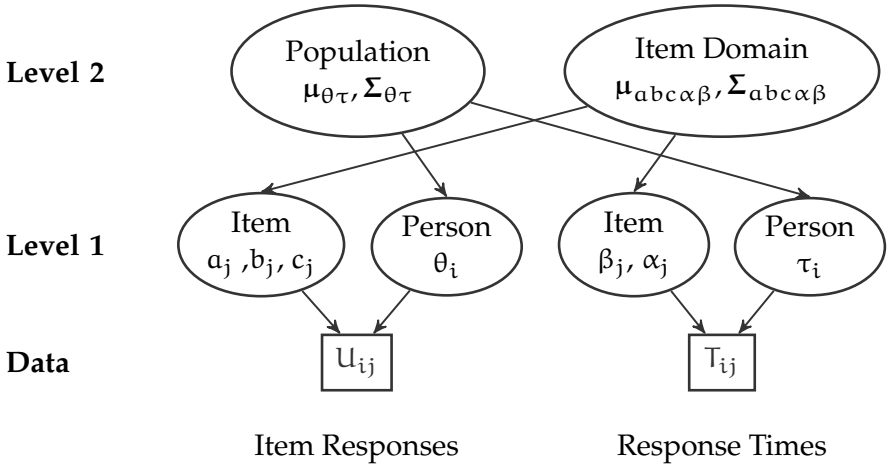

Figure 2.2. Graphical illustration of the hierarchical model of van der Linden (2007).

van der Linden & Glas, 2010) and substantive research (e.g., van der Linden, 2008; van der Linden & Guo, 2008).

Van der Linden (2007) only mentions in passing the issue of missing values: He notes that in his model "both speed and power aspects [are] captured by the variables T_{ij} (or D_{ij}) and U_{ij}, respectively" (p. 16), where T_{ij} denotes the RT, D_{ij} the missing indicator variable which shows which items the test taker completed, and U_{ij} the response variable. However, he does not further elaborate the role of missing values in his model.

2.4 Objectives

Missing values occur in cognitive tests and may have a systematic impact on conclusions drawn from cognitive test data. Different approaches for dealing with missing data exist. These rely on information about item responses, missing values, and covariates and try to adequately model the missing data process using this information. While models for RT exist, they have not, yet, been used for modeling missing values. We aim at filling this gap by bringing together research on missing values with research on RTs. Specifically, we use the SA model of van der Linden (2007) to account for missing responses due to time limits. We will show the usefulness of the model, discuss implications for the evaluation of the performance of test takers, and give an outlook to further model extensions

USING THE SPEED-ACCURACY MODEL FOR MODELING MISSING RESPONSES DUE TO TIME LIMITS One may understand the propensity of not reaching items as a measure of working speed. Then, the model of van der Linden is closely related to

the mNRI model for NRIs (Rose et al., 2010, Figure 2.1). In both models, ability and a measure of speed are considered. Both models include two variables for speed and accuracy simultaneously but separately, and the relationship of ability and speed (or the tendency to [not] reach items) is estimated. Both models make the assumption that ability and speed/missing propensity are stationary within the test. They are also similar in the (implicit or explicit) assumption of independence of responses and RTs or missing indicators given ability and speed. Note that in contrast to Rose et al. (2010), van der Linden (2007) also estimates a measurement model for speed as well as the joint distribution of the item parameters of the response and the RT model.

While in the model by Rose et al. (2010) speed is indicated by the number of NRIs, indicators for speed in van der Linden's model are the RTs per item. The number of NRIs can be seen as a rough proxy for RT at the level of the whole test. Thus, the information contained in the number of NRIs is also contained in the RTs. The RTs provide even more detailed information, that is, information on how much time a person has spent on each item. Thus, the model by van der Linden may be able to account for nonignorable missing values due to time limits.

INVESTIGATING THE PERFORMANCE OF THE MODEL AND COMPARING IT TO PREVIOUS MODELS In this paper, we investigate whether the SA model of van der Linden may indeed be sufficient to account for missing values due to not reaching the end of the test because of time limits. We investigate the performance of the SA model in comparison to the model by Rose et al. for accounting for missing values. If the SA model proves suitable, this approach may not only help to model missing values due to time limits, but may also provide information about the missing response process. Furthermore, it may help bringing together the research traditions of modeling missing values and those of modeling RTs.

2.5 Method

We conducted a simulation study in which item responses and missing values were generated following the SA model. For data analyses, we applied the SA model as well as the mNRI model by Rose et al. (2010). Note that as the SA model is used as the data-generating model, it is the true model with respect to comparisons of models made in subsequent analyses. We decided to use the SA model as the data-generating model for two reasons: First, the SA model is the more informative model from which data of both approaches may be generated, i.e., it is not possible to generate RTs from the number of NRIs alone without further assumptions. Second, and more important, missing responses due to time limits of the test are thought to be determined by the total of the times taken for each item. That is why the SA

model will allow incorporating a theoretically sound missing data mechanism. As a consequence of this simulation design the SA model will fit the generated data better than the mNRI model. However, we do not aim at such a comparison, but rather at a) evaluating whether the SA model is able to correctly estimate ability parameters in the presence of nonignorable missing data and b) investigating whether the SA and the mNRI model result in similar parameter estimates, in particular, to what extent ability estimates agree between these approaches.

2.5.1 Data Generation

For data generation, we chose parameters that represent typical low-stakes LSAs. Employing the SA model as the data-generating model, we generated data for $N = 1000$ persons responding to $K = 30$ items. For setting the parameters of the SA model, we relied on empirical results from the application of the SA model to empirical data, while the applications were not specifically considering missing values (e.g. Klein Entink, Fox, & van der Linden, 2009; van der Linden, 2007; van der Linden, Breithaupt, Chuah, & Zhang, 2007; van der Linden & Guo, 2008; van der Linden, Scrams, & Schnipke, 1999).

For the person parameters, θ and τ, we chose the following settings: $(\theta, \tau) \sim \mathcal{MVN}(\boldsymbol{\mu}_\mathcal{P}, \boldsymbol{\Sigma}_\mathcal{P})$ with $\boldsymbol{\mu}_\mathcal{P} = (0, -3.50)$ and $\boldsymbol{\Sigma}_\mathcal{P} = \begin{pmatrix} 1 & cov(\theta,\tau) \\ cov(\theta,\tau) & 0.25 \end{pmatrix}$. This corresponds to findings from empirical data, for example, in van der Linden et al. (1999). The correlation between the person parameters $cor(\theta, \tau)$ varies a lot in empirical data, ranging from $cor(\theta, \tau) = .30$ (van der Linden, 2007; van der Linden et al., 2007) to $cor(\theta, \tau) = .04$ (van der Linden et al., 1999) to negative values down to $cor(\theta, \tau) = -.76$ (Klein Entink, Fox, & van der Linden, 2009). We reflected this range of results by choosing different levels of correlations in our simulation, that is, $cor(\theta, \tau) = (-.50, .0, .50)$.

As the measurement model for item responses we chose the Rasch model, as it is used, for example in PISA until 2012, or in NEPS. Hence, only item difficulties b_j need to be estimated. For the measurement model of RTs, we fixed the discrimination $\alpha_j = \alpha = 1.875$ to be the same across all items. The value 1.875 was chosen in accordance with empirical results of van der Linden (2006). For the remaining item parameters, b and β, we assumed a multivariate normal distribution: $(b, \beta) \sim \mathcal{MVN}(\boldsymbol{\mu}_\mathcal{I}, \boldsymbol{\Sigma}_\mathcal{I})$ with $\boldsymbol{\mu}_\mathcal{I} = (0, 0)$ and $\boldsymbol{\Sigma}_\mathcal{I} = \begin{pmatrix} 1 & cov(b,\beta) \\ cov(b,\beta) & 0.14 \end{pmatrix}$. In empirical data, correlations $cor(b, \beta)$ between difficulty and time intensity of items have been found to vary between $cor(b, \beta) = .30$ (van der Linden, 2007) and $cor(b, \beta) = .51$ (Klein Entink, Fox, & van der Linden, 2009) and can even be as high as $cor(b, \beta) = .65$ (van der Linden et al., 1999). We

reflected this range of correlations and added a zero correlation as a reference in our simulation design, resulting in simulated correlations of $cor(b, \beta) = (0, .60)$.

We combined each of the two item parameter correlations with each of the three person parameter correlations, resulting in six simulation conditions. For each condition, we generated 100 replicate datasets. The person and item parameters were fixed and used in all 100 replications. We used the formulas in Equations 2.1 and 2.2 to generate item responses and RTs for each replication in each condition based on the generated item and person parameters. This resulted in 9 x 100 = 900 complete data sets without any missing values. The median RT of the items ranged from 14.17 to 71.58s. Within this data generation, no time limits were assumed.

We then considered a test setting in which the time limit for the test is 30 min. This corresponds to the usual time limit of test forms in LSAs (e.g., in NEPS, test forms ranging from about 24 to 36 items are presented with a 30 min time limit, see e.g., Duchhardt and Gerdes, 2012; Pohl, Haberkorn, Hardt, and Wiegand, 2012; Senkbeil, Ihme, and Adrian, 2014). We induced missing values based on the cumulative RT of the items. The items were assumed to be in the same order for every person (e.g., as in NEPS, Pohl et al., 2012) and the RTs were cumulated across the position of the items in the test. All items with a cumulated RT exceeding the time limit were assumed to be not reached and hence responses to these items were coded as missing. This resulted in incomplete data sets with 4.73 to 5.71% of missing values.

We subsequently assessed the effects of sample size, number of items, as well as rate of NRIs. To do so, we chose the condition with $cor(\theta, \tau) = .50$ and $cor(b, \beta) = .60$, as this was one of the conditions with the most severe threat to parameter estimation. We varied the number of examinees (adding a condition of $N = 500$), the number of items (adding a condition of $K = 10$), and the rate of NRIs (adding a condition of 15%). We controlled the amount of NRIs by setting stricter time limits (1,200s for a missing rate of 15% under conditions with $K = 30$; 600 and 400s for missing rates of 5% and 15%, respectively, under conditions with $K = 10$). This resulted in seven additional conditions (two sample size conditions times two item number conditions times two rate of NRIs conditions minus the baseline condition) evaluated in additional analyses.

2.5.2 Data Analysis

We analyzed the generated data using the SA model as well as the mNRI model. First, in order to evaluate how the SA model deals with missing values in general we applied the SA model a) to the complete data (SAcomp) as well as b) to the incomplete data (SAinc) and compared the results. Second, in order to evaluate the

difference between the SA model and the mNRI model for dealing with missing values, we applied both to the incomplete data. In order to get comparable results, we estimated all models using Bayesian estimation with Gibbs sampling in JAGS (Plummer, 2003), making use of the rjags package (Plummer, 2016) in R version 3.3.2 (R Development Core Team, 2017). Missing values in JAGS are imputed based on the specified model. We used noninformative priors, keeping the priors for the same parameters constant in all models. The settings for priors and syntax for the analyses can be found in Appendix A.1.

For both, the SA model and the mNRI model, we used three chains and no thinning. For the SA model, we chose a total of 45,000 iterations per chain, with a burn-in of 5,000, yielding a total of 120,000 iterations for posterior analyses. For the mNRI model, we ran 15,000 iterations per chain using a burn-in period of 5,000; altogether 30,000 iterations were saved.

We evaluated convergence of the model via trace plots and the Gelman–Rubin Potential Scale Reduction Factor (PSRF, Gelman & Rubin, 1992). We checked autocorrelation by assessing plots of the autocorrelation function along with the effective sample size (ESS as described in, e.g., Drummond, Nicholls, Rodrigo, & Solomon, 2002). For evaluating the performances of the models in retrieving accurate parameter estimates, we examined the estimates of person ability and speed, item parameters, as well as the correlation between ability and speed and between item difficulty and time intensity.

2.6 Results

In the following, we will first present the results for conditions with $N = 1000$, $K = 30$, and a rate of NRIs of 5%. We will then show the results of the effects of sample size, number of items, and rate of NRIs.

2.6.1 Convergence and Efficiency

Across all conditions and models, no convergence problems were encountered. All trace plots indicated good mixing of the chains and convergence. For both models, PSRF values remained far below 1.05 for all parameters and thus were, in line with Gelman and Shirley (2011), considered acceptable.

Autocorrelation in the MCMC chains varied largely across parameters when data were analyzed with the SA model. ESS ranged from 473 and 526 (for a time intensity parameter) to 26,691 and 26,694 (for an item parameter variance estimate) when the SA model was applied to complete and incomplete data sets, respectively. This

indicates that parameter space had been sufficiently explored to assess posterior means and standard deviations (Kruschke, 2014, p.184). For the mNRI model, ESS ranged from 5,587 (for a difficulty parameter) to 10,653 (for an ability estimate).

2.6.2 Performance of the Speed-Accuracy Model for Complete Data

As expected, the SA model for complete data yielded unbiased group-level parameter estimates across all conditions with bias for all parameter types remaining below 10%. Figure 2.3 shows the difference of the individual ability estimates averaged across all replications in the SA model for complete data (SAcomp) from the true parameters for the nine conditions.

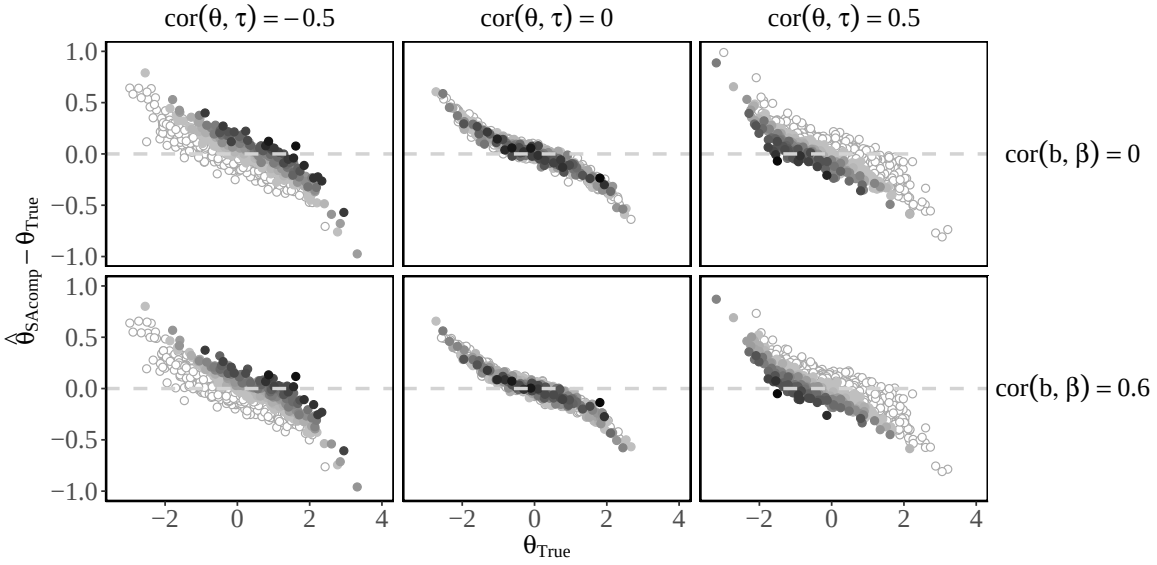

Figure 2.3. Difference in ability estimates using the SA model for complete data compared to the true ability values as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

There is a noticeable shrinkage effect (see also Fox, 2010) in all conditions. Although no missing values were induced for these analyses, in the graphs we marked the average number of NRIs for each simulee to be induced later. The estimates do not systematically differ across conditions with different item parameter correlations. There are, however, differences across conditions with different person parameter correlations. This is due to the adjustment that is made by incorporating speed in the model. For a person parameter correlation of zero, there is no systematic difference in person parameter estimation of persons with different numbers of missing values. For a correlation unequal to zero, the relationship of the difference in person parameter estimation with the speed variable becomes evident. A negative correlation

means that more proficient test takers work slower and therefore have a tendency to produce more missing values. Also, for the same ability level, slower test takers are those that have more missing values (see Figures 2.3 and 2.4) and by the negative correlation between ability and speed this results in a higher ability estimate. Note that in the SA model for complete data, there are no missing values; Figures 2.3 and 2.4 only show the missing values that will be induced. As can be seen in Figure 2.4, there is no systematic bias in ability estimation for different values of true speed.

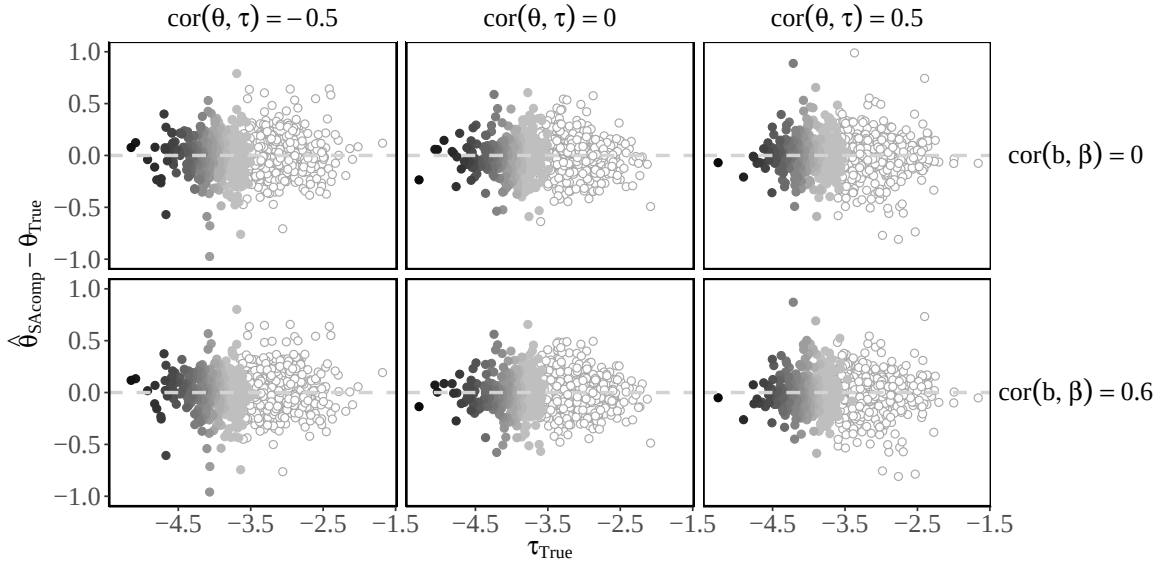

Figure 2.4. Difference in ability estimates using the SA model for complete data compared to the true ability values as a function of true speed. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

We found similar effects for the estimation of speed. There is a shrinkage effect in the estimation of speed due to unreliability and there is an effect of the correlation between ability and speed (Figures A.3 and A.4 in Appendix A.2). Summarizing the results, it can be concluded that, in the case of complete data, the SA model was able to adequately recover the true parameters. That is, the complete data model may serve as a reference for comparison for subsequent analyses.

2.6.3 Performance of the Speed-Accuracy Model to Deal with Not-Reached Items

There is no systematic bias in group-level parameter estimates of the SA model for incomplete data, bias for all parameter types remained below 10%. In the following, we compare the results of the SA model applied to incomplete and complete data, respectively. By doing this, we control the shrinkage effect due to using only 30 items, as this number is the same in both cases. Differences between the results of the two

analyses can, thus, be attributed to the existence of missing values. Figure 2.5 shows the difference in ability estimates averaged across all replications between the SA model for incomplete (SAinc) and the SA model for complete data (SAcomp) as a function of true ability and the number of missing values in all conditions.

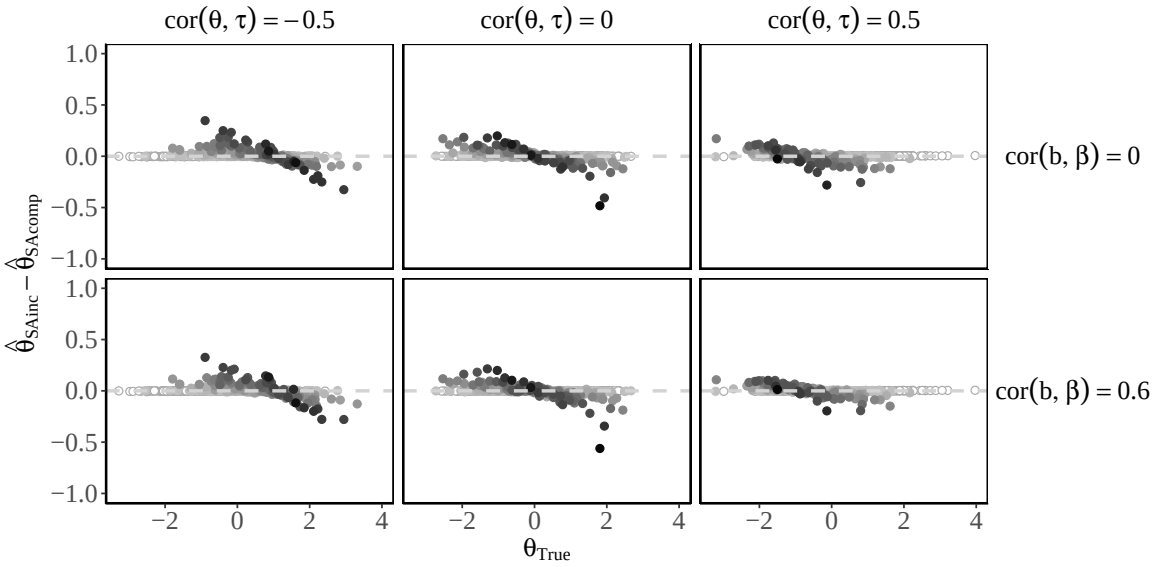

Figure 2.5. Difference in ability estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true ability. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

The results show that for simulees with no missing values, there is no difference in parameter estimates. There is a difference in parameter estimates for simulees with missing values; with greater differences being observed for simulees with a higher number of NRIs. This difference may be explained by a shrinkage effect that is due to the fact that for these respondents, information from fewer than 30 items is available. This is reflected in the posterior standard deviation for simulees with a high number of NRIs: for persons with an average number of NRIs greater than 15, posterior standard deviations ranged from 0.36 to 0.89 – as compared to a range from 0.36 to 0.43 for the same person parameter estimates, however, estimated using complete data. The greater uncertainty of these estimates is associated with an increased shrinkage effect. Person parameters are thus estimated closer to the overall ability mean. This effect is even more obvious when plotting the difference in ability estimates as a function of speed (see Figure 2.6). Since simulees with a lower speed produce more missing values, fewer item responses are available, resulting in greater standard errors of ability estimates and a larger shrinkage effect. We found similar results for the estimation of speed (see Figures A.5 and A.6 in Appendix A.2)

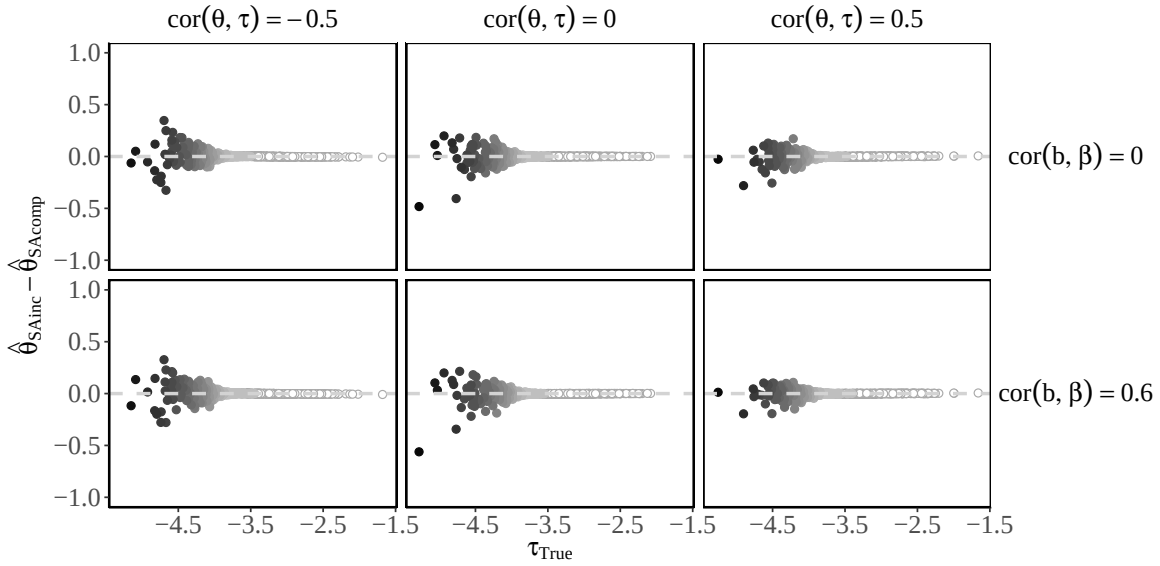

Figure 2.6. Difference in ability estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true speed. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

2.6.4 Illustrating the Shrinkage Effect due to Missing Values

In order to underpin the conclusion that the difference in parameter estimates between the SAcomp and SAinc model go back to shrinkage effects due to missing values, we re-ran the analyses on the performance of the SA model for incomplete data with missing values induced completely at random (SA MCAR). To do so, we used the complete data sets of 1000 examinees responding to 30 items and introduced missing values being MCAR. The number of missing values for each person was drawn from a discrete uniform distribution ranging from 0 to 25. The number of missing values for each person was fixed across replications; however, the specific items bearing missing values were chosen randomly for each replication. By doing so, we ensured that the average number of missing values for each person across replications displayed a range comparable to the range of the average number of NRIs in the main simulation. Figure 2.7 depicts the difference in ability estimates between using the SA model on a data set with missing values being MCAR and the SA model on the complete data set as a function of true ability and the number of missing values for the condition with $\mathrm{cor}(\theta, \tau) = .50$ and $\mathrm{cor}(b, \beta) = .60$. As can be seen, the resulting pattern resembles the pattern of differences between the SA incomplete and the SA complete model, where the difference increased as a function of the number of missing values. Since missing values for the SA MCAR model were induced completely at random, systematic bias in person parameter estimates can

be ruled out as an explanation for these differences; instead, the differences can be attributed to the shrinkage effect as a result of the reduced amount of information for individuals with fewer observed responses and a high number of missing responses. From these results we conclude that the SA model can account for missing values due to time limits. In the next step, we compare the SA model and the mNRI model for dealing with missing data.

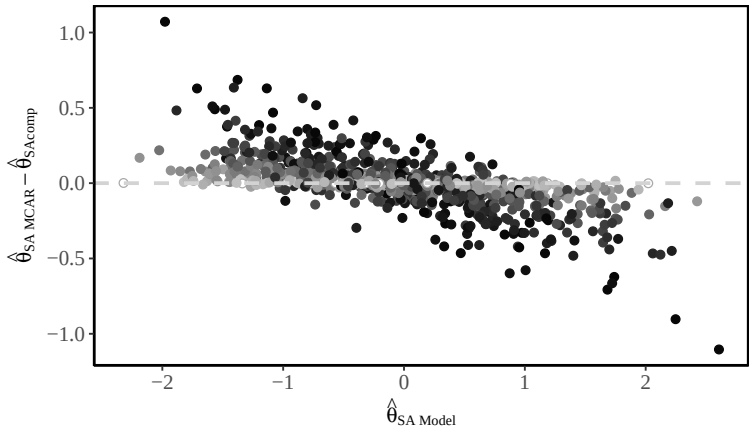

Figure 2.7. Difference in ability estimates between the SA model for incomplete data with missing values induced completely at random (SA MCAR) and the SA model for complete data (SAcomp) as a function of true ability for the condition with $\mathrm{cor}(\theta, \tau) = .50$ and $\mathrm{cor}(b, \beta) = .60$. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

2.6.5 Performance of the Manifest Missing Response Model for Incomplete Data

The simulation results show that bias in hyperparameters is comparable to the SA model. There are, however, differences in person parameter estimation. So far, in the analyses we have found a) a shrinkage effect due to using only 30 items and b) a shrinkage effect due to missing values. These are to be expected for this model as well, as it uses the same data as the SA model for incomplete data. In order to separate the shrinkage effects from differences between the models, we compared the mNRI model to the SA model for incomplete data (which was shown to appropriately recover the parameters of the model). Figure 2.8 and 2.9 shows the difference in ability estimates averaged across all replications between the mNRI model and the SA model for incomplete data. There is no impact of item parameter correlation. There is, however, one of person parameter correlation. For $\mathrm{cor}(\theta, \tau) = 0$, there is no difference between the two models. For a correlation of $\mathrm{cor}(\theta, \tau) \neq 0$, there are noticeable differences between the two models, which depend on the true ability and on the number of missing values.

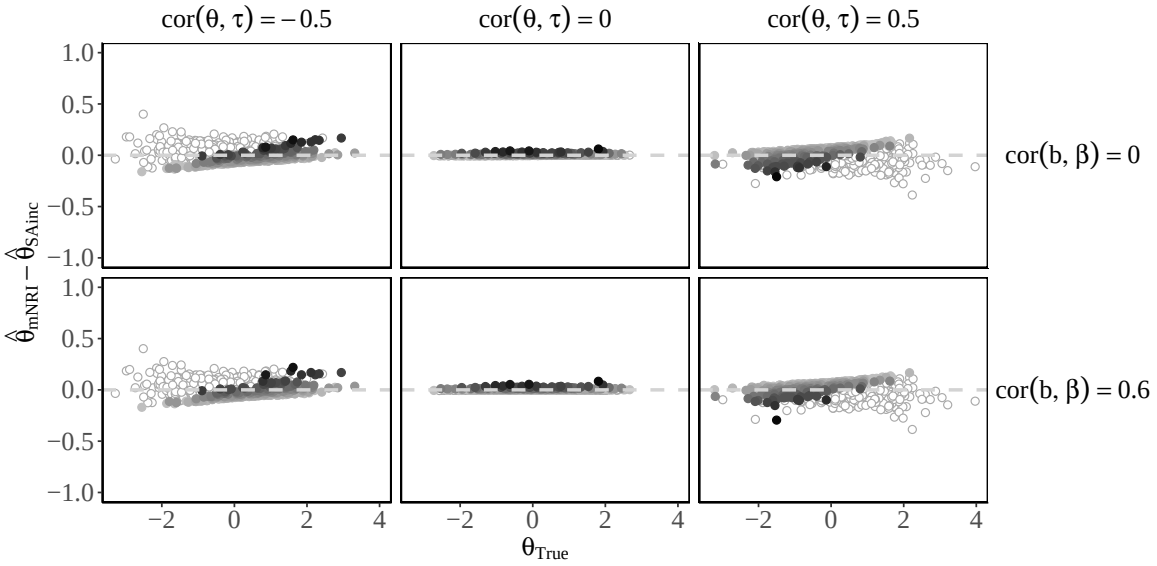

Figure 2.8. Difference in ability estimates from the mNRI model and the model for incomplete data (SAinc) as a function of ability. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

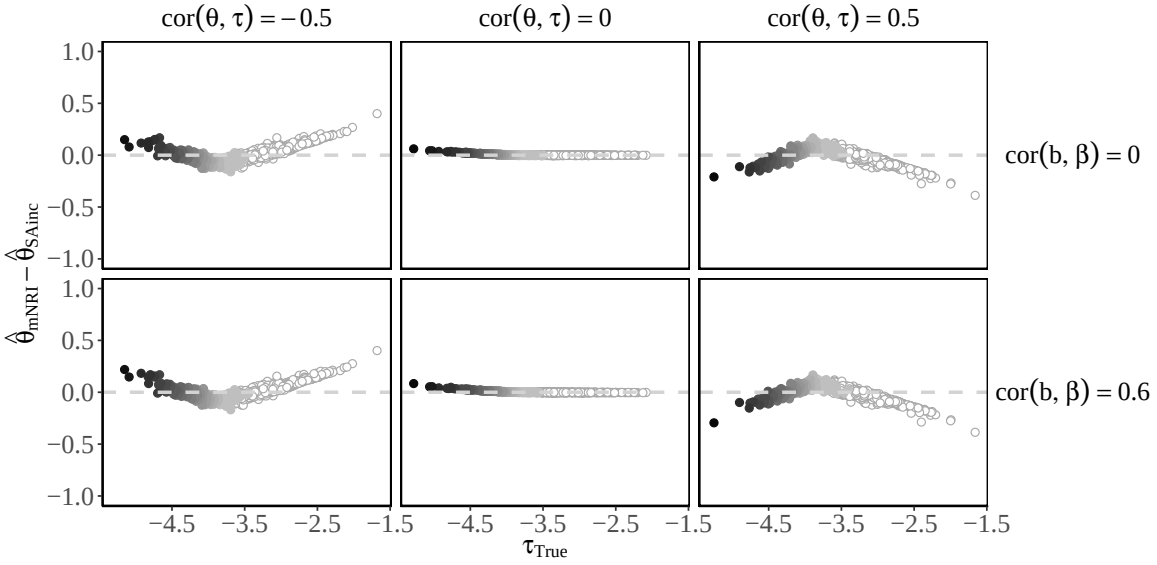

Figure 2.9. Difference in ability estimates between the mNRI model and the SA model for incomplete data (SAinc) as a function of true speed. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

There are two mechanisms at work here: First, as all persons with sufficient speed reach all items (i.e., have no missing values), there is a truncation of the distribution of speed: The manifest missing variable does not distinguish between persons who work with sufficient speed; all of these persons have zero NRIs. As such,

for $cor(\theta, \tau) = -.50$, the ability of persons with a high speed level is overestimated, while it is underestimated for $cor(\theta, \tau) = .50$. This is due to the adjustment made to ability due to working speed. As in the mNRI model persons with high speed have the same number of NRIs (i.e. zero), these persons are – in contrast to the data-generating model – treated the same. In the data-generating model, for $cor(\theta, \tau) = .50$ persons with high speed are those with high ability; in the mNRI model, this is not accounted for on the upper speed level.

Second, the mNRI model underestimates the association between ability and speed. The estimated correlation of the number of NRIs (as a proxy for speed) and ability was on average .32, .02 and -.33 in the condition with $cor(\theta, \tau) = -.50, 0$, and $.50$, respectively. This may be due to the fact that a) a manifest instead of a latent variable is used (see Pohl et al., 2014, for a similar result using manifest and latent missing propensity) and/or b) there is truncation and as such a variance reduction, and/or c) the number of NRIs is a nonlinear transformation of speed (see Figure 2.10) which may violate the linearity assumption of the relationship of ability and speed. As a consequence, the adjustment of ability estimates based on speed is different than in the SA model for incomplete data. In Figures 2.8 and 2.9, we do see that ability estimates of respondents with low speed (i.e., having many missing values) is overestimated (underestimated) in the condition of $cor(\theta, \tau) = -.50$ ($cor(\theta, \tau) = .50$).

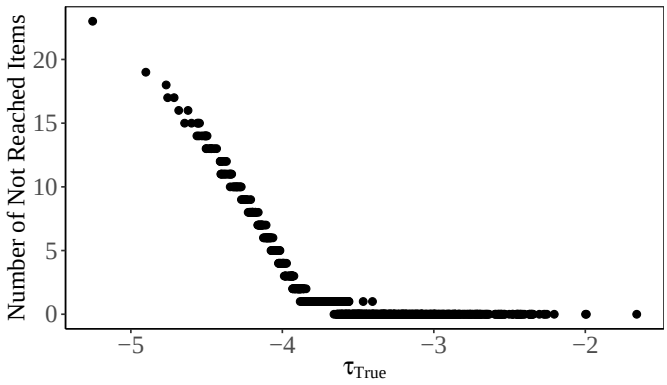

Figure 2.10. Number of not-reached items per person across as a function of true speed for the condition with $N = 1000$, $K = 30$, $cor(\theta, \tau) = .50$, $cor(b, \beta) = .60$, and a rate of not-reached items of 5%.

2.6.6 Additional Analyses

We conducted additional analyses to study the effects of sample size, number of items, as well as rate of NRIs on parameter estimation. Both models, the SA model for incomplete data and the mNRI model, converged across all conditions and replications. Again, no systematic bias was found in group-level parameter estimates of the SA model. Only for conditions with a small number of items ($K = 10$), item

parameter variances were estimated substantially higher than the respective true values (1.11 as compared to 1.00 and 0.25 as compared to 0.14 for $var(b)$ and $var(\beta)$, respectively). Comparable effects on item parameter variance estimates occurring for smaller number of items have been reported by Fox and Marianti (2016). Appendix A.3 shows the result of person parameter estimation using the SAcomp model as compared to the true ability parameters (Figure A.7) and using the SAinc model as compared to the SAcomp model (Figure A.8). There was no effect of sample size. As was to be expected, an increase in the amount of missing values resulted in an increased shrinkage effect as less information was available for persons with more missing values. There was also an effect of the number of items: There were larger shrinkage effects for conditions with $K = 10$, since under these conditions less information on the examinees' ability is available.

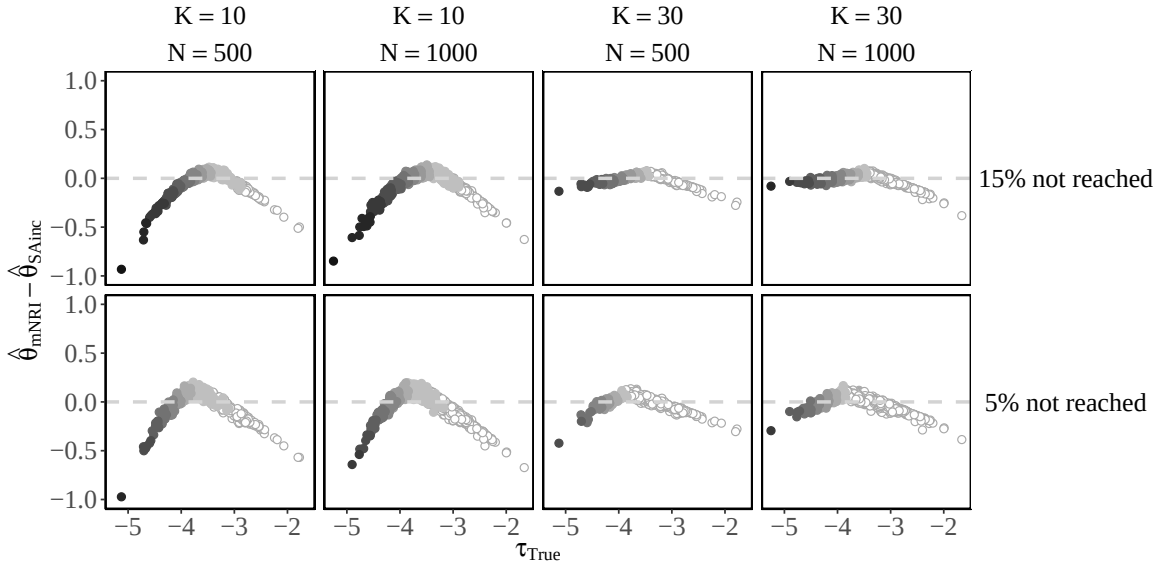

Figure 2.11. Difference in ability estimates between the mNRI model and the SA model for incomplete data (SAinc) as a function of true speed. White circles represent simulees without missing values and filled circles represent simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

Figure 2.11 shows the difference in ability estimates when using the mNRI approach as compared to the SA model for incomplete data. Again, there is no effect of sample size. There is, however, an effect of the number of items: with more items, there is more information available for estimating person ability. As a result, the impact of speed becomes smaller relative to the impact of item responses. There is also an effect of the amount of missing values, with more missing values resulting in smaller differences between the two approaches. This is an effect of an increase in discrimination regarding differences in speed and a reduction in the truncation effect

in the mNRI model. With an increasing rate of NRIs, the number of NRIs displays more variation and, as such, differences in speed are reflected better by the number of NRIs. As a consequence, the relationship between speed and ability can better be captured in the mNRI model by the relationship of the number of NRIs and ability. Whereas in conditions with a rate of 5% NRIs the average correlation between ability and the number of NRIs was as low as -.35 as compared to the generated correlation between ability and speed of $cor(\theta, \tau) = .50$, in conditions with a rate of 15% NRIs an average correlation of -.42 was estimated. [5]

2.7 Empirical Data Analysis

In order to evaluate the applicability of the approach, we analyzed data from the Canadian sample of PISA 2015 (OECD, 2017). We applied a) the mNRI model and b) the SA model to science cluster number 7 administered in the second position of the CBA. In total, analyses were based on $N = 840$ examinees responding to $K = 17$ items within a time limit of 1,800s. Six percent of the test takers did not reach all items. In total, the science cluster under consideration displayed a rate of NRIs of 2%. For reasons of simplicity, partial credit items were dichotomized and examinees with missing data other than NRIs were removed from the analyses. We analyzed the data employing the same MCMC setup as in the simulation study described above, saving 6,000 and 8,000 iterations as a sample of the posterior distribution for the mNRI model and the SA model, respectively. Convergence was assessed on the basis of trace plots as well as PSRF values of all parameters. Judging by these criteria, both models converged.

Figure 2.12 shows that not all test takers stopped working on the test when the test time limit of 1,800s was reached. This indicates that the time limit may not have been enforced rigorously. Most importantly, we see that only for persons with one or two NRIs, the time limit was reached. Almost all test takers with more than one NRI did not use the time they had. Thus, the test time limit seems to be not the only reason for NRIs in the test, but missing values at the end of the test may also occur due to quitting. Note that speed as estimated in the SA model and the number of NRIs correlate -.16. This reflects the results from the descriptive statistics, indicating that NRIs do not only occur due to low speed and reaching the time limit.

Figure 2.13 shows the difference in ability estimate between using the mNRI model and the SA model. For persons without missing values, we see a similar pattern as in the simulation study. As NRIs are a result of time limits for only very few test takers, for persons with missing values, this pattern deviates from the pattern

[5] These results refer to a condition with $N = 30$ items and $N = 1000$ persons.

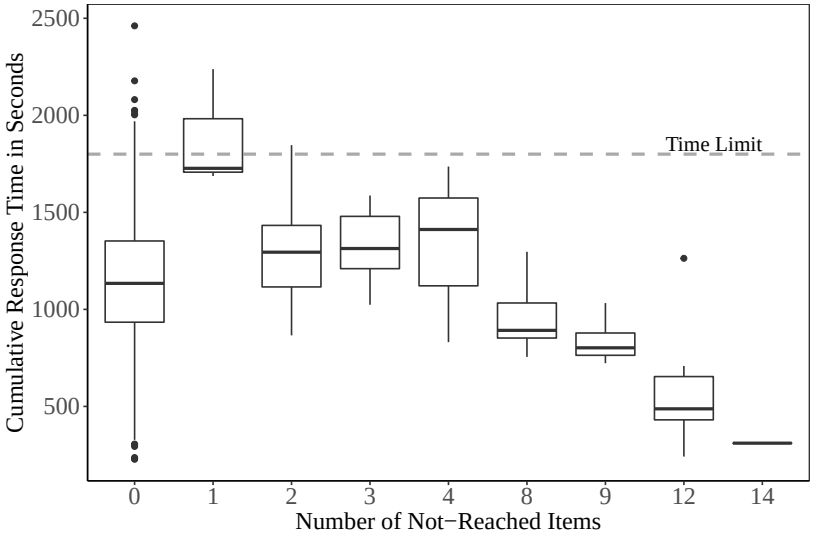

Figure 2.12. Cumulative response time distribution for different numbers of not-reached items.

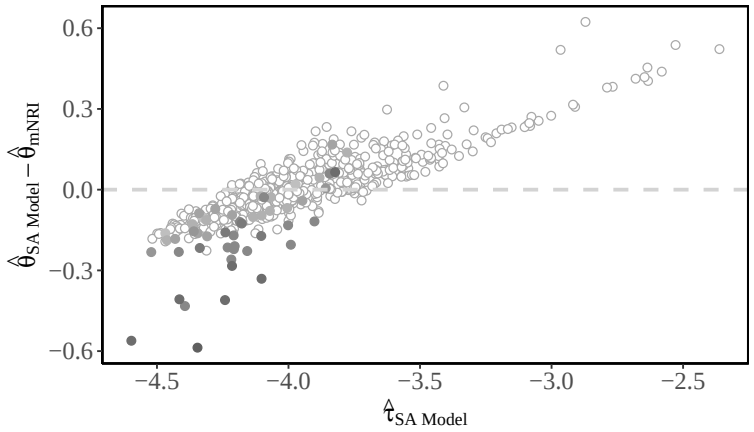

Figure 2.13. Difference in estimated ability scores between using the mNRI model and the SA model in the empirical application. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

in the simulation study (see, e.g., Figure 2.9). This can also be seen in the parameter estimates of the two models. Using the SA model, we estimated a correlation of -.40 between ability and speed. The correlation of ability and number of NRIs in the mNRI model was estimated to be -.01. Summarizing the results, the SA model can account for NRIs that occur due to reaching the time limits. However, NRIs did not only occur due to time limits in this data set.

2.8 Discussion

In this study, we integrated research on RT modeling with research on modeling missing responses. We proposed using the SA model to model and account for missing values due to time limits in the test. Our study showed that there are similarities between the mNRI model for NRIs (Glas et al., 2015; Rose et al., 2010) and the SA model of van der Linden (2007). In particular, we identified the potential of the SA model to account for missing values due to time limits on tests. In a simulation study that models this scenario, we showed a) that the SA model can recover parameters in the case of missing values due to time limits and b) that the SA model results in different person parameter estimates than the mNRI model. If missing values due to not reaching the end of the test occur because test takers work at different speed levels, then the SA model can describe the missing data process. In addition, the SA model incorporates differences in working speed also for those test takers who do reach the end of the test (i.e., who have no missing values).

Note that we explicitly aimed at estimating effective ability and effective speed. Even for test takers without missing values, using the SA model or even just a unidimensional model for responses, effective ability is estimated. The approach adopted here for NRIs due to time limits also estimates effective ability for test takers with missing values, resulting in the same target ability for all groups of test takers, those with and those without missing values.

Of course, we cannot determine with certainty whether the assumed missing data mechanism is the one at work leading to different numbers of NRIs for respondents with different working speeds. From a theoretical point of view, the mechanism seems quite plausible and in the real data analyses we found evidence in supporting of this mechanism for some test takers. While the mNRI model does not describe the mechanism of how missing values occur and includes the missing propensity for adjustment purposes only, the SA model describes a mechanism that explains missing values in terms of time spent on previous items. Empirical studies using response times do hint at the plausibility of such mechanisms (e.g., Goldhammer & Kröhne, 2014).

Note that in our simulation study, we only considered NRIs that occur due to time limits. We did not consider NRIs that occur because the test taker quits responding before reaching the last item. In the empirical analysis, we found evidence that this is another plausible mechanism in practice. While NRIs due to time limits can be accounted for by the SA model, early quitting behavior needs to be accounted for differently. RTs and other log data provide comprehensive information that may help to distinguish between different nonresponse mechanisms.

Moreover, the SA model proposed by van der Linden assumes stationarity of speed and is most appropriately applied to data stemming from tests with a generous time limit. This, however, is not necessarily the case in LSAs that administer tests to groups of students: Testing situations in which test takers encounter (tight) time restrictions and are either running out of time, or perceive it to be so, might lead some test takers to speed up towards the end of the test in order to finish the test within the allocated time. When working speed is used to account for missing values due to NRIs, the very fact that some test takers were not able to reach the end of the test is an indicator that testing time has been not sufficient for all participants. As a consequence, some test takers might have adjusted their working speed in order to reach the end of the test (e.g., Yamamoto & Everson, 1997). Under these conditions, the assumption of stationarity of speed is not plausible and it appears necessary to allow for within-person variation of speed (Fox & Marianti, 2016; Goegebeur, De Boeck, Wollack, & Cohen, 2008).

Just by the position of a missing response in the test, within the test (omitted) or at the end of the test (not reached), one cannot infer whether the item had really been attempted or not. It may well be that some omitted items within the test have not been attempted at all (resulting in low nonresponse time) or that NRIs at the end of the test have been attempted (resulting in higher nonresponse time for these items). So far, models for missing values have relied on the position of the missing items within the test for drawing inferences on whether the item had been attempted or not. In some LSAs (e.g., NAEP, Allen, Donoghue, & Schoeps, 2001), the treatment of missing values is even based on this distinction (omitted items are scored partially correct and NRIs are treated as if they were not administered). By using RTs, one could potentially infer better whether items with missing responses have been attempted or not, compared to making this determination just based on the position of the missing response within the test. How not attempted items within a test and different kinds of missing values can be evaluated with the help of RTs is one important and promising future research task. While Weeks et al. (2016) set out to explore how this can be achieved, their study remains mainly descriptive. A more model-based approach is needed that describes mathematically the interdependence between the time spent and the time remaining on the one hand and response vs. omission propensity on the other hand. The present study shows that the SA model proposed by van der Linden, together with some assumptions about how time on tasks and time limits relate to NRIs, can be used to model missing data by utilizing more information than the missingness indicators alone.

2.8.1 Implications for the Practice of Dealing with Not-Reached Items due to Time Limits

In low-stakes LSAs – as they are currently implemented – persons differ in their working speed. As a consequence, test takers are on a different position with regard to their speed-accuracy trade-off (van der Linden, 2007). As such, we do estimate the effective ability and effective speed, which differs between test takers. This is true no matter whether we actually assess RT or not, or whether persons reach the end of the test or not. Unless in very specific experimental settings (Goldhammer, 2015), which are however not feasible in LSAs, it is not possible to correct for chosen speed and to estimate optimal ability (i.e., the ability observed when speed is chosen so that the exact given testing time is used; not more or less). Thus, in line with van der Linden (2007), we argue for the estimation of effective ability, as this quantity can be estimated in the majority of testing situations. With the use of RTs and by modeling the association between speed and ability, we can describe these different aspects of performance.

We suggest describing the performance of groups of test takers (for example grouped by language, country, or school type) by all aspects of performance: ability and speed (and/or missing propensity in case of other reasons for missing values) and use all of these for evaluating the performance (see also Pohl & von Davier, 2018). This allows to develop a richer description of differences in performance and to disentangle the different constructs involved. For example, Cosgrove (2011) investigated the decrease in the PISA trend results of Ireland from 2000 to 2009 and found that students showed much larger amounts of missing values in later PISA assessments. They concluded that the decrease in PISA score may be a result of lack of motivation that led to more omitted responses. If the different aspects, that is, effective ability and speed (and/or possibly missing response propensity) would have been estimated and presented separately, these changes over time would have been more evident, and the apparent performance differences could have been understood in the light of other changes (see Sachse et al., 2019, for an investigation thereof). When comparing, for example, the performance of different countries in a cognitive domain, one may want to compare these on both effective ability and effective speed. For country rankings, policy makers are interested in the comparison of only one score for each cognitive domain. If a single score is of interest, we suggest using a composite score based on the estimated aspects of performance. Substantive researchers may then decide how to combine ability and speed estimates by developing a composite score that reflects the dimension they want to focus on most. One advantage of such an approach would be that this composition of a total

score would be the same for all test takers. Furthermore, the approach can also deal with varying time restrictions, as has been present in the PISA data. As the measure of speed does not dependent on the total time used, more or less rigorous enforcement of time limits may be accounted for.

This is different in the approach of scoring missing values as incorrect. Scoring missing values due to not reaching the end of the test as incorrect is also a constructed measure incorporating accuracy of responses and speed into a single score. However, a) the different aspects of performance cannot be disentangled. As such, subgroups with the same estimated average score may differ in effective ability and effective speed. The score may be a result of high effective ability or high effective speed. b) Speed is only corrected for in the scoring for persons that do reach the time limit, but not for test takers that complete the test within the limit. However, even test takers that are within the time limit differ in their speed. Thus, different target abilities would be estimated for both groups (see also Pohl & von Davier, 2018). This is not the case in the SA model. If speed should be part of the construct to be measured, it should be incorporated in the same way for all test takers. Additionally, c) scoring NRIs as incorrect results in violations of model assumptions (local stochastic independence, measurement invariance) and was recognized to introduce bias more than 40 years ago (Lord, 1974). Last but not least, d) differences in test time limits are not accounted for by incorrect scoring. Thus, we think that it is valuable and incorporates the different advantages of the previous approaches to first disentangle the different aspects of performance using the SA model and building composite scores in an additional step.

3

A Multi-Process Item Response Model for Not-Reached Items due to Time Limits and Quitting

This chapter is published as Ulitzsch, E., von Davier, M., & Pohl, S. (2019b). A multi-process item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*. doi:10.1177/0013164419878241.

So far, modeling approaches for not-reached items have considered one single underlying process. However, missing values at the end of a test can occur for a variety of reasons: On the one hand, examinees may not reach the end of a test due to time limits and lack of working speed. On the other hand, examinees may not attempt all items and quit responding due to, e.g., fatigue or lack of motivation. We use response times retrieved from computerized testing to distinguish missing data due to lack of speed from missingness due to quitting. On the basis of this information, we present a new model that allows to disentangle and simultaneously model different missing data mechanisms underlying not-reached items. The model a) supports a more fine-grained understanding of the processes underlying not-reached items and b) allows to disentangle different sources describing test performance. In a simulation study we evaluate estimation of the proposed model. In an empirical study we show what insights can be gained regarding test-taking behavior using this model.

In large-scale assessments (LSAs), examinees do not always attempt all items they were assigned to answer. When an examinee fails to attempt a sequence of items presented at the end of a test, the resulting missing responses are referred to as not-reached items (NRIs). NRIs can occur for a variety of reasons. Examinees may not reach the end of a test due to lack of speed when tests are administered with time limits. This is supported by results from experimental research suggesting that increased test-taking time results in lower NRI rates (e.g. Mandinach, Bridgeman, Cahalan-Laitusis, & Trapani, 2005; Wild & Durso, 1979). However, the onset of NRIs does not seem to depend solely on test time. Examinees may quit the assessment prematurely due to, e.g., fatigue or lack of motivation. This is particularly the case in low-stakes assessments where low motivation is likely to affect examinee test-taking behavior (Chen, von Davier, Yamamoto, & Kong, 2015; Cosgrove, 2011; Liu, Rios, & Borden, 2015; Wise & DeMars, 2005). Indeed, in LSAs, NRIs are even observed in assessments administered without time constraints, such as in the Programme for the International Assessment of Adult Competencies (PIAAC, OECD, 2013).

NRIs occurring due to lack of speed and NRIs occurring due to quitting represent different types of missingness processes which tend to occur under different testing situations, correspond to different test-taking strategies, and might be related differently to ability. Thus, disentangling and modeling different types of NRIs can be beneficial for understanding examinee performance as well as for informing decisions regarding the adequate treatment of missing values due to NRIs. In this context, considering additional data retrieved from computer-based assessment facilitates the understanding of examinee behavior and thus of potential mechanisms underlying NRIs. For instance, cumulative response times (RTs) contain information on the time passed up to the last item attempted before ending the assessment. This allows to distinguish examinees who worked at a slow pace and reached the time limit before reaching the end of the test from those who displayed cumulative RTs far below the time limit without attempting all items administered (Pohl et al., 2019). Based on the information contained in RT data, Pohl et al. (2019) illustrated that within the same data set NRIs are plausible to occur due to different mechanisms – i.e., lack of speed and quitting. In this article, we argue that these are potentially different mechanisms that should be modeled as such. We propose a framework to disentangle and simultaneously model these missingness mechanisms.

3.1 Dealing with Not-Reached Items in Large-Scale Assessments

Current practices for handling NRIs in LSAs are rather heterogeneous. While in the majority of LSAs NRIs are either ignored (e.g., in the National Educational Panel

Study (NEPS), Pohl & Carstensen, 2012) or scored as incorrect, mixed approaches exist. For instance, in the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), NRIs are ignored for item parameter estimation and scored as incorrect for person parameter estimation (Foy, 2017, 2018). In PIAAC, NRIs are ignored if sufficient information about examinee proficiency is available – that is, if there are more than five item responses per domain. Otherwise, the examinees' self-stated reasons for not completing the assessment are considered when handling missing responses due to NRIs. For examinees quitting the assessment after giving responses to fewer than five items per domain, all NRIs are treated as incorrect if the self-stated reason for not responding is related to cognitive skills. If examinees give reasons unrelated to competence, NRIs are ignored in the analysis (OECD, 2013). This treatment of NRIs acknowledges that NRIs can occur due to different mechanisms. The approach is however rather heuristic in that it a) relies on self-stated reasons for not reaching the end of the test and b) distinguishes different types of NRIs only for examinees who responded to less than five items.

Scoring NRIs as wrong assumes the probability to solve an NRI to be zero – regardless of the examinee's ability level (see Lord, 1983; Rose, 2013; Rose et al., 2017). Ignoring NRIs assumes ignorability of the missingness mechanism. For ignorability to hold, data need to be missing at random (MAR), that is, missingness needs to be conditionally independent of the unobserved data given the observed data. Furthermore, the (unobserved) parameters governing the distribution of NRIs need to be distinct from ability (Mislevy & Wu, 1996; Rubin, 1976). There is, however, a substantial body of research suggesting that not reaching the end of a test is indeed related to ability (Debeer et al., 2017; Glas & Pimentel, 2008; Lawrence, 1993; List et al., 2019; Pohl et al., 2014; Rose et al., 2010). This indicates that the mechanisms underlying NRIs are nonignorable. Not properly accounting for the mechanisms that produce nonignorable missing data poses a threat to valid inferences and may potentially lead to biased person and item parameter estimates or distort relationships between ability and explanatory variables as well as country rankings (Glas & Pimentel, 2008; Köhler et al., 2017; Pohl et al., 2014; Rose, 2013). In order to properly account for nonignorable NRIs, a model for the mechanisms underlying their occurrence is needed.

3.1.1 Model-Based Approaches for Nonignorable Missing Values

In recent years, model-based approaches for handling nonignorable NRIs have been developed. In this class of models, information about NRIs is integrated into item

response theory (IRT) models – either employing a latent or manifest variable – and thus accounted for when estimating ability.

For modeling ability, customary IRT models are employed. In the case of a Rasch model, the probability of a correct response on response indicator u_{ij}, containing person i's response on item j can be modeled as a function of person ability θ_i and the item's difficulty b_j

$$p(u_{ij} = 1) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}. \tag{3.1}$$

Missing values due to not reaching the end of the test are not coded as wrong, but rather treated as MAR by including terms for missing data in the likelihood function used to estimate parameters.

Rose et al. (2010) suggested to employ a manifest missing data model, since, due to the monotone missing pattern resulting from NRIs, considering the number of reached items is sufficient to account for NRIs. Within manifest NRI approaches, information on the number of reached items k_{tot_i} is included as a manifest variable in the background model. This can be achieved by either regressing ability θ_i on k_{tot_i} or by applying multi-group IRT models where stratification on k_{tot_i} serves as a grouping variable (Rose et al., 2010). Manifest approaches are computationally less intensive compared to latent variable approaches for NRIs and since 2015 are considered in the population model in the Programme for International Student Assessment (PISA) for the generation of plausible values (OECD, 2017).

Within latent variable approaches for nonignorable missing values due to NRIs, the information on NRIs is included in form of a second dimension describing the propensity to reach the end of the test (Glas & Pimentel, 2008; List et al., 2019). Missingness indicators d_{ij}, being defined as 1 if u_{ij} is observed, 0 if u_{ij} is the first NRI, and coded as missing otherwise, constitute the measurement model for this propensity. $p(d_{ij} = 1)$ is then modeled as a function of examinee i's propensity to reach the end of the test and item j's response difficulty. Linear restrictions are imposed on the response difficulty parameters implying a monotonously decreasing probability of observing a response.

In both latent and manifest variable approaches for modeling the onset of NRIs, correlations different from zero between ability and the number of reached items/ the propensity to reach the end of the test indicate that the onset of NRIs is related to the construct being measured and thus nonignorability of the missingness mechanism. Including information about nonignorable NRIs has been shown to yield less biased and more accurate parameter estimates as compared to ignoring or scoring missing values as wrong (Glas & Pimentel, 2008; Rose et al., 2017; Rose et al., 2010).

3.1.2 Using Response Times to Model Not-Reached Items

If NRIs occur due to lack of speed, examinees reach the time limit before managing to reach the end of the test. For this case it has been shown that the missing data process underlying NRIs can be described by examinee speed (Pohl et al., 2019). With the widespread availability of RT data retrieved from computerized testing, a direct measure of speed becomes available (van der Linden, 2006). Pohl et al. (2019) were the first to suggest utilizing this information on the missingness process by employing van der Linden's (2007) hierarchical speed-accuracy (SA) framework to model the occurrence of NRIs in low-stakes assessments. They showed that the SA model a) can successfully model NRIs due to time limits, b) provides a closer description of the missing data processes than model-based approaches for nonignorable missing values, and c) can also deal with varying enforcement of time limits – given, that NRIs are the result of lack of speed.

In the SA model, first-level models are specified separately for the responses and associated RTs. For the response indicators u_{ij} van der Linden has recommended employing customary IRT models. For the RTs t_{ij}, denoting the time examinee i required to generate an answer to item j, a lognormal model with separate person and item parameters is chosen. That is, logarithmized RTs are assumed to follow a normal distribution. In the lognormal model, logarithmized RTs are considered to be a function of the examinee's speed τ_i and the item's time intensity β_j:

$$\begin{gathered} \ln(t_{ij}) = \beta_j - \tau_i + \epsilon_{ij}, \\ \text{with} \\ \epsilon_{ij} \sim \mathcal{N}(0, \alpha_j^{-2}). \end{gathered} \tag{3.2}$$

α_j represents the inverse of the RTs' standard deviation and can be interpreted as a time discrimination parameter. That is, the larger α_j, the larger the proportion of the RT variance that stems from differences in speed across examinees. On a second level, joint multivariate normal distributions of person and item parameters are specified.

Pohl et al. (2019) delineated that approaches that consider the number of NRIs and the SA model are closely related. First, both approaches include an additional variable that represents the missingness mechanism in the model. While the SA model includes a direct measure of speed, model-based approaches for nonignorable missing values include the tendency to reach the end of the test, as measured by the number of NRIs. If the number of NRIs is a result of lack of speed under testing conditions with time limits, the propensity to reach the end of the test can be

understood as a proxy of working speed. In this case, the SA model presents a better and finer-grained description of the missingness process (Pohl et al., 2019).

Second, both approaches assume a single mechanism underlying NRIs. Obviously, when the SA model is applied to account for NRIs, it is assumed that NRIs occurred due to lack of speed. Although model-based approaches that consider the number of NRIs do not explicitly rely on this assumption, they still assume the same missingness mechanism for all NRIs. When mechanisms leading to NRIs differ across examinees, that is, when multiple mechanisms are underlying NRIs such as lack of speed and quitting, this assumption is violated. Hence, in the case that missingness at the end of a test occurs not only due to speed, but also due to motivational reasons, neither controlling for speed nor for the tendency to reach the end of the test is sufficient to properly model NRIs.

3.2 Objective

We propose a new framework that takes into account that multiple mechanisms can underlie NRIs. Doing so requires a) distinguishing examinees who quit the assessment from those who did not work with sufficient speed and reached the time limit and b) to establish a model that considers both mechanisms simultaneously.

The remainder of this article is organized as follows: First, we present an approach that distinguishes between and simultaneously models two different types of NRIs. Second, parameter recovery of the proposed model is investigated using a simulation study. Third, the relevance of the model for understanding the missingness processes is illustrated in an empirical example.

3.3 Speed-Accuracy+Quitting Model

The proposed speed-accuracy+quitting (SA+Q) framework, depicted in Figure 3.1, is an extension of the hierarchical SA model that also accounts for quitting. Following Pohl et al. (2019), NRIs due to lack of speed are modeled by considering examinee speed. For simplicity, we model ability θ employing a Rasch model as given by Equation 5.1. This is in accordance with the analysis frameworks of major LSAs (e.g. NEPS, Pohl & Carstensen, 2012). Note that the model can be extended to other measurement models (see Ulitzsch, von Davier, & Pohl, 2019c). For speed, we employ the lognormal model suggested by van der Linden (2006) as given by Equation 3.2. We model the quitting process by considering the number of items reached before quitting. These are determined by employing RTs to distinguish between examinees displaying NRIs due to lack of speed or due to quitting and

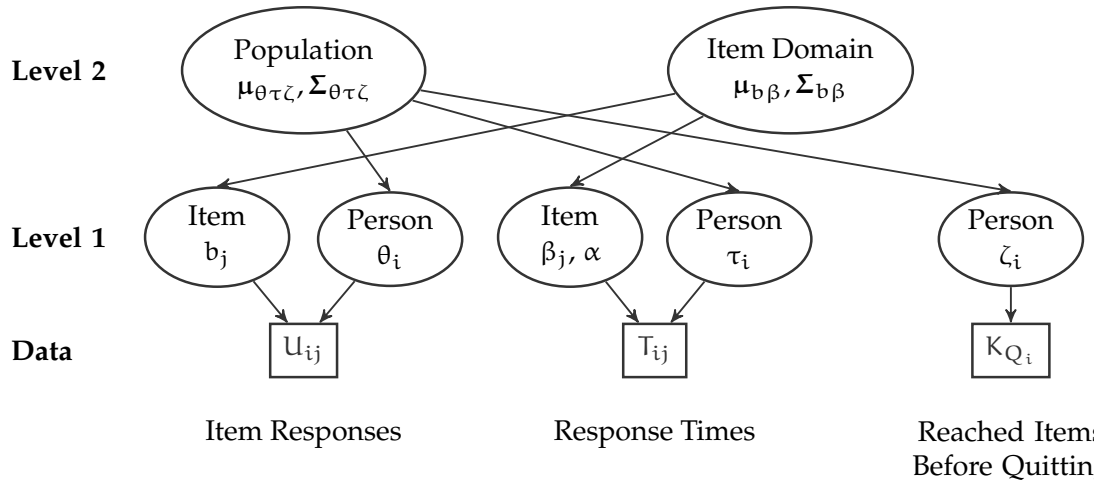

Figure 3.1. Hierarchical framework for the joint modeling of speed, accuracy, and test endurance.

constitute the measurement model for a newly introduced variable giving examinee test endurance. If the end of the test or the time limit was reached, the number of items reached before quitting is set to be missing, as no information about quitting is available in this case.

3.3.1 Identifying Quitting

In the present study, we assume that examinee i has quit the assessment when a) he or she did not reach the end of the test (that is, k_{tot_i} is smaller than the number of items administered K) and b) his or her total RT t_{tot_i} falls below the time limit (that is, t_{tot_i} is smaller than T_{lim}). Based on this information, an indicator c_i of observed quitting behavior can be constructed as follows

$$c_i = \begin{cases} 1 & \text{if } k_{tot_i} < K \text{ and } t_{tot_i} < T_{lim} \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

By construction, $\mathbf{c}$ distinguishes between examinees who quit the assessment ($c_i = 1$) and those who reached the time limit or finished the test before quitting ($c_i = 0$).

3.3.2 Modeling Quitting

To model the quitting process, we utilize the information contained in the number of reached items up to the point where the assessment has been quit k_{Q_i}. When quitting behavior has been observed, that is, when $c_i = 1$, k_{Q_i} is given by the observed number of reached items k_{tot_i}. We suggest to employ a Poisson lognormal model for k_{Q_i}. These are common for count data on the test or task level (Doebler & Holling, 2016;

Jansen, 1994, 1995), such as the number of correct items in a task or – as mentioned in passing by Jansen (1995) – the number of completed items in a test. More specifically, we model the probability that examinee i quits the assessment after attempting k items as a Poisson process with mean λ_i, where $\ln(\lambda_i)$ corresponds to the person parameter ζ_i:

$$\begin{gathered} p(k_{Q_i} = k) = \frac{e^{-\lambda_i}\lambda_i^k}{k!}, \\ \text{with} \\ \ln(\lambda_i) = \zeta_i. \end{gathered} \tag{3.4}$$

ζ_i denotes examinee i's test endurance and thus governs the item position at which examinee i is most likely to quit the assessment – given that the assessment has not been quit before. As such, ζ_i can be understood as a survival parameter. In the context of NRIs, the onset of NRIs poses the event of interest occurring within a sequence of item positions (List et al., 2019). The survival function $S(k)$ depends on the Poisson lognormal model for k_{Q_i} and gives the probability that examinee i will continue the assessment beyond k items, as follows:

$$S(k) = p(k_{Q_i} > k) = 1 - F(k) = 1 - \sum_{l=0}^{k} \frac{e^{-\lambda_i}\lambda_i^l}{l!}, \tag{3.5}$$

with $F(k)$ denoting the Poisson cumulative distribution function.

The presented model incorporates the assumption that examinees will continue the assessment only for a definite number of items and thus will quit the assessment at some point. Note, however, that quitting behavior is not fully observable since some examinees either manage to complete the test or reach the time limit before quitting. Hence, under $c_i = 0$, k_{Q_i} is exposed to right censoring. That is, the observed number of reached items k_{tot_i} corresponds to the number of reached items before quitting only under $c_i = 1$. Otherwise, k_{tot_i} marks the item position k_{C_i} at which k_{Q_i} has been right-censored. The relationship between k_{tot_i}, k_{Q_i}, and the censoring variable k_{C_i} is given by

$$k_{tot_i} = \min(k_{Q_i}, k_{C_i}). \tag{3.6}$$

Table 3.1 illustrates this relationship for three examinees administered a test of length $K = 20$ with a a time limit of T_{lim} =1,800 seconds. Examinee 1 reached the end of the test within the allocated time without showing quitting behavior. Hence, k_{Q_1} is censored due to reaching the end of the test at $k_{C_1} = k_{tot_1} = K = 20$. Examinee 2 did

not reach the end of the test due to lack of speed. Hence, k_{tot_2} gives the item position k_{C_2} at which quitting behavior has been censored due to lack of speed. Examinee 3 did reach neither the end of the test nor the time limit and is thus assumed to have quit the assessment. Thus, k_{tot_3} corresponds to the item position at which the assessment has been quit k_{Q_3}.

Table 3.1. Illustration of the relationship between k_{tot_i}, k_{Q_i}, and k_{C_i}

	k_{tot_i}	t_{tot_i}	c_i	k_{Q_i}	k_{C_i}
1	20	1,450s	0	NA	20
2	16	1,800s	0	NA	16
3	16	1,450s	1	16	NA

Note: k_{tot_i}: observed number of reached items, t_{tot_i}: total response time, k_{Q_i}: number of reached items before quitting, k_{C_i}: censoring item position, c_i: quitting indicator.

Note that usually in LSAs only a few examinees show quitting behavior. That is, there are few observations with $c_i = 1$. This results in sparse data with respect to the number of reached items before quitting k_{Q_i} on the one hand and a large portion of k_{Q_i} to be assumed to have been censored on the other. If speed is related to test endurance and there is censoring of k_{Q_i} due to reaching the time limit, speed is informative with respect to the censoring of k_{Q_i}. Under such conditions, speed is related to both the parameter governing the distribution of k_{Q_i} and the probability that k_{Q_i} is censored. Modeling test endurance and speed jointly accounts for the informative censoring of k_{Q_i} (Baker, Fitzmaurice, Freedman, & Kramer, 2005). Likewise, modeling ability and speed jointly with test endurance accounts for nonignorable missingness due to quitting on response as well as on RT indicators.

3.3.3 Second Level Models

In analogy to the SA model, on the second level, the joint distributions of the first-level person and item parameters are modeled. Following van der Linden (2007), person parameters are assumed to be multivariate normal with mean vector

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\tau, \mu_\zeta) \tag{3.7}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} & \sigma_{\theta\zeta} \\ \sigma_{\theta\tau} & \sigma_\tau^2 & \sigma_{\tau\zeta} \\ \sigma_{\theta\zeta} & \sigma_{\tau\zeta} & \sigma_\zeta^2 \end{pmatrix}. \tag{3.8}$$

Assessing the joint distribution of person parameters provides valuable insights into the processes underlying NRIs, their relationship to ability as well as to each other. Non-zero correlations with test endurance indicate nonignorability of the associated missingness process.

For the sake of simplicity, for the measurement model of RTs, time discrimination parameters α_j are constrained to be equal across items, that is

$$\alpha_j = \alpha \text{ for all } j \tag{3.9}$$

This constraint can be understood as an analogue to the Rasch model in IRT (van der Linden, 2006) and thus mirrors the Rasch parameterization implemented for item responses in major LSAs (see, e.g., Pohl & Carstensen, 2012).

The joint distribution of item parameters is, in accordance with van der Linden (2007), assumed to be multivariate normal with mean vector

$$\boldsymbol{\mu}_{\mathcal{I}} = (\mu_b, \mu_\beta) \tag{3.10}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{I}} = \begin{pmatrix} \sigma_b^2 & \sigma_{b\beta} \\ \sigma_{b\beta} & \sigma_\beta^2 \end{pmatrix}. \tag{3.11}$$

When a Rasch model is employed for response indicators, the model can be identified by setting the expectations of θ and τ to zero.

Assuming joint distributions for person and item parameters yields the following likelihood

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{j=1}^{k_{tot_i}} p(u_{ij}|b_j, \theta_i) f(t_{ij}|\beta_j, \tau_i, \alpha) p(k_{Q_i}|\zeta_i)^{c_i} S(k_{C_i}|\zeta_i)^{1-c_i} \\ g(\theta_i, \tau_i, \zeta_i|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) h(b_j, \beta_j|\boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}). \tag{3.12}$$

The first four terms incorporate the assumption of conditional independence of responses, RTs, and number of reached items before quitting given the second-order variables in the model. The third and fourth term take the right censoring of quitting behavior into account: For examinees who quit the assessment ($c_i = 1$), the probability that examinee i quits the assessment after attempting k_{Q_i} items as given by Equation 3.4 contributes to the likelihood function. For examinees with unobserved quitting behavior ($c_i = 0$), the likelihood function considers the probability that examinee i will continue the assessment beyond the censoring position k_{C_i} as given by the survival function in Equation 3.5. $g(\theta_i, \tau_i, \zeta_i|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ and $h(b_j, \beta_j|\boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}})$ denote

the multivariate normal densities of the person and item parameters, respectively. To facilitate estimation of the SA+Q model, we employ Bayesian estimation techniques.

3.4 Parameter Recovery

We conducted a simulation study to investigate whether true parameter values can satisfactorily be recovered in estimation under realistic conditions. The SA model and extensions thereof have been shown to yield good parameter recovery under realistic conditions (e.g., Fox & Marianti, 2016; Molenaar et al., 2015; Pohl et al., 2019). We therefore focused especially on possible challenges for estimation imposed by censoring of quitting behavior and the resulting data sparseness on k_{Q_i}.

3.4.1 Data Generation

Data were generated using R version 3.5.1 (R Development Core Team, 2017). We employed the SA+Q model as the data-generating model. Using the mvrnorm function from the MASS package (Venables & Ripley, 2002), person and item parameters were randomly drawn from multivariate normal distributions with variances and covariances set to values similar to those of the data application reported below. Population values of the data-generating model are reported in Table 3.2. We employed a Rasch model for the item responses and set time discrimination parameter for all items to $\alpha = 1.75$ (van der Linden, 2007).

Missing values were induced based on a) cumulative RTs across item positions and b) the number of reached items before quitting. Cumulative RTs give the time passed when the respective item is responded to. The number of reached items before quitting was generated for each examinee according to the Poisson lognormal model for the quitting process. All items with either a cumulative RT exceeding the time limit or whose position exceeded the number of reached items before quitting were assumed to be not reached and coded as missing.

To evaluate the effects of censoring of quitting behavior, we considered multiple censoring mechanisms and varied four factors that are relevant for data sparseness:

(a) the sample size ($N = 350$; $N = 700$), representing low and medium sample sizes per item encountered in LSAs with balanced incomplete block designs (see Gonzalez & Rutkowski, 2010).

(b) the test length ($K = 20$ with $T_{lim} = 1{,}800$s; $K = 40$ with $T_{lim} = 6{,}600$s).

(c) the rate of NRIs (2.5%; 5%; 10%), reflecting the upper three quarters of a typical range of percentages of NRIs. For instance, in the PISA 2012 computer-based

assessment, percentages of NRIs across booklets ranged from 0.42% to 11.19% (OECD, 2014).

(d) the missingness mechanisms underlying NRIs (NRIs caused solely by quitting; half of the NRIs resulting from quitting and half from lack of speed)[1]. While the former represents a testing condition without or a very generous time limit (e.g., as in PIAAC), the latter represents a more speeded testing situation where some examinees run out of time before reaching the end of the test or quitting (e.g., as in PISA). We included these conditions in order to a) assess whether the proposed model yields unbiased and efficient parameter estimates under various testing situations and b) to disentangle possible effects of different censoring mechanisms on estimation accuracy and efficiency. Under the first condition, the number of reached items $\mathbf{k}_Q$ is censored due to test length. Under the second condition, censoring occurs due to both, test length and lack of speed.

Under this design, conditions with 5% (2.5%) missing values and no censoring due to speed display the same missingness rates due to quitting as conditions with 10% (5%) missing values with half of it going back to quitting. This allows to assess whether the model performs differently when – for the same amount of information available on test endurance – the overall missingness rate increases and k_{Q_i} is exposed to different censoring mechanisms. We controlled the amount and types of NRIs by varying the expectations of test endurance μ_ζ and speed μ_τ (see Table 3.2). Note that low rates of missingness due to quitting might go back to relatively high proportions of examinees exhibiting such behavior. This becomes evident in Table 3.2, displaying the corresponding average proportions of simulated examinees exhibiting quitting behavior across all cells of the simulation design. In total, the simulation design led to 2 x 2 x 3 x 2 = 24 conditions. For each cell of the simulation design, we generated 100 data sets.

3.4.2 Estimation Procedure

We employed Bayesian estimation with Gibbs sampling. All analyses were conducted in JAGS version 4.3.0. (Plummer, 2003) using the rjags package (Plummer, 2016) for R version 3.5.1 (R Development Core Team, 2017). Settings for noninformative priors were chosen following recommendations provided by Fox (2010) and Gelman and Hill (2007).

[1]Note that in the case of NRIs going back entirely to lack of speed, there is no need to model quitting behavior and the SA model is sufficient for modeling the mechanism underlying NRIs.

Table 3.2. Population parameters of the data-generating model

Person Parameters					Item Parameters			
	θ	τ	ζ	$\boldsymbol{\mu}_{\mathcal{P}}$		$\mathfrak{b}$	β	$\boldsymbol{\mu}_{\mathcal{I}}$
θ	1.00			0.00	$\mathfrak{b}$	1.50		0.00
τ	-.40	.10		μ_τ	β	0.40	.25	0.00
ζ	-.15	.25	0.65	μ_ζ				

		Missingness Mechanisms					
		quitting			speed & quitting		
K	% NR	% Q	μ_τ	μ_ζ	% Q	μ_τ	μ_ζ
20	2.5%	8.15%	-2.50	4.15	4.28%	-3.75	4.40
	5%	15.08%	-2.50	3.85	7.50%	-3.85	4.15
	10%	26.49%	-2.50	3.50	12.75%	-4.00	3.85
40	2.5%	7.83%	-2.50	4.85	5.30%	-3.75	5.00
	5%	14.50%	-2.50	4.55	7.20%	-3.85	4.85
	10%	28.44%	-2.50	4.15	12.60%	-4.00	4.55

Note: N: number of examinees; K: number of items; % NR: overall missingness rate due to not-reached items; % Q: percentage of examinees quitting; θ: ability; τ: speed; ζ: test endurance; $\mathfrak{b}$: item difficulty; β: time intensity; $\boldsymbol{\mu}_{\mathcal{P}}$ and $\boldsymbol{\mu}_{\mathcal{I}}$ give mean vectors of person and item parameters, respectively. Mean speed μ_τ and the mean test endurance μ_ζ are varied in the simulation design to control the amount of mechanisms underlying not-reached items.

For person and item parameter variances and covariances, we employed inverse Wishart priors with

$$\boldsymbol{\Sigma}_{\mathcal{P}} \sim \text{IW}_{3+1}(\mathbf{I}_3) \tag{3.13}$$

and

$$\boldsymbol{\Sigma}_{\mathcal{I}} \sim \text{IW}_{2+1}(\mathbf{I}_2), \tag{3.14}$$

where $\mathbf{I_3}$ and $\mathbf{I_2}$ represent identity matrices of dimension 3 and 2, respectively. These are default prior settings for inverse Wishart priors implemented in statistical software for Bayesian analyses (van Erp, Mulder, & Oberski, 2018). Note that inverse Wishart priors tend to be informative about variances when these are close to zero and the sample size is small (Alvarez, Niemi, & Simpson, 2014; Schuurman, Grasman, & Hamaker, 2016). Since usually the number of items is small, for item parameter variances and covariances, prior settings will have a larger impact.

We set μ_θ and μ_τ to zero for model identification. For the remaining item and person parameter means, we chose noninformative normal priors with mean zero and variance 1000^2. A noninformative gamma prior with shape 0.5 and rate $\frac{1}{100^2}$ was employed for squared time discrimination α^2. JAGS code for the SA+Q model is provided in Appendix B.1.

Each generated data set was analyzed running three MCMC chains with 100,000 iterations each. We employed a thinning factor of 5 and discarded the first 40,000 iterations as burn-in, saving 36,000 iterations as a sample of the posterior distribution. We determined the number of iterations in pre-analyses, inspecting potential scale reduction factor (PSRF) values, trace plots, and effective sample sizes. In the case of nonconvergence (i.e., PSRF values higher than 1.10, Gelman & Shirley, 2011), we increased the number of iterations by 50,000 per chain out of which 30,000 were discarded as burn-in. This procedure was repeated up to 250,000 iterations per chain in total.

3.4.3 Evaluation Criteria

We evaluated statistical performance in terms of convergence, bias in and efficiency of parameter estimates, as well as coverage of the true parameter values by 95% highest density intervals. Convergence was assessed on the basis of PSRF values, with PSRF values below 1.10 being considered acceptable (Gelman & Rubin, 1992; Gelman & Shirley, 2011). Coverage between .91 and .98 was considered good (Muthén & Muthén, 2002).

3.4.4 Results

CONVERGENCE Table 5.1 gives the proportions of replications converging after 100,000 up to 250,000 iterations across all cells of the simulation design. Reaching convergence was more challenging under conditions with low quitting rates. This effect was more pronounced in conditions with larger number of items. Accordingly, conditions with $K = 40$ and an overall missingness rate of 2.5% out of which half went back to quitting were most challenging with respect to convergence, with up to 58% of the replications not converging after 250,000 iterations. The number of iterations needed to reach convergence decreased rapidly with higher quitting rates, such that under conditions with 10% missingness solely due to quitting, at least 95% of the replications converged after 250,000 iterations at most. No considerable differences concerning convergence could be observed for the same amount of missing values due to quitting when either missing values due to speed were present or not.

Parameters typically yielding high PSRF values were test endurance mean and variance estimates. This is due to the censoring of quitting behavior, with the distribution of test endurance needing to be extrapolated based on information from examinees assumed to belong to the lower quartiles of the distribution. When quitting behavior is observable only for few examinees, information on the distribution of test endurance is sparse, impacting convergence. Replications that did not converge were excluded from further analyses.

COVERAGE Coverage values for all parameter types and conditions are available in Appendix B.2. For item parameter variances, covariances, and means coverage was satisfactory across all conditions. Coverage values falling below .91 occurred rarely and the lowest coverage value was still as high as .82. For person parameters, when either the number of examinees or items was sufficiently high ($K = 40$ or $N = 700$), coverage fell below .91 only for test endurance variance $var(\zeta)$ and mean μ_ζ estimates under conditions with missingness rates due to quitting below 5% (that is, under conditions with less than 10% NRIs due to speed and quitting as well as under conditions with an overall missingness rate below 5% only due to quitting). Under these conditions, for $var(\zeta)$, the lowest observed coverage values were .88 and .67 for conditions with $K = 20$ and $K = 40$, respectively. For μ_ζ, the lowest observed coverage values were .76 and .75 for conditions with $K = 20$ and $K = 40$, respectively, under conditions with less than 5% missingness caused by quitting.

PARAMETER ESTIMATION To evaluate bias in and efficiency of parameter estimates, we assessed the median along with 90% ranges of the means of the posterior distribution. Figures 3.2 and 3.3 depict results for person parameter variances and

Table 3.3. Proportion of replications converging after 100,000 to 250,000 iterations

K	N	% NR	Mechanisms	Iterations			
				100,000	150,000	200,000	250,000
20	350	2.5%	quitting	.13	.31	.52	.64
			speed & quitting	.07	.26	.38	.46
		5%	quitting	.28	.55	.73	.85
			speed & quitting	.13	.31	.44	.60
		10%	quitting	.46	.78	.91	.96
			speed & quitting	.25	.50	.66	.81
	700	2.5%	quitting	.08	.32	.53	.64
			speed & quitting	.08	.22	.35	.48
		5%	quitting	.29	.50	.67	.77
			speed & quitting	.12	.26	.44	.51
		10%	quitting	.59	.87	.97	.99
			speed & quitting	.17	.35	.54	.75
40	350	2.5%	quitting	.07	.23	.41	.57
			speed & quitting	.03	.11	.29	.42
		5%	quitting	.20	.43	.62	.74
			speed & quitting	.10	.20	.41	.55
		10%	quitting	.55	.79	.95	.98
			speed & quitting	.13	.28	.51	.64
	700	2.5%	quitting	.05	.25	.40	.57
			speed & quitting	.10	.23	.38	.47
		5%	quitting	.19	.37	.62	.77
			speed & quitting	.10	.24	.35	.45
		10%	quitting	.57	.84	.92	.97
			speed & quitting	.14	.34	.59	.71

Note: % NR: overall missingness rate due to not-reached items; quitting and speed & quitting denote conditions under which all not-reached items go back to quitting and not-reached items occurred due to both lack of speed and quitting, respectively; K: number of items; N: number of examinees.

covariances as well as mean test endurance posterior means.[2] Across all conditions, median person parameter variance and covariance estimates were close to the true data-generating values. The only exception were parameters concerning the distribution of test endurance (μ_ζ and $var(\zeta)$), which were sensitive to bias under conditions with missingness rates due to quitting below 5% as well as under conditions with few items. Under conditions with $K = 20$, median parameter estimates of $var(\zeta)$ ranged from 0.53 up to 0.71 as compared to the true value of 0.65. Bias for μ_ζ under conditions with $K = 20$ was less severe. Under conditions with $N = 350$ and 10%,

[2]Note that plots for parameter recovery of mean test endurance μ_ζ (Figure 3.3) are organized according to the data-generating values employed to achieve missingness rates due to quitting ranging from 1.25% (speed & quitting, 2.5%) to 10% (quitting, 10%).

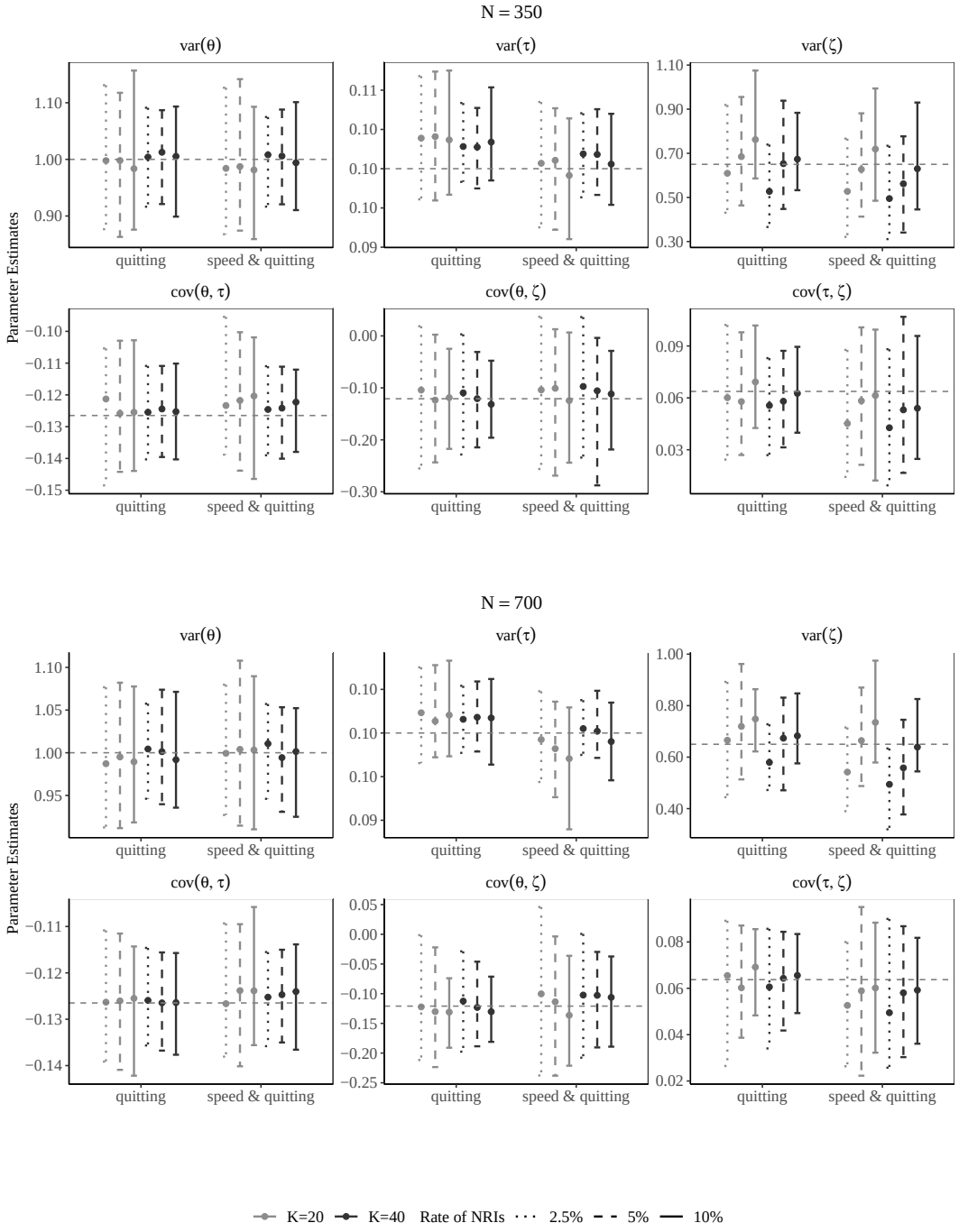

Figure 3.2. Medians and 90% ranges of person parameter variance and covariance estimates over all 100 replications per condition. The dashed horizontal line indicates the respective true parameter. Note that y-axes differ in scale. θ: ability; τ: speed; ζ: test endurance; N: number of examinees; K: number of items; NRIs: not-reached items; quitting and speed & quitting denote conditions under which all not-reached items go back to quitting and not-reached items occurred due to both lack of speed and quitting.

respectively, 1.25% missingness due to quitting, median parameter estimates of 3.55 and 4.19 as compared to the data-generating value of 3.50 and 4.40, respectively, were observed. Under conditions with $K = 40$ and a missingness rate due to quitting of at least 5%, median estimates of $\nu\alpha r(\zeta)$ and μ_ζ were well recovered. Median bias in parameter estimates did not vary largely across the sample sizes under consideration. Nevertheless, variability of parameter estimates decreased with increasing sample size.

Results for bias and efficiency of item parameter means, variances, and covariances are given in Appendix B.3. Due to the small number of items, estimates were shrunken towards the prior mean of the inverse Wishart prior (see Alvarez et al., 2014; Daniels & Kass, 1999).

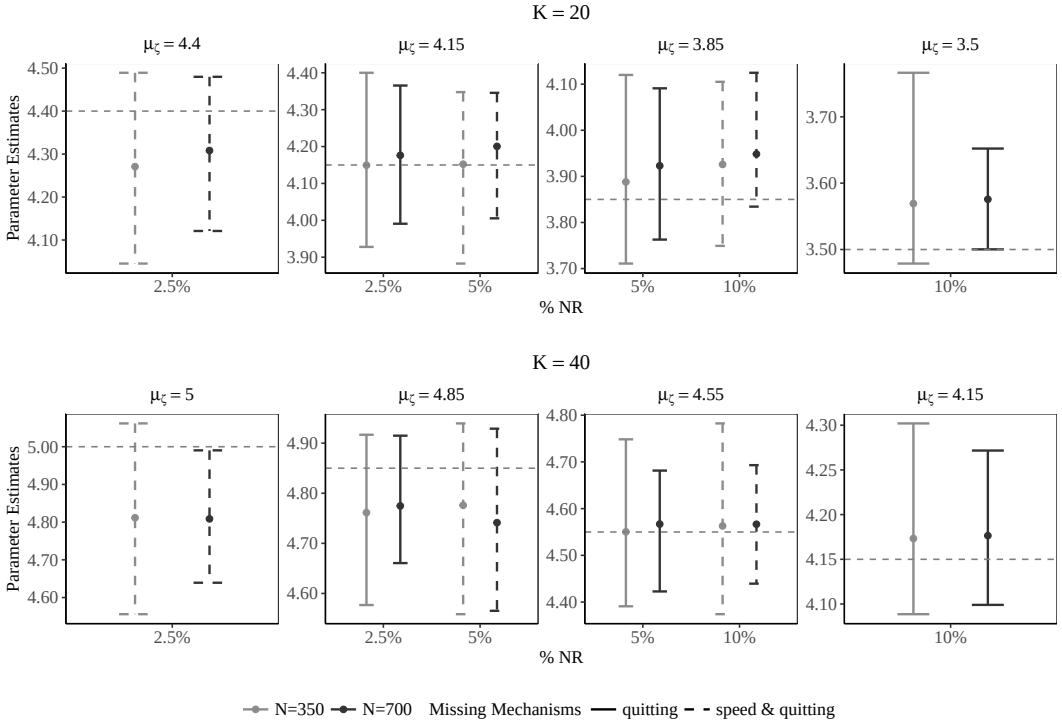

Figure 3.3. Medians and 90% ranges of mean test endurance μ_ζ over all 100 replications per condition. The dashed horizontal line indicates the respective true parameter. Plots are organized according to the data-generating values employed to achieve missingness rates due to quitting ranging from 1.25% (speed & quitting, 2.5%) to 10% (quitting, 10%). Note that y-axes differ in scale. N: number of examinees; K: number of items; % NR: overall missingness rate due to not-reached items; quitting and speed & quitting denote conditions under which all not-reached items go back to quitting and not-reached items occurred due to both lack of speed and quitting, respectively.

SUBSEQUENT ANALYSES In subsequent analyses we aimed at investigating whether the recovery of $var(\zeta)$ under conditions with few items ($K = 20$) improves with an increasing amount of NRIs. To do so, we increased the missingness rate under the condition with $K = 20$, $N = 350$ and missingness solely caused by quitting to 20% by setting μ_ζ to 3. We employed 100 replications and analyzed all generated data sets employing three chains with 250,000 iterations each. No convergence issues occurred. Indeed, the additional condition yielded a median estimate of $var(\zeta)$ of 0.67 as compared to the true value of 0.65, thereby supporting the conclusion of the simulation study that test endurance variances can better be recovered when data sparseness on k_{Q_i} is less severe.

3.5 Empirical Example

We used data from the Spanish sample of the PISA 2015 assessment to illustrate the use of the SA+Q model for the understanding of the occurrence of NRIs. We analyzed data from examinees who were administered science cluster number 7 at the second position out of four 30 minute blocks. For reasons of simplicity, we eliminated examinees who showed item omissions from further analyses. The final sample consisted of 326 examinees responding to 17 items. The data set under consideration displayed a missingness rate of 7.13%, going back to those 21.47% of examinees who did not reach the end of the cluster. The majority of examinees with NRIs (58.57%) reached all but 4 items at most.

3.5.1 Total Response Time Distributions

In a first step, to get a better understanding of possible mechanisms, we followed Pohl et al. (2019), and examined distributions of total RT t_{tot_i}. Figure 3.4 displays t_{tot_i} as a function of the number of NRIs. The time limit T_{lim} of 1,800 seconds is marked with a dashed horizontal line. The results suggest that the time limit was enforced with varying rigor by test administrators, since for some examinees t_{tot_i} exceeded the time limit. More importantly, t_{tot_i} varied largely across examinees with NRIs. Such variation was not to be expected under conditions where NRIs occurred entirely due to lack of speed, where all t_{tot_i} associated with NRIs should be close to T_{lim}. Instead, the fact that t_{tot_i} associated with NRIs was close to T_{lim} only for some examinees, while it was considerably below T_{lim} for others can be understood as evidence that different mechanisms – i.e., lack of speed and premature quitting – are underlying NRIs.

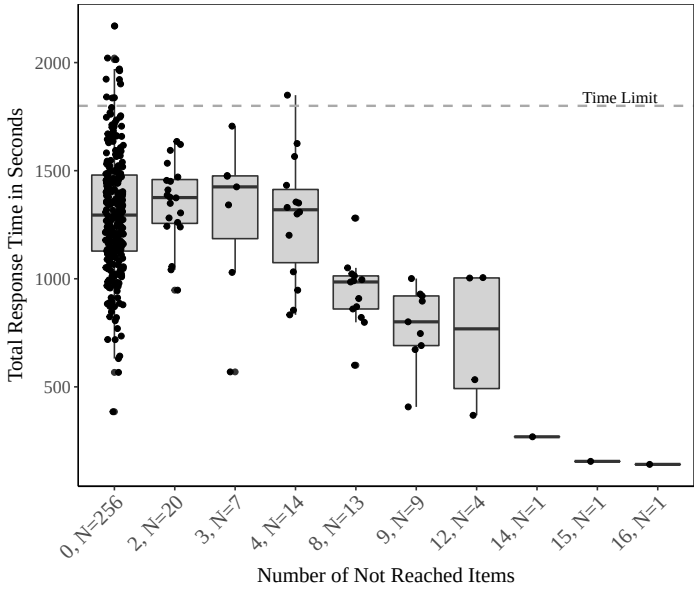

Figure 3.4. Total response time distributions for PISA science cluster number 7 administered in Spain, by number of not-reached items. The dashed horizontal line marks the time limit T_{lim} of 1,800 seconds.

3.5.2 Investigating the Occurrence of Not-Reached Items

For simplicity, when classifying examinees as quitters, we ignored that in PISA time limits were enforced with varying rigor. To deal with the fact that RT was not recorded for the last item seen when no response has been generated, we employed a heuristic approach and adjusted the decision boundary for t_{tot_i} associated with NRIs by a typical item-level RT. We classified NRIs as due to quitting when t_{tot_i} fell below the time limit by less than the 90th percentile of RTs across all items and examinees, that is, 1,800s $-$ 66.97s $=$ 1,733.03s.[3] Doing so led to classifying 68 out of 70 examinees who did not reach the end of the test as quitters. We employed the SA+Q model to analyze the data. With 250,000 iterations per chain, no PSRF values below 1.10 were encountered.

Results are displayed in Table 3.4. The negative correlation between ability θ and speed τ as well as between ability θ and test endurance ζ indicates that more able examinees tended to display lower general working speed and showed lower test endurance. That is, more able examinees were more likely to display NRIs due to both reaching the time limit and quitting. Note that the highest density interval for the correlation between θ and ζ includes zero which means that the correlation was not credibly different from zero. The positive correlation between speed τ and

[3]Note that other approaches may also be taken for classifying NRIs as due to quitting. Especially when more detailed information is available for each item, it might be possible to even better classify quitting. Since currently publicly available databases containing RT information do not provide RTs for the last item seen when no response has been generated, the heuristic decision rule employed here might serve as a first guideline for applying the SA+Q framework to such data.

test endurance ζ indicates that examinees who worked faster had the tendency to generate answers to more items before quitting the assessment. Furthermore, the fact that τ and ζ are not highly correlated underlines that missingness due to speed and missingness due to quitting should be seen as different processes. μ_ζ gives the expectation of the mean logarithmized number of items reached before quitting. The value of 3.52 corresponds to 33.78 items.

Table 3.4. Variances and means of as well as correlations among person and item parameters

	Person Parameters			
	θ	τ	ζ	$\boldsymbol{\mu}_\mathcal{P}$
θ	1.13			0
	[0.87; 1.39]			
τ	-.37	.07		0
	[-.50; -.24]	[0.06; 0.09]		
ζ	-.16	.30	0.62	3.52
	[-.35; .03]	[.12; .47]	[0.34; 0.94]	[3.31; 3.74]
	Item Parameters			
	$\mathfrak{b}$	β	$\boldsymbol{\mu}_\mathcal{I}$	
$\mathfrak{b}$	1.41		-0.36	
	[0.59; 2.47]		[-0.99; 0.18]	
β	0.41	0.21	4.14	
	[.02; .76]	[0.09; 0.37]	[3.94; 4.39]	

Note: Highest density intervals are given in squared brackets. θ: ability; τ: speed; ζ: test endurance; $\mathfrak{b}$: item difficulty; β: time intensity; $\boldsymbol{\mu}_\mathcal{P}$ and $\boldsymbol{\mu}_\mathcal{I}$ give mean vectors of person and item parameters, respectively.

3.6 Discussion

The speed-accuracy+quitting model proposed in this article allows to disentangle and simultaneously model different missing data mechanisms underlying NRIs. Namely, the SA+Q model distinguishes between NRIs stemming from lack of speed and NRIs due to quitting and thereby allows to substantively meaningful describe different processes underlying NRIs. This is achieved by further integrating research on missing data and research on RTs (see Pohl et al., 2019). The SA+Q model considers RT data to handle NRIs by a) utilizing the additional information contained in RTs to distinguish between examinees displaying NRIs due to lack of speed and those who quit and b) extending van der Linden's SA model by a Poisson lognormal survival model describing the quitting process in terms of examinee test endurance. The SA+Q model can be employed to model the occurrence of NRIs under conditions

where all NRIs go back to quitting as well as the occurrence of NRIs occuring due to both lack of speed and quitting.

The SA+Q model represents a refined model-based approach for dealing with nonignorable missing data. As delineated above, previously suggested model-based approaches for NRIs (Glas & Pimentel, 2008; Rose et al., 2010; van der Linden, 2007) rely on the assumption of a single missingness mechanism. The SA+Q model complements model-based approaches for NRIs in that it considers that multiple mechanisms can underlie their occurrence. As such, the SA+Q model overcomes limitations of current state-of-the-art approaches for NRIs.

We employed data from PISA 2015 to illustrate how the approach can provide insights into the processes underlying NRIs. We showed that there is strong evidence that NRIs in LSAs indeed can be attributed to different missingness processes. In this context the SA+Q model supports a finer-grained understanding of the occurrence of NRIs by further assessing examinee characteristics associated with test endurance. As such, the SA+Q model can be used to evaluate and inform substantive theories on test-taking behavior and strategies.

The model gives reasonable estimates under conditions with a missingness rate of at least 5% due to quitting (or approximately 15% of examinees exhibiting quitting behavior) and a higher number of items ($K = 40$). Since the SA+Q model estimates test endurance based on information on the number of reached items before quitting, the number of examinees quitting the assessment might be of greater importance for retrieving unbiased estimates than the missingness rate due to quitting. Generally, a higher number of iterations is needed when little information on quitting behavior is available. Note that model-based approaches yield ability estimates considerably different from those retrieved when ignoring nonignorable missing values only under high missingness rates (Pohl et al., 2014; Rose, 2013; Rose et al., 2010), such that the application of the SA+Q might be useful mainly under conditions with higher rates of NRIs.

Due to the model's complexity, we recommend keeping measurement models as simple as possible, e.g., by employing a Rasch model for item responses and/or fixing time discrimination parameters in the measurement model of RTs to be equal across items. More complex measurement models that might better fit the data at hand can be incorporated in the SA+Q framework. For RTs, for instance, alternative parameterizations have been suggested assuming distributions of RTs different than lognormal (Klein Entink, van der Linden, & Fox, 2009) or introducing additional parameters that reflect the way an item distinguishes between examinees of different speed levels (Klein Entink, Kuhn, Hornke, & Fox, 2009). Note, however, that adding additional model complexity by choosing more complex measurement models for

responses and/or RTs might further challenge estimation of the SA+Q model and increase the number of items or examinees needed to achieve convergence as well as unbiased and efficient parameter estimation.

3.6.1 Limitations and Future Research

While the SA+Q model allows to incorporate different missingness mechanisms, it heavily relies on extrapolation of the distribution of the number of reached items before quitting and the underlying test endurance variable. When the majority of examinees manages to reach the end of the test, the number of reached items before quitting is strongly affected by right censoring. Thus, as it is the case with previously developed latent model-based approaches for NRIs (e.g., Glas & Pimentel, 2008), the distribution of reached items before quitting is extrapolated from observations assumed to belong to lower quartiles of the distribution. This renders it difficult to assess whether distributional assumptions are met. In the simulation study for evaluating estimability, rather low missingness rates due to quitting went back to a relatively high proportion of examinees exhibiting such behavior. For instance, under conditions with a missingness rate of 10% due to quitting, quitting behavior was observable for roughly a quarter of examinees. This, however, must not be the case in real data. It might well be that high NRI rates can be attributed to few examinees quitting the assessment at early stages. The capability of the SA+Q model to yield an accurate description of the quitting process under such conditions remains to be evaluated. Likewise, it still remains to be assessed whether statistical performance of the SA+Q model can be improved under conditions with large sample sizes, as often encountered on the country level in LSAs. It might well be that the SA+Q model performs well under conditions with a low proportion of examinees exhibiting quitting behavior when the absolute number of examinees is sufficiently high. Until then, the requirements concerning missingness rates due to quitting established on the basis of the simulation study may serve as lower boundaries. In addition, for count data on the test level, distributions different from lognormal, such as a gamma distribution, have been suggested for the random, person-specific mean of the Poisson distribution (Doebler & Holling, 2016; Jansen, 1994). Further research and possibly also experimental studies are needed to evaluate the appropriateness of different distributional assumptions for test endurance.

Furthermore, it remains to reason whether the quitting process can sufficiently be described by only considering item positions. First, it might well be that rather than the number of attempted items, passed test time better measures test endurance. As long as time intensities do not show large differences across items, item positions

might serve as a proxy for passed test time (Fox & Marianti, 2016). Otherwise, the SA+Q model might not be able to sufficiently capture the quitting process. Second, subpopulations with different quitting strategies might exist. For instance, while some examinees might quit due to lack of motivation, others might quit when they consider the test too difficult or even out of frustration when noticing that the time remaining is not sufficient to complete the assessment. Likewise, there might be qualitative differences in quitting mechanisms between examinees who quit at earlier and later stages of the assessment. To meaningfully incorporate different forms of quitting behavior into the model, further substantive research is needed to describe and understand these mechanisms.

Moreover, the SA+Q model assumes stationarity of speed. This assumption is reasonable when tests are administered with generous time limits, such that examinees are unlikely to run out of time (van der Linden, 2007). However, the presence of NRIs due to lack of speed indicates that the allocated time might not have been sufficient for all examinees. When examinees perceive their current speed level as being insufficient to reach the end of the test, they might try to adjust their pace (Yamamoto & Everson, 1997), rendering the stationarity assumption implausible. In future studies it seems therefore necessary to allow for within-person variation of speed in the SA+Q model (see Fox & Marianti, 2016, for an extended model).

In general, estimation of the SA+Q model was found to be rather challenging. Severe convergence issues were encountered under conditions with less than 10% missingness due to quitting. When convergence was reached, this oftentimes was only the case with high numbers of iterations. Future research should therefore address facilitating the estimation procedure of the SA+Q model.

In the empirical example, we encountered conditions that pose further challenges to adequate modeling of NRIs and provided heuristic solutions to address these issues. Future research is needed on how to better deal with a) the fact that oftentimes no RT information is available for the last item seen when no response has been generated as well as b) varying enforcement of time limits when applying the SA+Q framework. In addition, in order for the SA+Q model to be readily applicable to empirical data, it needs to be considered that NRIs are often not the only source of missingness going back to examinee behavior. In most LSAs, both NRIs as well as item omissions can be encountered (Pohl et al., 2014). In order to additionally handle (nonignorable) omission processes, the SA+Q model could be combined with model-based approaches for item omissions (e.g., O'Muircheartaigh & Moustaki, 1999; Rose, 2013; Ulitzsch et al., 2019c).

4

Using Response Times for Joint Modeling of Response and Omission Behavior

This chapter is published as Ulitzsch, E., von Davier, M., & Pohl, S. (2019c). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*. doi:10.1080/00273171.2019.1643699.

For adequate modeling of missing responses, a thorough understanding of the nonresponse mechanisms is vital. As a large number of major testing programs are in the process or already have been moving to computer-based assessment, a rich body of additional data on examinee behavior becomes easily accessible. These additional data may contain valuable information on the processes associated with nonresponse. Bringing together research on item omissions with approaches for modeling response time data, we propose a framework for simultaneously modeling response behavior and omission behavior utilizing timing information for both. As such, the proposed model allows a) to gain a deeper understanding of response and nonresponse behavior in general and, in particular, of the processes underlying item omissions in LSAs, b) to model the processes determining the time examinees require to generate a response or to omit an item, and c) to account for nonignorable item omissions. Parameter recovery of the proposed model is studied within a simulation study. An illustration of the model by means of an application to real data is provided.

5

A Hierarchical Latent Response Model for Inferences about Examinee Engagement in Terms of Guessing and Item-Level Nonresponse

This chapter is published as 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*. doi:10.1111/bmsp.12188.

In low-stakes assessments, test performance comes with little or no consequences for examinees themselves, so that examinees may not be fully engaged when answering the items: Instead of engaging in solution behavior, disengaged examinees might randomly guess or generate no response at all. When ignored, examinee disengagement poses a severe threat to the validity of results obtained from low-stakes assessments. Statistical modeling approaches in educational measurement have been proposed that account for nonresponse or for guessing, but do not consider both types of disengaged behavior simultaneously. We bring together research on modeling examinee engagement and research on missing values and present a hierarchical latent response model for identifying and modeling the processes associated with examinee disengagement jointly with the processes associated with engaged responses. To that end, we employ a mixture model that identifies disengagement on the item-by-examinee level by assuming different data-generating processes underlying item responses and omissions, respectively, as well as response times associated with engaged and disengaged behavior. By modeling examinee engagement with a latent response framework, the model allows assessing how examinee engagement relates to ability and speed as well as identifying items that are likely to evoke disengaged test-taking behavior. An illustration of the model by means of an application to real data is presented.

6
Discussion

Understanding the occurrence of missing responses is of utmost importance throughout the whole process of LSA operations, ranging from test construction through administration to the analysis and interpretation of results. The current work developed, evaluated, and applied model-based frameworks that utilize RTs for a) facilitating a better understanding of whether and how examinees differ when interacting with a test in general and the mechanisms underlying missing responses in particular, b) considering these differences when analyzing LSA data, and c) better handling missing values due to examinee behavior by modeling the mechanisms underlying their occurrence. For doing so, the present work brought together research on modeling item omissions and NRIs with research utilizing RTs when modeling observed responses.

Four approaches were presented that focus on different aspects and mechanisms underlying missing responses. Chapters 2 and 3 focused on modeling the processes underlying NRIs. While Chapter 2 delineated and evaluated the capability of van der Linden's (2007) SA framework for modeling the mechanism underlying NRIs due to lack of speed, Chapter 3 built on that work and aimed at disentangling and jointly modeling multiple mechanisms underlying NRIs, namely lack of speed and quitting. In Chapters 4 and 5, model-based approaches for modeling omission behavior were provided. Chapter 4 focused on jointly modeling omission behavior and response behavior, thus providing a better understanding of how these two types of behavior differ. Chapter 5 built on previous theoretical work relating item omissions to examinee disengagement and provided a model-based approach that allows for identifying and modeling examinee disengagement in terms of both omission and guessing behavior.

In the following, a holistic view on the presented approaches is taken and advantages of considering RTs for modeling and understanding missing responses in LSAs as well as limitations along with directions for future research are discussed. Recommendations for practitioners aiming at applying the presented frameworks are given. Last, placing the present work in a broader perspective, it is further

discussed how its approaches contribute to "unpacking" and better understanding examinee performance. Based on these considerations, implications for analyzing and reporting LSA data are derived.

6.1 Advantages of Using Response Times for Modeling Missing Responses

In the present work, it has repeatedly been shown that utilizing RTs for modeling missingness mechanisms comes with strong advantages. First and foremost, considering the additional information contained in RTs when modeling missing responses supports a more nuanced way of modeling and assessing missingness mechanisms. Second, understanding the occurrence of missing responses provides a better understanding of examinee test-taking behavior in general. Third, from a measurement perspective, considering additional information on how examinees interacted with the assessment provides less biased and more reliable parameter estimates. These advantages are evident from all models presented and shall be reviewed more closely in the following.

6.1.1 Modeling and Understanding Missingness Mechanisms

All frameworks presented aim at modeling and thus understanding missingness mechanisms. By leveraging and modeling the additional information on test-taking behavior contained in RTs, the presented frameworks allow for a more nuanced modeling of missingness mechanisms as compared to previous approaches for handling NRIs and item omissions. In contrast to the manifest model-based approach for NRIs by Rose et al. (2010), the frameworks for modeling NRIs presented in Chapters 2 and 3 allow considering differences in test-taking behavior (i.e., speed and test endurance) for all examinees – not just for those who did not reach the end of the test. In addition, the speed-accuracy+quitting (SA+Q) framework presented in Chapter 3 allows for disentangling and modeling NRIs due to speed and quitting simultaneously.

As compared to rather heuristic RT-based scoring approaches as well as approaches for modeling omissions based on information retrievable from paper-and-pencil-based assessment, the speed-accuracy+omission (SA+O) and speed-accuracy+engagement (SA+E) frameworks for modeling omission and/or guessing behavior presented in Chapters 4 and 5 provide richer, more nuanced models of the mechanisms underlying omissions and/or guessing. In addition, the SA+E framework presented in Chapter 5 allows considering and assessing omission and guessing behavior simultaneously.

6.1.2 Investigating Differences in Test-Taking Behavior

Above and beyond modeling and assessing mechanisms underlying missing responses, an important aspect of all frameworks presented is the opportunity to provide a general understanding of how examinees interact with assessments. This is achieved by embracing the potential of missing responses as a rich source of information on how examinees interact with the assessment instead of considering missing responses as a mere nuisance need to be dealt with.

The insights into test-taking behavior that can be gained based on the frameworks presented throughout Chapters 2 to 5 have been illustrated using data from two major LSAs – PIAAC and PISA. From the empirical examples it became evident that all types of test-taking behavior considered in the present work can be encountered to a considerable degree in LSAs and that, at the same time, examinees with different levels of ability differ in the tendency to show such behavior. Employing data from the Canadian and Spanish sample from PISA 2015, both Chapter 2 and 3 could illustrate that within the same data set, both lack of speed and quitting pose potential mechanisms underlying NRIs. In the empirical example in Chapter 3, examinees with higher ability showed both lower levels of general working speed and lower test endurance, indicating that these examinees were more likely to run out of time or to quit the assessment at earlier stages. The empirical example in Chapter 4 illustrated that examinees operate on different speed levels when generating responses and omitting items, indicating different underlying processes. This conclusion could be drawn based on findings for the Chilean sample from PIAAC 2012, where the two speed levels associated with responses and omissions, respectively, did not reveal a perfect linear relationship to each other and yielded different correlations with ability and omission propensity. In data from the Austrian sample from PISA 2015 examined in Chapter 5, only approximately four fifth of responses were classified as engaged, while the remaining responses were either omitted or guessed. Disengaged examinees were found to display lower levels of effective ability. These results suggest that less able examinees are highly at risk to show disengaged behavior is LSA. In addition, the models presented in Chapters 4 and 5 allow assessing item characteristics associated with omission and guessing behavior. In the empirical examples, difficulty, time intensity, and open-response formats were item characteristics associated with items more likely to evoke omission and guessing behavior.

By providing insights into how examinees interact with assessments, the presented approaches provide tools for informing and evaluating substantive theories on examinee test-taking behavior. It should, however, be noted that in the empirical

examples only subsets of items and examinees were considered. Since the empirical examples served illustrative purposes, subsets with high incidence of item omissions and NRIs were chosen. For the sake of simplicity, analyses were based on single item blocks or testlets rather than the whole assessment. As such, the empirical examples should be seen as illustrations of the insights that can be gained on the basis of the presented frameworks as well as providing guidance on how to interpret the frameworks' model parameters. For evaluating and informing theories on test-taking behavior, more thorough investigations of LSA data are needed.

6.1.3 Enhancing Ability and Item Parameter Estimation

Even under conditions where, from a substantive perspective, investigating test-taking behavior is not of major interest in itself, the presented frameworks bear great advantages from a measurement perspective. By considering additional information on how examinees interacted with the assessment, all frameworks presented in this work support retrieving less biased and more efficient person and item parameter estimates. This has been illustrated by showing the impact of considering additional information about examinee test-taking behavior on ability and item parameter estimation. For doing so, the presented frameworks were compared to models that either do not consider or make stronger assumptions concerning the specific behavior under consideration. It could be shown that a) not considering the types of test-taking behavior under consideration as well as b) ignoring that examinees with different levels of ability differ in test-taking behavior can induce bias to ability and/or item parameter estimates. In addition, it has been illustrated how considering additional information on test-taking behavior yields more reliable ability estimates – especially for examinees with higher rates of missing responses.

6.2 Limitations and Directions for Future Research

Although the presented approaches pose advanced methods for modeling the mechanisms underlying missing responses in LSAs, multiple issues need to be addressed before the presented approaches are readily applicable under real-life LSA conditions. First, there is a strong need for an integrated approach, considering both omissions and NRIs simultaneously. Second, all frameworks come with assumptions on examinee test-taking behavior that might not always be met by empirical data. Out of the assumptions inherent to all frameworks presented, the neglect of a) qualitative differences in test-taking behavior across examinees as well as b) varying behavior across the test pose the most pertinent ones to be addressed. Third, future research should aim at validating and getting a better understanding of the constructs being

captured by the presented approaches. Fourth, for the frameworks to be applicable under real-life LSA conditions, challenges for model application encountered under such conditions need to be addressed. In addition, the presented frameworks opened up avenues for a myriad of future directions. Among the most promising extensions of the scope of application of the presented frameworks are a) considering additional data on test-taking behavior as well as b) adjusting the presented frameworks for modeling missing responses in noncognitive assessments. In the following, possible points of departure and opportunities for addressing these issues in future research will be discussed.

6.2.1 Modeling Omissions and Not-Reached Items Jointly

For providing model-based approaches for modeling missing responses, different missingness mechanisms were considered separately in this work. In data stemming from LSAs, however, it is likely that different types of missing responses occur within the same data set. Hence, there is strong need for an integrated framework allowing for considering mechanisms underlying omitted items and NRIs simultaneously.

At present, the approach by Rose (2013) is the only model-based approach that allows simultaneously modeling nonignorable item omissions and NRIs. Yet, this approach relies solely on information retrievable from paper-and-pencil-based assessment, and does, as such, not consider examinee time allocation strategies and quitting behavior. By considering RTs, the presented approaches pose more sophisticated methods for modeling item omissions and NRIs, although, so far, these frameworks only allow considering item omissions and NRIs separately. Building on the presented approaches with the objective of modeling mechanisms underlying omissions and NRIs simultaneously is therefore a fruitful topic for future research.

Challenges for such an integrated framework are a) modeling NRIs due to lack of speed in the context of different pacing behavior associated with omission and/or guessing behavior and engaged responses as well as b) increased model complexity when jointly modeling omission and/or guessing behavior and quitting.

As has been shown in the empirical examples in Chapters 4 and 5, examinees tend to require different amounts of time to generate engaged responses and to omit and/or rapidly guess. Hence, in the presence of omission and guessing behavior, the mechanism underlying NRIs due to unfavorable time allocation strategies is rather complex and poses a combination of different aspects of test-taking behavior. The SA+O and SA+E frameworks presented in Chapters 4 and 5 consider different time allocation strategies associated with engaged and disengaged responses as well as omissions along with the probability of showing omission or guessing behavior. It

therefore should be possible to employ these frameworks to model the processes underlying unfavorable time allocation strategies that lead examinees to not reach the end of the test. A systematic investigation, however, is still pending.

Considering quitting, omission, and/or guessing behavior simultaneously could be achieved by combining the SA+Q model with either the SA+O or the SA+E model. Such model combinations should be based on assumptions concerning omission and response processes: In the case that all responses are assumed to reflect the level of examinee ability, it is recommended to extend the SA+Q model by examinee omission propensity and omission speed as defined in the SA+O model for the purpose of modeling omission and quitting behavior simultaneously. In the case that guessing behavior is likely to have occurred and given that researchers have reason to assume that omission and guessing behavior stem from similar processes, it is recommended to combine the SA+Q framework with the SA+E model. Although combining the approaches presented in this work is technically straightforward, this results in rather complex, high-dimensional models. Assessing conditions under which such models perform well is still open for investigation.

6.2.2 Identifying Subpopulations Differing in Test-Taking Behavior

All frameworks presented in this work aim at modeling differences in how examinees interact with the assessment. When doing so, such differences are assumed to stem from quantitative differences in speed, omission propensity, engagement, and test endurance. It might, however, well be that examinees stem from subpopulations that qualitatively differ in their test-taking strategies. For further model developments, this implies that rather than locating examinees on latent continua, distinct subpopulations qualitatively differing in how they approach the test need to be identified.

Concerning item omissions (and guessing behavior), such subpopulations might be examinees omitting items (and/or guessing) due to being disengaged and examinees showing such behavior for, e.g., test-strategic reasons to maximize their score. Mixture extensions of the SA+O or the SA+E model would allow for detecting such subpopulations. In addition, when asked for explaining their motives for omitting items, examinees oftentimes name different reasons for item omissions, such as lack of confidence in the correct answer, fatigue, or lack of motivation (Jakwerth et al., 2003; OECD, 2013; van Barneveld et al., 2013). It is thus possible that in the course of the assessment, the same examinee might omit items for different reasons. Extending the SA+E model by allowing for different types of omissions, i.e., jointly considering

disengaged and engaged omissions, might support detecting and modeling such processes.

Concerning quitting behavior, rather than locating all examinees on a single test endurance continuum and extrapolating the level of test endurance for those who did not quit the assessment, it might be fruitful to conceptualize quitters as a subpopulation of examinees qualitatively differing in how they approach the assessment from examinees willing to complete the assessment. In addition, different reasons might underlie quitting behavior, demanding a finer-grained analysis of the group of quitters. Indeed, when asked about their motives to quit, examinees state different reasons, ranging from feeling overtaxed with the assessment to lack of motivation to complete it (OECD, 2013).

6.2.3 Allowing for Varying Test-Taking Behavior Across the Test

All frameworks assume stationarity of the test-taking behavior considered. That is, it is assumed that examinees operate with constant ability, speed, omission speed, and engagement throughout the assessment. This assumption might not hold under the test conditions considered in this work: When faced with time limits, examinees might adjust their pace to reach the end of the test, e.g., by increasing their speed at the expense of lower accuracy or omitting more items and taking less time for their decision to do so. Likewise, in low-stakes settings, fatigue effects might result in lower engagement towards the end of the test. While recent extensions of the SA model allow accounting for varying working speed across the test (Fox & Marianti, 2016), these do not incorporate a model component for adjustments in the level of effective ability accompanying varying levels of speed. Still, the approach presented by Fox and Marianti (2016) is a possible point of departure for tackling the stationarity assumptions incorporated in the frameworks presented in this work.

6.2.4 Model Validation

The frameworks presented in this work introduce additional person variables assumed to underlie the occurrence of NRIs, item omissions as well as rapid guesses and aim at providing a depiction of how examinees interact with assessments. In this context, two issues are of pertinent importance: First, there is strong need to validate the newly introduced person variables. Second, it needs to be addressed whether the frameworks considered in this work capture constructs that are of relevance for real-life behavior.

All frameworks presented in this work entail interpretations of the constructs being captured by newly introduced person variables. This is especially evident in the SA+E and the SA+Q frameworks in which the newly introduced person variables are interpreted as representing examinee engagement and test endurance. However, whether or not such interpretation is justified might be highly context-dependent and is still open for investigation. For validating the newly introduced person variables, self-reports on test-taking behavior and motivation can be of great use. For instance, while in the SA+E framework guessing and omission behavior are assumed to represent disengaged test-taking behavior, the model might also capture processes different from disengagement such as test-taking strategies aimed at maximizing test scores by guessing or omitting. Assessing how examinee disengagement as identified by the model relates to self-reports on effort and/or test-taking motivation might therefore provide further insight into the construct being measured by the SA+E model (Ulitzsch, Penk, von Davier, & Pohl, manuscript in preparation). Likewise, relating test endurance to self-reports on why the assessment has been quit might provide further insight into how to interpret the test endurance variable introduced in the SA+Q framework.

Once sound knowledge on the constructs being captured by the presented frameworks is established, these may be of great value for addressing the objectives of LSAs: Major LSAs such as PISA or PIAAC aim at assessing competencies necessary for professional, social, and civic life (OECD, 2013). Usually, (effective) ability is employed as a predictor of such external criteria. Yet, research has recently started to also consider test-taking behavior as an additional predictor, with the rationale being that test-taking behavior provides a measure of noncognitive skills based on real-life behavior (Balart, Oosterveen, & Webbink, 2018; Hitt, Trivitt, & Cheng, 2016; Zamarro, Cheng, Shakeel, & Hitt, 2018; Zamarro, Hitt, & Mendez, 2016). However, although such studies could show that test-taking behavior predicts important real-life outcomes both on the individual as well as on the country level, they also pointed out that a closer investigation of the noncognitive skills captured in test-taking behavior is needed. Addressing both the specific noncognitive skill captured by the presented frameworks as well as how the aspects of test-taking behavior considered in this work relate to real-life outcomes therefore pose important topics for future research.

6.2.5 Dealing with Operational Challenges of Large-Scale Assessments

Currently, the presented frameworks might not always be feasible for application under operational LSA settings. First, estimation of all frameworks presented was rather time consuming. The computational burden would certainly be exacerbated in

operational practice in LSAs due to a) substantially larger data sets than considered in the present work (e.g., some 500,000 examinees in PISA 2015, OECD, 2017) as well as b) even higher model complexity due to additionally considering background variables. Further research in Bayesian estimation algorithms as well as technical advances could bring relief concerning said computational burden.

Second, estimation of the presented frameworks was conducted employing Bayesian estimation. Currently, the frameworks of analysis implemented and software employed in major LSAs rely on maximum likelihood estimation. While estimating the presented approaches with maximum likelihood might reduce computational burden to some degree, the high-dimensionality of the presented models as well as data sparseness on missingness and quitting indicators might pose challenges for estimation that still need to be evaluated.

Third, research objectives employing LSA data oftentimes involve multiple groups or trajectories of competencies over time. To address such research questions adequately, measurement invariance needs to be established. For the proposed approaches, this assumption is not trivial since tests are usually not designed in order to similarly evoke omission, guessing, or quitting behavior as well as to be equally time intense across groups and time.

6.2.6 Considering Additional Data on Test-Taking Behavior

All frameworks presented in this work leverage the rich information on examinee behavior provided by RTs. While these allow researchers to get a better understanding of examinee behavior by providing information on *how long* examinees interacted with an item until generating a response or deciding to omit it, computer-based LSAs also provide log data containing information on *how* examinees interacted with items. In its public database, for instance, PISA provides information on the number of actions (that is, clicks, double clicks, key presses and drag/drop events) for each item-by-examinee interaction. Furthermore, in addition to the total RT per item, the first time to action, defined as the time between the first showing of the item and the first action recorded for the item, is provided (OECD, 2017). Oftentimes, for more complex tasks, e.g. simulations as implemented in PIAAC or tasks on collaborative problem solving as implemented in PISA, action sequences are available. Analyzing such data in addition to observed final responses may facilitate understanding how examinees arrive at a certain response by understanding "how individuals plan, evaluate, and select operations" (He & von Davier, 2016, p. 72). This can be achieved by, e.g., assessing whether examinees interact systemically and efficiently with items (Zhu, Shu, & von Davier, 2016) or by identifying subpopulations that differ in how

they arrive at a response (Greiff, Molnár, Martin, Zimmermann, & Csapó, 2018; Greiff, Wüstenberg, & Avvisati, 2015).

Combining approaches considering additional information on the number of actions (e.g., De Boeck & Scalise, 2019) as well as action sequences (e.g., Greiff et al., 2018; He & von Davier, 2016; Zhu et al., 2016) with frameworks utilizing RTs for modeling missing responses is a highly promising topic for future research. Considering this additional information has high potential for further, in-depth insight into omission mechanisms. A possible research direction could be to disentangle different omission processes by identifying different patterns of examinee-by-item interactions. For instance, while omissions with short RTs and no interactions might point towards disengaged omission behavior, omissions with longer RTs and interaction patterns resembling those encountered on (incorrect) observed responses might indicate that examinees tried to solve the item but due to, e.g., lack of confidence in the correct answer decided to omit it. Likewise, for disentangling different mechanisms underlying quitting behavior, subgroups of examinees could be identified that differ in how they interact with the items they respond to and, as such, might also differ in whether, when, and for what reason they quit the assessment.

6.2.7 Modeling Missingness Mechanisms in Noncognitive Assessments

In LSAs, missing values due to examinee behavior do not only occur in the cognitive assessment but are also encountered on the noncognitive background questionnaires. In PIAAC 2012 Cyprus, for instance, item-level missingness rates on the background questionnaire were as high as 18% (OECD, 2013). Adjusting the models presented in this work to suit the characteristics of noncognitive assessments can be a powerful tool for understanding examinees' interaction with questionnaires and the occurrence of missing responses in particular.

A possible starting point for adjustments is, for instance, the fact that in the literature, RTs on questionnaire items are perceived as representing the difficulty of endorsing an item. That is, in addition to the examinee's speed of responding and the item's time intensity, RTs are assumed to be governed by the distance between the examinee's trait level and the item location. This corresponds to the assumption that examinees who either strongly agree or disagree with a statement can express this belief rather quickly, while examinees for whom it is difficult to decide whether or not to endorse a statement need more time for their decision (distance-difficulty hypothesis, see, e.g., Ferrando & Lorenzo-Seva, 2007; Kuncel & Fiske, 1974). Building on van der Linden's SA model, Ferrando and Lorenzo-Seva (2007) provided a framework integrating these theoretical considerations when

modeling RTs from noncognitive assessments that could be combined with the frameworks presented in this work.

Once such adjusted frameworks are available, relating behavior in the cognitive assessment to behavior on the noncognitive assessment poses a highly promising topic for future research. For instance, theoretical work on examinee engagement has emphasized the possibility that it might well be that examinees approaching the cognitive assessment disengagedly might also be more prone to interact carelessly with the subsequent noncognitive questionnnaires (Wise, 2015). As such, it has been suggested to employ omission rates on background questionnaires as indicators of disengaged behavior on both the background questionnaire as well as the preceding cognitive assessment (Boe et al., 2002; Zamarro et al., 2016). Assessing disengagement in noncognitive assessments employing adjusted versions of the SA+E model and relating disengaged behavior on cognitive and noncognitive assessments would provide the basis for further elaborating on that hypothesis as well as provide further insights in the stability of examinee behavior across different types of assessments.

6.3 Recommendations for Model Application

To ensure that the presented models give reasonable estimates under conditions typically encountered in LSAs, the statistical performance of the presented models has been evaluated in comprehensive simulation studies. From these simulation studies, the following recommendations can be derived: With sample sizes of at least $N \geqslant 500$ and a test length of at least $K \geqslant 10$, the SA+O and SA+E models give reasonable estimates under conditions with at least 5% omissions and disengagement rates of 10%, respectively. The SA+Q model needs longer tests of at least $K \geqslant 40$ items and missingness rates due to quitting of 5% (or approximately 15% of examinees exhibiting quitting behavior) to give reasonable estimates. However, since the SA+Q model heavily relies on extrapolating the distribution of test endurance, the model's assumptions might only be justified when the majority of examinees quit the assessment. For conditions with few examinees and/or items, higher missingness rates are needed to estimate the models presented in Chapters 3 to 5. Chapters 4 and 5 provided guidelines for model checking that can easily be applied to the other frameworks presented in this work.

Concerning the choice between the two approaches for NRIs and omissions, respectively, the following recommendations are given: In the presence of NRIs, it is recommended to assess NRIs jointly with cumulative RTs and check whether examinees exhibited quitting behavior. If so, application of the SA+Q for considering quitting behavior is recommended. Otherwise, the SA model is sufficient for

modeling NRIs. For modeling omissions, choosing between the SA+O and SA+E model can be based on theoretical considerations concerning possible mechanisms underlying omissions and observed responses as well as model comparisons.

Note that all presented frameworks are extensions of the SA model and, as such, retain its flexibility. This means that different "plug-ins" for the component models can be implemented, such as different measurement models for item responses, models with different distributional assumptions for RTs (e.g. Klein Entink, van der Linden, & Fox, 2009), or different theoretical considerations on item parameters, i.e., perceiving these as either fixed or random effects (De Boeck, 2008). Likewise, further developments of the SA model as well as model-based approaches for missing responses can be integrated with the presented frameworks, such as hypotheses on varying speed (Fox & Marianti, 2016) or multidimensional omission processes (Köhler et al., 2015b). When doing so, it should be kept in mind that the presented frameworks are rather complex and adding further model components might challenge estimation.

6.4 Implications

The approaches proposed in this work allow modeling, assessing, and explicating differences in important aspects of examinee test-taking behavior and support insights into examinee and item characteristics associated with the occurrence of item omissions, disengaged guessing behavior, and NRIs. In the following, the potential of these properties for improving LSA operations will be discussed. It is argued that the approaches proposed in this work can be of great utility across all stages of LSA operations, ranging from test construction and administration to analysis and reporting.

6.4.1 Implications for Test Construction

The proposed approaches can be of great value in test construction as they provide the opportunity for diagnosing the appropriateness of time limits as well as for identifying possible issues with items associated with an increased incidence of omissions, rapid guesses, or quitting behavior. First, as noted by van der Linden (2011b), knowledge gained on the basis of the SA framework (and extensions thereof) can be utilized to exercise control over the probability of examinees running out of time and, as such, on the probability of observing NRIs due to lack of speed. In the case of undesirably high rates of NRIs due to lack of speed, knowledge on the items' time intensities of a given assessment on the one hand along with knowledge on the

level of speed examinees operate on the other can be utilized for adjusting the time limit or shortening the test (see van der Linden, 2011b).

Second, while reasons for aberrant behavior such as omitting items, guessing, or quitting the assessment altogether are manifold, such behavior might point towards problems with the specific task or wording of items associated with an increased incidence of such behavior. For instance, in the empirical example in Chapter 5, one item was especially prone to evoke disengaged test-taking behavior, with an omission rate of approximately 35% and an additional 15% of item-by-examinee interactions being classified as perfunctory answers. Likewise, in the data application in Chapter 3, some items were much stronger associated with quitting behavior than the preceding or subsequent items. Such results do not necessarily imply problems with the respective items and might very well go back to examinee test-taking strategies, motivation, or "aversion" to, e.g., items with specific formats or contents. Yet, such results might yield test constructors to re-evaluate the respective items and check whether these are hard to understand or too complex, with the consequence being that examinees interact with these items differently as intended by test constructors. Likewise, pre-field studies can be conducted to identify test conditions and features preventing quitting and disengaged test-taking behavior for both the target population or subgroups such as examinees with special needs. The assessment may then be adjusted accordingly. For examinees with special needs, such pre-field studies can be of great utility for identifying test conditions and features posing an undue burden and evoking quitting or disengaged test-taking behavior. Such studies can then facilitate choosing suitable accommodations for examinees with special needs.

6.4.2 Implications for Test Administration

Above and beyond improving items and testing conditions prior to test administration, the presented frameworks may also be employed to influence change of unwanted test-taking behavior during the assessment. This could be achieved by monitoring test-taking behavior by means of real-time estimation of, e.g., omission propensity, and/or engagement and issuing warnings to examinees once pre-defined thresholds of acceptable aberrances in test-taking behavior are exceeded. Such warnings could encourage examinees to give their best or further clarify test instructions. Likewise, examinees could be encouraged to re-engage with the test directly after quitting. Wise, Bhola, and Yang (2006) have provided an example for monitoring systems employing heuristic RT-based scoring methods for identifying disengaged guesses. Their results are highly promising, showing that issuing warnings indeed

supports reducing disengaged guessing behavior and increases the validity of test scores.

6.4.3 Implications for Analysis of Large-Scale Assessment Data

Missing responses due to omitted items or NRIs force researchers to explicate their beliefs a) on the nature of the test-taking behavior underlying omitted items and NRIs as well as b) on how differences in that behavior should be considered when assessing differences in examinee performance. These beliefs become evident in researchers' decision on how to deal with missing responses in data analysis. In Chapter 1, assumptions and limitations of approaches currently implemented in operational LSA settings were thoroughly discussed. Chapters 2 to 5 provided frameworks for overcoming these limitations and thus improving the handling of missing responses in data analysis.

Currently, in most LSAs missing responses are either ignored or scored as (partially) incorrect. Ignoring missing responses due to item omissions and NRIs entails the assumption that missing responses are not informative in the sense that the processes underlying such missing responses are not related to ability. In the present work, it has repeatedly been shown that this assumption is highly likely to be violated and that violations of this assumption heavily impact ability estimation. In addition, as delineated in Chapter 2, estimating ability solely based on responses given to reached items without considering speed and ignoring missing responses occurring due to lack of speed disadvantages examinees who, on the expense of lower accuracy, worked with a speed level sufficient to reach the end of the test. Conversely, this practice advantages examinees who, by working with a slower pace, did not reach the end of the test and had more time available for the items they succeeded to attempt in the given time.

Scoring NRIs as incorrect penalizes examinees who did not allocate their time such that they could reach the end of the test or who were not persistent enough to complete all items administered (Rohwer, 2013). However, as delineated in Chapter 2, this confounds differences in test-taking behavior and effective ability since low test scores might stem from rather different scenarios. For instance, examinees might either have shown low ability but worked sufficiently fast to finish the test on time, or they might not have reached the end of the test due to working too slowly or quitting but displayed higher levels of ability on approached items. When scoring omissions as incorrect, it is assumed that examinees omitted items because they did not know the answer (Rohwer, 2013). If this is not the case and examinees omitted items for different reasons, e.g., lack of motivation, fatigue, or refusal to participate,

scoring omitted items as incorrect, too, confounds differences in test-taking behavior and the achieved level of effective ability.

The presented frameworks overcome these limitations as they allow for considering differences in test-taking behavior in general and behavior underlying the occurrence of missing responses in particular when estimating ability. As such, the presented frameworks can critically improve the handling of missing responses in LSA data analysis.

6.4.4 Implications for Reporting on Results of Large-Scale Assessments

When examinees differ in test-taking behavior, it is highly possible that observed differences in performance do not only stem from differences in competencies but also from differences in how examinees approached the assessment. Based on data from major LSAs, it could be shown a) that in general, examinees indeed differ in the way they approach the assessment as well as b) that examinees with different levels of ability show different test-taking behavior. The results of the empirical illustrations of this work thus imply that observed examinee performance indeed results from the effective ability achieved at levels of effective speed, test endurance, omission propensity and/or engagement that differ across examinees.[1] In the present work, differences in test-taking behavior were assessed on the within-country level only. In addition, in LSAs there is strong variation in missingness rates and omission propensity as well as time spent on the assessment also on the country level (OECD, 2017; Sachse et al., 2019). This suggests that on the country level, too, levels of effective ability are achieved under test-taking behavior that systematically varies across countries.

Currently, such differences are left unconsidered in the reporting of LSA results. Rather, LSAs report results as rank tables consisting of single scores that confound differences in the level of competency with differences in test-taking behavior. In this work it is argued for taking a different, multidimensional perspective on performance when reporting on LSA results and explicitly considering behavioral aspects contributing to differences in observed performance. One way of such multidimensional reporting on test performance could be to report on a profile of different aspects of performance rather than merely on accuracy, as suggested by Pohl (March 2019) and Pohl, Ulitzsch, and von Davier (manuscript in preparation).

[1] As it is the case for differences in effective speed, it is plausible that the level of ability displayed by an examinee is related to the level of test endurance, omission propensity and/or engagement with which he or she approaches the assessment. For instance, examinees that approach the test disengagedly in the sense that they tend to omit and guess also might not display their best possible levels of ability on those items they answer according to their level of effective ability. The same might be true for examinees quitting the assessment.

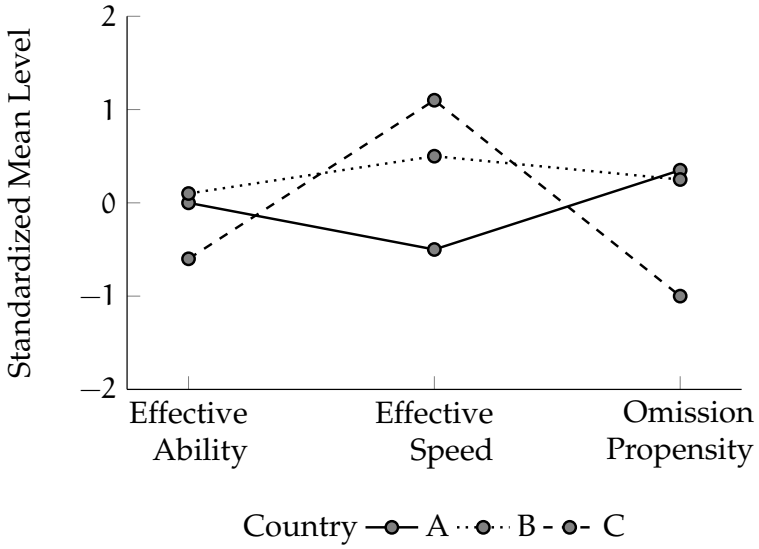

Figure 6.1. Schematic representation of profile-based reporting on effective ability, effective speed, and omission propensity for three hypothetical countries.

Following Pohl et al. (manuscript in preparation), Figure 6.1 illustrates this schematically by giving profiles of country-level effective ability, effective speed, and omission propensity, as defined in Chapter 4 for three hypothetic countries.[2] While, on average, examinees from Country A and B showed the same effective ability and interacted with the assessments with equal levels of omission propensity, examinees from Country B achieved the displayed level of effective ability at a higher speed level. That is, examinees from Country B were able to achieve the same level of accuracy in their responses faster and have as such showed better performance. Examinees from Country C showed a somewhat lower ability, however, achieved at much higher speed levels and omitted less items than examinees from Countries A and B. Depending on how researchers and policy-makers weigh the importance of effective ability, effective speed, and omission propensity when assessing performance, different conclusions might be drawn on Country C's overall performance: For instance, if the level of displayed ability is considered to be most important, regardless how fast responses were generated, examinees from Country C would be considered as performing worse than examinees from Countries A and B. However, if speed is considered to be an important aspect of performance, performance of examinees from Country C might be considered as equally well or even better as compared to the performance of examinees from Countries A and B. Similar considerations hold concerning whether and how omission behavior should be "punished" when assessing performance. Although such profile-based reporting does not allow to determine the level of effective ability that would have been displayed if examinees would have

[2]This figure is created based on preliminary findings from Pohl et al. (manuscript in preparation).

interacted with the assessment in the same manner, it disentangles different aspects contributing to examinee performance. As such, profile-based reporting informs consumers of LSA reports on differences in test-taking behavior as an important contributor to differences in test scores, allows explicating these differences, and thereby might encourage more cautious interpretations of LSA results.

For consumers of LSA reports, e.g., policy makers or the public, such a multidimensional perspective on performance provides the opportunity to get a more nuanced understanding of LSA results. First, a multidimensional perspective on performance supports investigating whether differences between groups as well as changes in performance across time go back to differences/changes in proficiencies or differences/changes in test-taking behavior (see Sachse et al., 2019).

Second, the more nuanced way of reporting entailed in profile-based reports would follow the demands on improving communication of LSA results of recent policy papers. Singer and Braun (2018, p. 39), for instance, have argued that as long as LSA results "are primarily reported as league tables, a mix of nationalism, fears about global competitiveness, and human nature inevitably lead policy-makers in countries with poor or declining performance toward unitary 'silver bullet' solutions based on highly aggregated data." A multidimensional perspective on performance that explicates differences in test-taking behavior related to observed differences in country-level performance would offer an alternative to reports based on rank tables that might foster more careful interpretations of LSA results, and as a consequence, more prudent derivations of policy measures.

Third, treating test-taking behavior as an additional aspect of performance rather than neglecting its pivotal importance for observing performance differences might be of great utility when assessing the predictive validity of performance on LSAs for real-world outcomes (see Pohl & von Davier, 2018). Differences in how examinees interact with the assessment might mirror important aspects of differences in real-life behavior. In fact, previous research suggests that test-taking behavior on low-stakes cognitive assessments and questionnaires such as omission behavior or careless answering on noncognitive items captures important noncognitive skills and predicts real-life outcomes. On the individual level, Hitt et al. (2016) have reported omission behavior on questionnaires to predict future educational attainment and income above and beyond cognitive ability. Likewise, based on data from PISA, Balart et al. (2018) have reported performance decline throughout the cognitive assessment to be negatively related to economic growth on the country level, even when controlling for performance at the beginning of the cognitive assessment. These results are highly promising as they evidence that taking a multidimensional perspective on

performance might be accompanied with an increased predictive validity of LSA results.

Fourth, once sound knowledge on the noncognitive skills captured by test-taking behavior is established, multidimensional reporting might support deriving more targeted policy measures based on LSA results. By providing the possibility to assess cognitive (as indicated by the displayed level of effective ability) and noncognitive aspects of performance (as indicated by different characteristics of test-taking behavior), profile-based reporting can support diagnosing whether low performance is a result of low competency or unfavorable test-taking behavior and thus allows policy makers to derive measures targeted at either the improvement of cognitive or noncognitive skills. In addition, further insight into the noncognitive skills captured by test-taking behavior may also inform the decision on which aspects of test-taking behavior are most relevant to be considered in profile-based reporting.

It should be noted that, since LSAs typically assess multiple countries on multiple competency domains, profile-based reports might be rather overwhelming for policy makers and consumers of LSA reports outside the scientific community. A simpler alternative might be to cluster countries into groups with similar profiles and report on group membership, or to develop a composite score weighing different aspects of test performance according to considerations on their importance (Pohl et al., manuscript in preparation).

6.5 Conclusion

Examinees differ in how they interact with assessments. In low-stakes LSAs, missing responses pose an evident kind of such differences. Understanding the underlying mechanisms is paramount for making appropriate decisions on how to deal with missing responses and drawing valid inferences on examinee proficiencies. This work brought together research on modeling missing responses with research on modeling RTs for the purpose of providing tools that allow for a nuanced modeling and understanding of test-taking behavior associated with the occurrence of missing responses in LSAs. The frameworks presented provide the opportunity a) to assess and account for the occurrence of missing responses as well as b) to explicate differences in examinee behavior that contribute to differences in performance in general and differences in the occurrence of missing responses in particular. Against this background, this work argues for a multidimensional perspective that explicates the different aspects contributing to examinee performance when analyzing, reporting, and communicating results from LSAs.

References

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*, *47*(3), 434–452. doi:10.1080/03057925.2017.1301399

Allen, N., Donoghue, J., & Schoeps, T. (2001). *The NAEP 1998 technical report*. Washington, DC: National Center for Education Statistics. NCES 2001-509.

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. Retrieved from https://arxiv.org/pdf/1408.4050.pdf

Baker, S. G., Fitzmaurice, G. M., Freedman, L. S., & Kramer, B. S. (2005). Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. *Biostatistics*, *7*(1), 29–40. doi:10.1093/biostatistics/kxi038

Balart, P., Oosterveen, M., & Webbink, D. (2018). Test scores, noncognitive skills and economic growth. *Economics of Education Review*, *63*, 134–153. doi:10.1016/j.econedurev.2017.12.004

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.

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462. doi:10.1007/BF03173192

Bhola, D. S. (1994). *An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error* (Doctoral dissertation, University of Nebraska-Lincoln). Retrieved from https://search.proquest.com/docview/304127164

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student task persistence in the third international mathematics and science study: A major source of acheievement differences at the national, classroom, and student levels.* University of Pennsylvania, Philadelphia. Center for Research and Evaluation in Social Policy. CRESP-RR-2002-TIMSS1.

Bolsinova, M., de Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126–1148. doi:10.1007/s11336-016-9537-6

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2016). Response moderation models for conditional dependence between response time and response accuracy. *British*

Journal of Mathematical and Statistical Psychology, *70*(2), 257–279. doi:10.1111/bmsp.12076

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, *113*(11), 2309–2344.

Cao, J. & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*(2), 209–230. doi:10.1007/S11336-007-9045-9

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). doi:10.18637/jss.v076.i01

Chen, H. H., von Davier, M., Yamamoto, K., & Kong, N. (2015). *Comparing data treatments on item-level nonresponse and their effects on data analysis of large-scale assessments: 2009 PISA study* (ETS Research Report No. RR-15-12). Educational Testing Service. Princeton, NJ. doi:10.1002/ets2.12059

Choe, E. M., Zhang, J., & Chang, H.-H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika*, *83*(3), 650–673. doi:10.1007/s11336-017-9596-3

Cosgrove, J. (2011). *Does student engagement explain performance on PISA? Comparisons of response patterns on the PISA tests across time*. Educational Research Centre. Dublin, Ireland. Retrieved from http://www.erc.ie/documents/engagement_and_performance_over_time.pdf

Culbertson, M. J. (April 2011). *Is it wrong? Handling missing responses in IRT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

Daniels, M. J. & Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, *94*(448), 1254–1263. doi:10.1080/01621459.1999.10473878

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. doi:10.1007/s11336-008-9092-x

De Boeck, P. & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*. doi:10.3389/fpsyg.2019.00102

De Boeck, P. & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology*, *10*, 1280. doi:10.3389/fpsyg.2019.01280

de Ayala, R., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*(3), 213–234. doi:10.1111/j.1745-3984.2001.tb01124.x

Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. doi:10.1111/jedm.12009

Debeer, D., Janssen, R., & Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, *54*(3), 333–363. doi:10.1111/jedm.12147

Doebler, A. & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson counts model. *Learning and individual differences*, *52*, 121–128. doi:10.1016/j.lindif.2015.01.013

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, *161*(3), 1307–1320.

Duchhardt, C. & Gerdes, A. (2012). *NEPS technical report for mathematics: Scaling results of starting cohort 3 in fifth grade (NEPS Working Paper No. 19)*. Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Erosheva, E. A. (2002). *Grade of membership and latent structure models with application to disability survey data* (Doctoral dissertation, Carnegie Mellon University). Retrieved from https://pdfs.semanticscholar.org/1fe4/64b6cae48%20d009697783bdbb72bcd4527608a.pdf

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. doi: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. doi:10.1111/j.1745-3984.2008.00062.x

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*(7), 1–14. doi:10.18637/jss.v020.i07

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. doi:10.1080/00273171.2016.1171128

Foy, P. (2017). *TIMSS 2015 user guide for the international database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Chestnut Hill, MA. Retrieved from https://timssandpirls.bc.edu/timss2015/international-database/downloads/T15_UserGuide.pdf

Foy, P. (2018). *PIRLS 2016 user guide for the international database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Chestnut Hill, MA. Retrieved from https://timssandpirls.bc.edu/pirls2016/international-database/downloads/P16_UserGuide.pdf

Frey, A., Spoden, C., Goldhammer, F., & Wenzel, S. F. C. (2018). Response time-based treatment of omitted responses in computer-based testing. *Behaviormetrika*, *45*(2), 505–526. doi:10.1007/s41237-018-0073-9

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models*. Cambridge University Press.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. doi:10.1214/ss/1177011136

Gelman, A. & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 163–174). Boca Raton, FL: Chapman Hall.

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling*, *57*(4), 523–541.

Glas, C. A. W. & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907–922. doi:10.1177/0013164408315262

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*(1), 65–87. doi: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*(3-4), 133–164. doi:10.1080/15366367.2015.1100020

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. doi:10.1177/0146621613517164

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers No. 133). OECD Publishing. Paris, France. doi:10.1787/19939019

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–6262. doi:10.1037/a0034716

Gonzalez, E. & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 3, pp. 125–156). Hamburg, Germany: IEA-ETS Research Institute.

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263. doi:10.1016/j.compedu.2018.07.013

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105. doi:10.1016/j.compedu.2015.10.018

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183. doi:10.1080/08957347.2016.1171766

Guo, J., Gabry, J., & Goodrich, B. (2018). *Rjags: R interface to Stan*. R package version 2.18.2. Retrieved from https://CRAN.R-project.org/package=rstan

He, Q. & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In A. L. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research. The 79th annual meeting of the psychometric society, Madison, Wisconsin, 2014* (pp. 750–777). New York, NY: Springer.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions). Cambridge, MA: National Bureau of Economic Research.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, *52*, 105–119. doi:10.1016/j.econedurev.2016.02.001

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Holman, R. & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. doi:10.1111/j.2044-8317.2005.tb00312.x

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). An investigation of why students do not respond to questions. In *NAEP validity studies*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Jansen, M. G. H. (1994). Parameters of the latent distribution in Rasch's Poisson counts model. In C. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 319–326). New York, NY: Springer.

Jansen, M. G. H. (1995). The Rasch Poisson counts model for incomplete data: An application of the EM algorithm. *Applied Psychological Measurement*, *19*(3), 291–302. doi:10.1177/014662169501900307

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, *52*(2), 93–100. doi:10.1080/00031305.1998.10480547

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. doi:10.1007/s11336-008-9075-y

Klein Entink, R. H., van der 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. doi:10.1348/000711008X374126

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. doi:10.1037/a0014877

Köhler, C., Pohl, S., & Carstensen, C. H. (2015a). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling*, *57*(4), 499–522.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015b). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*(5), 850–874. doi:10.1177/0013164414561785

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397–419. doi:10.1111/jedm.12154

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). Omitted and not-reached items in mathematics. In *1990 National Assessment of Educational Progress (CRE*

Technical Report No. 347). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Camebridge, MA: Academic Press.

Kuhn, J.-T. & Ranger, J. (2015). Measuring speed, ability, or motivation: A comment on Goldhammer (2015). *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 173–176. doi:10.1080/15366367.2015.1105065

Kuncel, R. B. & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement*, *34*(4), 743–755. doi:10.1177/00131644740.3400401

Kyllonen, P. & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4). doi:10.3390/jintelligence4040014

Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance* (ETS Research Report No. RR-93-49). Educational Testing Service. Princeton, NJ. doi:10.1002/j.2333-8504.1993.tb01560.x

Lee, Y.-H. (2007). *Contributions to the statistical analysis of item response time in educational testing* (Doctoral dissertation, Columbia University). Retrieved from https://search.proquest.com/docview/304858340?accountid=11004

Lee, Y.-H. & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Lee, Y.-H. & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*(1). doi:10.1186/s40536-014-0008-1

Lee, S. Y. (2018). *A mixture model approach to detect examinees with item preknowledge* (Doctoral dissertation, University of Wisconsin-Madison). Retrieved from https://search.proquest.com/docview/2068164818?accountid=11004

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. doi:10.1016/j.jmva.2009.04.008

List, M. K., Köller, O., & Nagy, G. (2019). A semiparametric approach for modeling not-reached items. *Educational and Psychological Measurement*, *79*(1), 170–199. doi:10.1177/0013164417749679

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134. doi:10.1080/01621459.1993.10594302

Little, R. J. & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79–94. doi:10.1080/10627197.2015.1028618

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247–264. doi:10.1007/BF02291471

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*(3), 477–482. doi:10.1007/BF02293689

Lüdtke, O., Robitzsch, A., & Wagner, J. (2018). More stable estimation of the STARTS model: A Bayesian approach using Markov chain Monte Carlo techniques. *Psychological Methods*, *23*(3), 570–593. doi:10.1037/met0000155

Mandinach, E. B., Bridgeman, B., Cahalan-Laitusis, C., & Trapani, C. (2005). *The impact of extended time on SAT® test performance* (Research Report No. 2005-8). The College Board. New York, NY. Retrieved from https://files.eric.ed.gov/fulltext/ED563027.pdf

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*(4), 523–547. doi:10.1007/BF02294327

Maris, G. & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. doi:10.1007/s11336-012-9288-y

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*(1), 1–27. doi:10.1111/jedm.12060

Meyer, J. P. (2010). A mixture rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538. doi:10.1177/0146621609355451

Mislevy, R. J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215. doi:10.1007/BF02295283

Mislevy, R. J. & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Report No. RR-96-30-ONR). Educational Testing Service. Princeton, NJ. doi:10.1002/j.2333-8504.1996.tb01708.x

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205–228. doi:10.1111/bmsp.12117

Molenaar, D. & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279–297. doi: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. doi:10.1080/00273171.2016.1192983

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). 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. doi:10.1111/bmsp.12042

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 39–51). Berlin, Germany: Springer.

Moustaki, I. & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *163*(3), 445–459. doi:10.1111/1467-985X.00177

Muthén, L. K. & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. doi:10.1207/S15328007SEM0904_8

National Center for Education Statistics. (2009, May 13). *NAEP technical documentation*. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 113–162). Boca Raton, FL: CRC Press.

OECD. (2009). *PISA 2006 technical report*. OECD Publishing. Paris, France. Retrieved from https://www.oecd.org/pisa/data/42025182.pdf

OECD. (2013). *Technical report of the survey of adult skills (PIAAC)*. OECD Publishing. Paris, France. Retrieved from https://www.oecd.org/skills/piaac/_TechnicalReport_17OCT13.pdf

OECD. (2014). *PISA 2012 technical report*. OECD Publishing. Paris, France. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2017). *PISA 2015 technical report*. OECD Publishing. Paris, France. Retrieved from https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

O'Muircheartaigh, C. & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal*

Statistical Society: Series A (Statistics in Society), *162*(2), 177–194. doi:10.1111/1467-985X.00129

Partchev, I. & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32. doi:10.1016/j.intell.2011.11.002

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: An illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and Writing*, *28*(1), 31–56. doi:10.1007/s11145-014-9518-z

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (pp. 1–10). Vienna, Austria.

Plummer, M. (2016). *rjags: Bayesian graphical models using MCMC*. R package version 4-6. Retrieved from https://CRAN.R-project.org/package=rjags

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS technical report for reading: Scaling results of starting cohort 3 in fifth grade (NEPS working paper no. 15)*. Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S. (March 2019). *Using timing information to model missing values in test data*. Paper presented at the fifth conference of the German Consortium in Statistics [Deutsche Arbeitsgemeinschaft Statistik]. Munich, Germany.

Pohl, S. & Becker, B. (2019). *Performance of missing data approaches under nonignorable missing data conditions*. Manuscript submitted for publication.

Pohl, S. & Carstensen, C. H. (2012, October). *NEPS technical report - Scaling the data of the competence tests* (NEPS Working Paper No. 14). Otto-Friedrich-Universität, Nationales Bildungspanel. Bamberg, Germany.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*(3), 423–452. doi:10.1177/0013164413504926

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response time models to account for not-reached items. *Psychometrika*, *84*(3), 892–920. doi:10.1007/s11336-019-09669-2

Pohl, S., Ulitzsch, E., & von Davier, M. (manuscript in preparation). *PISA results revisited: Disentangling different aspects that describe performance on competence tests*.

Pohl, S. & von Davier, M. (2018). Commentary: "On the importance of the speed-ability trade-off when dealing with not reached items" by Jesper Tijmstra and Maria Bolsinova. *Frontiers in Psychology*, *9*(1988). doi:10.3389/fpsyg.2018.01988

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325. doi:10.3102/1076998616636618

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47. doi:10.1111/emip.12102

R Development Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Ranger, J. & Kuhn, J.-T. (2012a). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31–47. doi:10.1007/s11336-011-9231-7

Ranger, J. & Kuhn, J.-T. (2012b). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*, *36*(3), 214–231. doi:10.1177/0146621612439796

Ranger, J. & Ortner, T. (2012). 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.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104. doi:10.1080/15305058.2016.1231193

Robitzsch, A. (2014). Zu nichtignorierbaren Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment. In B. Suchan, C. Wallner-Paschon, & C. Schreiner (Eds.), *PIRLS & TIMSS 2011 - die Kompetenzen in Lesen, Mathematik und Naturwissenschaft am Ende der Volksschule: Österreichischer Expertenbericht* (pp. 55–64). Graz, Austria.

Rohwer, G. (2013). *Making sense of missing answers in competence tests* (NEPS Working Paper No. 30). Otto-Friedrich-Universität, Nationales Bildungspanel. Bamberg, Germany.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Doctoral dissertation, Friedrich-Schiller-Universität Jena). Retrieved from https://d-nb.info/1036873145/34

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*(3), 795–819. doi:10.1007/s11336-016-9544-7

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report No. RR-10-11). Educational Testing Service. Princeton, NJ. doi:10.1002/j.2333-8504.2010.tb02218.x

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–174). New York, NY: Elsevier Science.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi:10.1093/biomet/63.3.581

Sachse, K., Mahler, N., & Pohl, S. (2019). Effects of changing nonresponse mechanisms on trends and group comparisons in international large-scale assessments. *Educational and Psychological Measurement*, *79*(4), 699–726. doi:10.1177/0013164419829196

San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*(3), 183–203. doi:10.1177/01466216052.82773

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi:10.1037/1082-989X.7.2.147

Schnipke, D. L. (April 1996). *How contaminated by guessing are item-parameter estimates and what can be done about it?* Paper presented at the Annual Meeting of the National Council on Measurement in Education. New York, NY.

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. doi:10.1111/j.1745-3984.1997.tb00516.x

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). New York, NY: Psychological Press.

Schuurman, N., Grasman, R., & Hamaker, E. (2016). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51*(2-3), 185–206. doi:10.1080/00273171.2015.1065398

Semmes, R., Davison, M. L., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement*, *35*(6), 433–446. doi:10.1177/0146621611407305

Senkbeil, M., Ihme, J. M., & Adrian, E. (2014). *NEPS technical report for computer literacy: Scaling results of starting cohort 3 in grade 6*. Leibniz Institute for Educational Trajectories.

Singer, J. D. & Braun, H. I. (2018). Testing international education assessments. *Science*, *360*(6384), 38–40. doi:10.1126/science.aar4952

Stan development team. (2017). *Stan modeling language: User's guide and reference manual* (version No. 2.17.0). Retrieved from https://github.com/stan-dev/stan/releases/download/v2.17.1/stan-reference-2.17.1.pdf

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). New York, NY: Academic Press.

Tijmstra, J. & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology*, *9*. doi:10.3389/fpsyg.2018.00964

Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (manuscript in preparation). *A validation of measures for identifying examinee disengagement*.

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*. doi:10.1111/bmsp.12188

Ulitzsch, E., von Davier, M., & Pohl, S. (2019b). A multi-process item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*. doi:10.1177/0013164419878241

Ulitzsch, E., von Davier, M., & Pohl, S. (2019c). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*. doi:10.1080/00273171.2019.1643699

UNESCO Institute for Statistics. (2018). *The impact of large-scale learning assessments*. Montreal, Canada. Retrieved from http://uis.unesco.org/sites/default/files/documents/impact-large-scale-assessments-2018-en.pdf

van den Wollenberg, A. L. (1979). *The Rasch model and time-limit tests: An application and some theoretical contributions* (Doctoral dissertation, Katholieke Universiteit te Nijmegen). Retrieved from https://repository.ubn.ru.nl/bitstream/handle/2066/147865/mmubn000001_026549123.pdf

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. doi: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. doi:10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20. doi:10.3102/1076998607302626

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272. doi:10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J. (2011a). Setting time limits on tests. *Applied Psychological Measurement*, *35*(3), 183–199. doi:10.1177/0146621610391648

van der Linden, W. J. (2011b). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60. doi:10.1111/j.1745-3984.2010.00130.x

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130. doi:10.1111/j.1745-3984.2007.00030.x

van der Linden, W. J. & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139. doi: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. doi: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. doi:10.1177/0146621609349800

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210. doi:10.1177/01466219922031329

van Barneveld, C., Pharand, S.-L., Ruberto, L., & Haggarty, D. (2013). Student motivation in large-scale assessments. In M. Simon & K. E. and Michel Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues and practice* (pp. 43–61). New York, NY: Routledge.

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. doi:10.1037/met0000162

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth ed.). ISBN 0-387-95457-0. New York, NY: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Verbic, S. & Tomic, B. (2009). Test item response time and the response likelihood. Retrieved from https://arxiv.org/ftp/arxiv/papers/0901/0901.4356.pdf

Verhelst, N. D., Verstralen, H. H., & Jansen, M. (1997). A logistic model for time-limit tests. In *Handbook of modern item response theory* (pp. 169–185). New York, NY: Springer.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2), i–35. doi:10.1002/j.2333-8504.2005.tb01993.x

von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (2013). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*. New York, NY: Springer.

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. doi: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. doi:10.3102/1076998618767123

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.

Wild, C. & Durso, R. (1979). *Effect of increased test-taking time on test scores by ethnic groups, age, and sex*. Educational Testing Service. GRE Board Research Report GREB No. 76-6R. Princeton, NJ. Retrieved from https://www.ets.org/Media/Research/pdf/GREB-76-06R.pdf

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252. doi:10.1080/08957347.2015.1042155

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. doi:10.1111/emip.12165

Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice*, *25*(2), 21–30. doi:10.1111/j.1745-3992.2006.00054.x

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wise, S. L. & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. doi:10.1111/j.1745-3984.2006.00002.x

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. doi:10.1080/08957347.2017.1353992

Wise, S. L., Kingsbury, G., Thomason, J., & Kong, X. (April 2004). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Diego, CA.

Wise, S. L. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S. L. & Ma, L. (April 2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Vancouver, Canada.

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. doi:10.1080/08957340902754650

Xu, D., Chatterjee, A., & Daniels, M. (2016). A note on posterior predictive checks to assess model fit for incomplete data. *Statistics in Medicine*, *35*(27), 5029–5039. doi:10.1002/sim.7040

Yamamoto, K. & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost (Ed.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Münster, Germany: Waxmann.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In *Technical report of the survey of adult skills (PIAAC)*. Retrieved from http://www.oecd.org/skills/piaac/Technical_Report_2nd_Edition_Chapters_17-23.pdf

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). Design and analysis in a cognitive assessment. *ETS Research Report Series*, *2003*(2), i–47. doi:10.1002/j.2333-8504.2003.tb01924.x

Zamarro, G., Cheng, A., Shakeel, M. D., & Hitt, C. (2018). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics*, *72*, 51–60. doi:10.1016/j.socec.2017.11.005

Zamarro, G., Hitt, C., & Mendez, I. (2016). *When students don't care: Reexamining international differences in achievement and non-cognitive skills*. The University of Arkansas, Department of Education Reform. EDRE Working Paper 2016-18.

Zhan, P., Jiao, H., Wang, W.-C., & Man, K. (2018). A multidimensional hierarchical framework for modeling speed and ability in computer-based multidimensional tests. Retrieved from https://arxiv.org/pdf/1807.04003.pdf

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement*, *53*(2), 190–211. doi:10.1111/jedm.12107

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. doi:10.1080/10705511.2018.1545232

A
Appendix to Chapter 2

A.1 JAGS Code and Prior Settings

Table A.1. Prior Settings

Speed-accuracy model	$\mathbf{\Sigma}_{\mathcal{P}} \sim \text{IW}_{2+1}(\mathbf{I}_2)$
	$\mathbf{\Sigma}_{\mathcal{J}} \sim \text{IW}_{2+1}(\mathbf{I}_2)$
	$\mu_b \sim \mathcal{N}(1, 1000^2)$
	$\mu_\beta \sim \mathcal{N}(1, 1000^2)$
	$\alpha^2 \sim \Gamma(0.01, 0.001)$
Manifest missing response model	$\sigma_\theta^2 \sim \text{IG}(0.01, 0.001)$
	$\gamma \sim \mathcal{N}(0, 1000^2)$
	$\mu_\beta \sim \mathcal{N}(1, 1000^2)$
	$\sigma_\beta^2 \sim \text{IG}(0.01, 0.001)$

Note: $\text{IW}_{2+1}(.)$: inverse Wishart prior with 2+1 degrees of freedom; $\mathcal{N}(.,\ .)$: normal prior; $\text{IG}(.,.)$: inverse gamma prior; $\Gamma(.,.)$: gamma prior; $\mathbf{I}_2$ represents an identity matrix of size 2.

A.1.1 JAGS Code: Speed-Accuracy Model

```
model {
  for (i in 1:N){
    for (j in 1:K){
      # item responses
      U[i,j] ~ dbern(prob[i, j])
      logit(prob[i,j]) <- PersPar[i,1] - ItemPar[j,1]
      # response times
      RT[i,j] ~ dlnorm(muOfLogX[i,j] , alpha.sqr )
      muOfLogX[i,j] <- ItemPar[j,2]- PersPar[i,2]
    }
  # prior for person parameter
  PersPar[i,1:2] ~ dmnorm(muP, invSigmaP)
  }
  # hyperprior for person parameter
  muP <- c(0,0)
  invSigmaP ~ dwish(M,3)
  SigmaP <- inverse(invSigmaP)
  correlP <- SigmaP[1,2]/(sqrt(SigmaP[1,1])*sqrt(SigmaP[2,2]))
  # prior for item parameter
  # prior for alpha
  alpha.sqr ~ dgamma(0.01, 0.001)
  alpha<-sqrt(alpha.sqr)
  for (j in 1:K){
    ItemPar[j,1:2] ~ dmnorm(muI, omegaI)
  }
  # hyperprior for item parameter
  muI[1] ~ dnorm(0, 0.000001)
  muI[2] ~ dnorm(1, 0.000001)
  invSigmaI ~ dwish(M,3)
  SigmaI <- inverse(invSigmaI)
  correlI <- SigmaI[1,2]/(sqrt(SigmaI[1,1])*sqrt(SigmaI[2,2]))
}
```

Figure A.1. N: number of persons, K: number of items. **M** represents an identity matrix of size 2. **U** is an N by K matrix containing the item responses and **RT** is an N by K matrix containing the associated response times.

A.1.2 JAGS Code: Manifest Missing Data Model

```
model {
  for (j in 1:N){
    for (i in 1:K){
      U[i, j] ~ dbern(prob[i, j])
      logit(prob[i, j]) <- theta[i] - b[j]
    }
    # prior for person parameter
    theta[i] ~ dnorm(muP[i], invSigmaP)
    muP[i] <- gamma[1] + gamma[2]*Z[i]
  }
  # prior for item difficulties
  for (j in 1:K) {
    b[j] ~ dnorm(muI, invSigmaI)
  }
  # identification and prior for beta
  gamma[1] <- 0
  gamma[2] ~ dnorm(0,0.000001)
  # hyperprior for person parameter
  invSigmaP   ~ dgamma(0.01, 0.001)
  SigmaP <- 1/invSigmaP
  # hyperprior for item parameter
  muI ~ dnorm(0, 0.000001)
  invSigmaI   ~ dgamma(0.01, 0.001)
  SigmaI <- 1/invSigmaI
}
```

Figure A.2. N: number of persons, K: number of items. **U** is an N by K matrix containing the item responses. **Z** is a vector of length n representing the number of not-reached items.

A.2 Differences in Speed Estimates

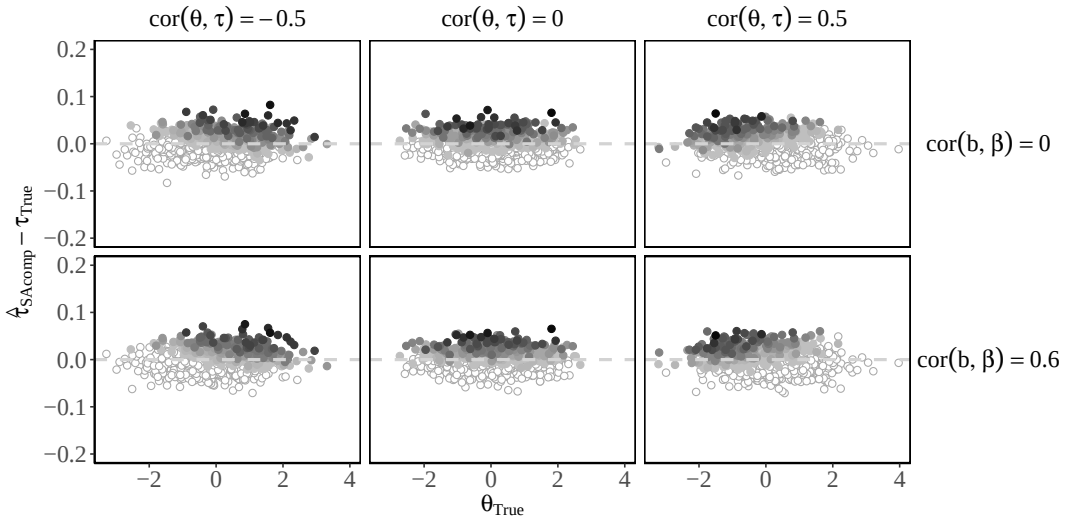

Figure A.3. Difference in speed estimates using the SA model for complete data compared to the true speed values as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

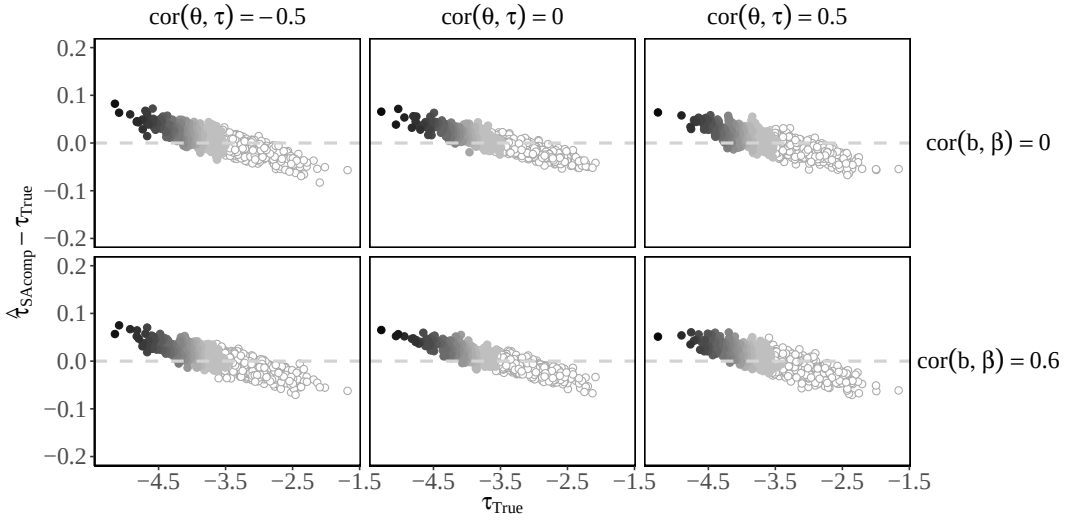

Figure A.4. Difference in speed estimates using the SA model for complete data compared to the true speed values as a function of true speed. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

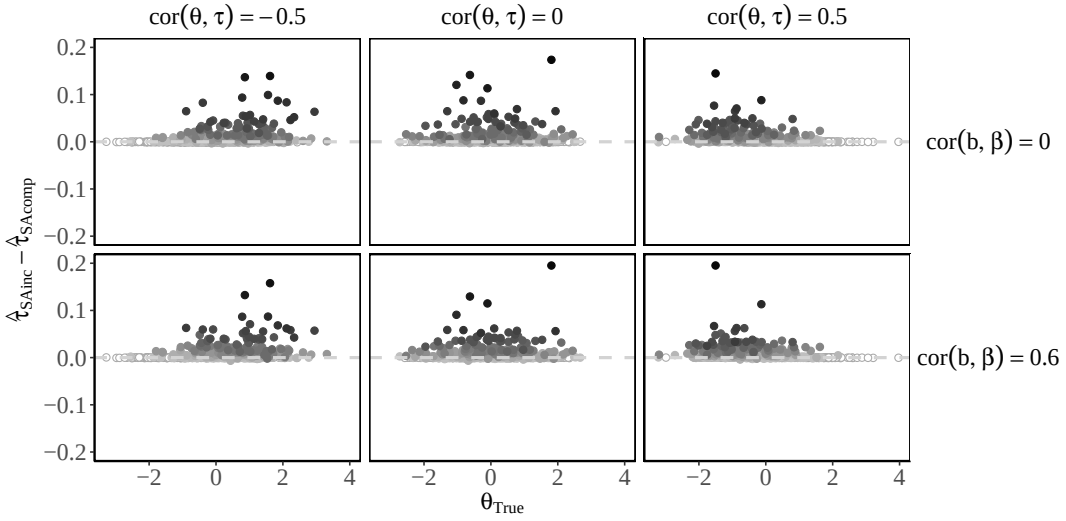

Figure A.5. Difference in speed estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

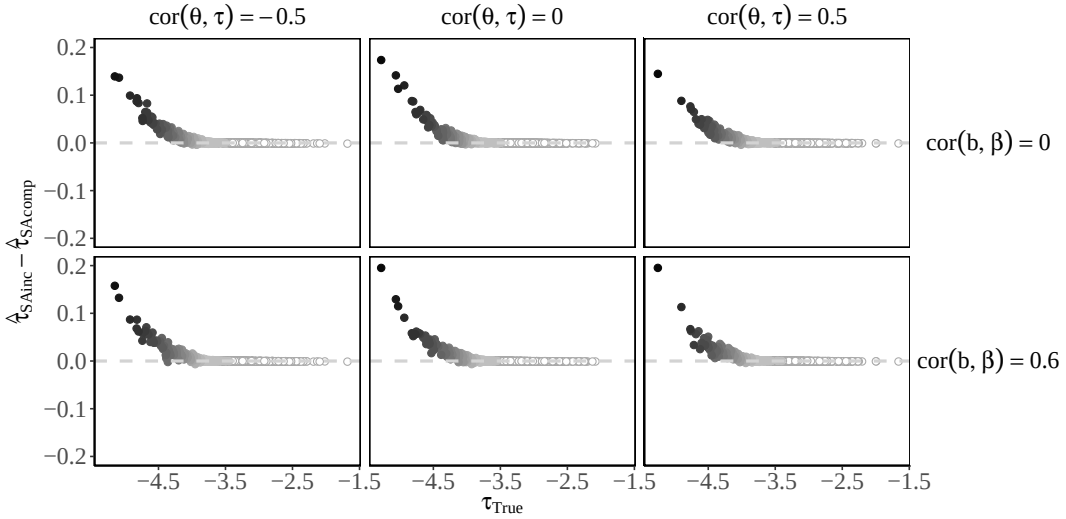

Figure A.6. Difference in speed estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true speed. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

A.3 Subsequent Analyses

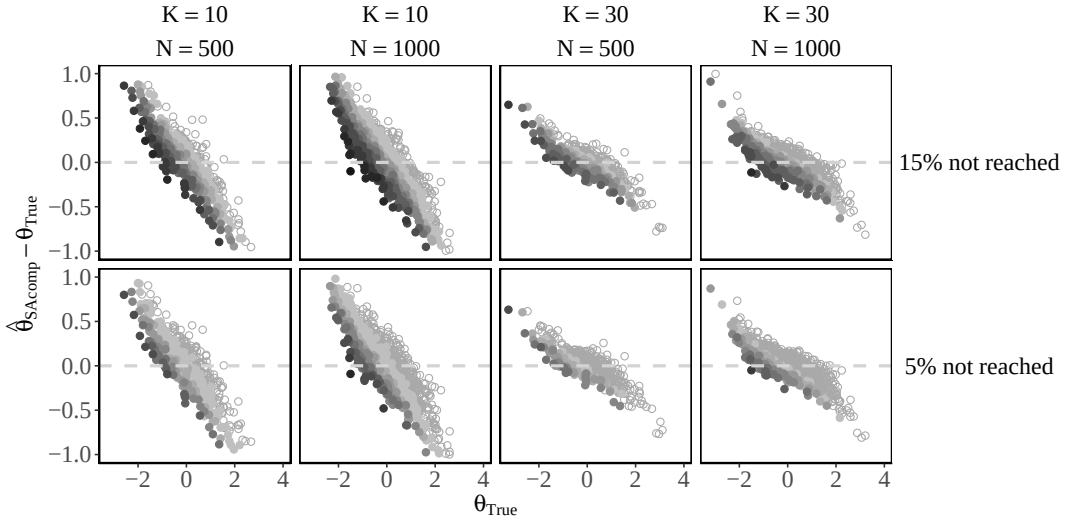

Figure A.7. Difference in ability estimates using the SA model for complete data (SAcomp) compared to the true ability values as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

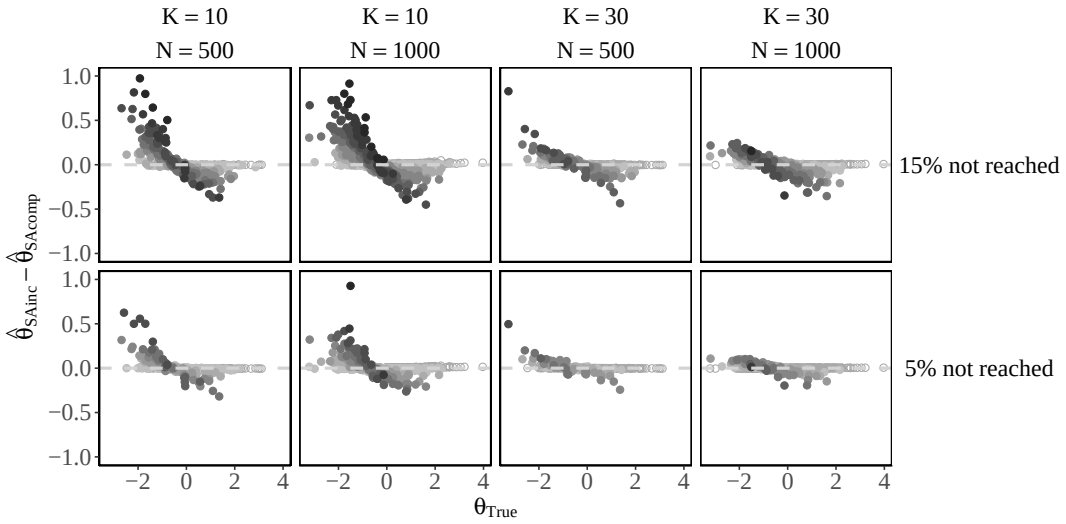

Figure A.8. Difference in ability estimates between using the SA model for incomplete data (SAinc) and using the SA model for complete data (SAcomp) as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color with darker colors denoting a higher number of not-reached items.

B
Appendix to Chapter 3

B.1 JAGS Code

```
model {
  for (i in 1:N){
    for (j in 1:K){
      # item responses
      U[i, j] ~ dbern(prob[i, j])
      logit(prob[i, j]) <- PersPar[i,1] - ItemPar[j,1]
      # response times
      RT[i,j] ~ dlnorm(muOfLogRT[i,j] , alpha.sqr )
      muOfLogRT[i,j] <-  ItemPar[j,2]-PersPar[i,2]
    }
    # number of reached items before quitting
    isCensored[i] ~ dinterval(kQ[i], c(0,kC[i]))
    kQ[i]~ dpois(exp(PersPar[i,3]))
    # prior for person parameter
    PersPar[i,1:3] ~ dmnorm(muP, invSigmaP)
  }
  # hyperprior for person parameter
  muP[1:2]<-c(0,0)
  muP[3]~ dnorm(0, 0.00001)
  invSigmaP ~ dwish(MP,4)
  SigmaP <- inverse(invSigmaP)
  for(i in 1:3){
    for(j in 1:3){
      correlP[i,j]<-SigmaP[i,j]/(sqrt(SigmaP[i,i])*sqrt(SigmaP[j,j]))
    }
  }
  # prior for item parameter
  for (j in 1:K){
    ItemPar[j,1:2] ~ dmnorm(muI, omegaI)
  }
  # hyperprior for itemparameter
  muI[1] ~ dnorm(0, 0.000001)
  muI[2] ~ dnorm(0, 0.000001)
  omegaI ~ dwish(MI,3)
  SigmaI <- inverse(omegaI)
  for(i in 1:2){
    for(j in 1:2){
      correlI[i,j]<-SigmaI[i,j]/(sqrt(SigmaI[i,i])*sqrt(SigmaI[j,j]))
    }
  }
  alpha.sqr~dgamma(0.5, 0.001)
  alpha<-sqrt(alpha.sqr)
}
```

Figure B.1. JAGS code for the speed-accuracy+quitting model. N: number of persons, K: number of items. **U** and **RT** are N by K matrices containing the item responses and associated response times, **kQ** and **kC** are vectors of length N containing the number of reached items before quitting and the censoring item position. **isCensored** contains information on observed quitting behavior for each examinee i and takes the values 1 and 2 for $c_i = 1$ and $c_i = 0$. **MP** and **MI** represent identity matrices of size 3 and 2.

B.2 Coverage

Table B.1. Coverage of item parameter means, variances, and covariances

K	N	% NR	Mechanisms	$var(b)$	$var(\beta)$	$cov(b,\beta)$	μ_b	μ_β
20	350	2.5%	quitting	.89	1.00	.99	.96	.98
			speed & quitting	.89	.98	.94	.89	.96
		5%	quitting	.95	.97	.92	.94	.98
			speed & quitting	.95	.94	.96	.93	.96
		10%	quitting	.96	.99	.98	1.00	.97
			speed & quitting	.95	1.00	.98	.93	.97
	700	2.5%	quitting	.96	1.00	.99	.90	.97
			speed & quitting	.95	.98	.95	.96	.95
		5%	quitting	.93	.98	.99	.90	.94
			speed & quitting	.90	.98	.96	.97	.97
		10%	quitting	.86	.97	.92	.98	.96
			speed & quitting	.96	.98	.95	.96	.97
40	350	2.5%	quitting	.94	.96	.93	.99	.97
			speed & quitting	.92	.97	.95	.96	.92
		5%	quitting	.95	.98	.95	.96	.97
			speed & quitting	.91	.97	1.00	.94	.95
		10%	quitting	.90	.96	1.00	.82	.97
			speed & quitting	.96	.95	.90	.95	.92
	700	2.5%	quitting	.93	.94	.93	.93	.95
			speed & quitting	.95	.97	.95	.94	.98
		5%	quitting	.94	.95	.97	.99	.98
			speed & quitting	.94	.96	.96	.97	.97
		10%	quitting	.96	.99	.98	.97	.99
			speed & quitting	.98	.94	.94	.99	.98

Note: % NR: overall missingness rate; quitting and speed & quitting denote conditions under which all not-reached items go back to quitting and not-reached items occurred to both lack of speed and quitting, respectively; K: number of items; N: number of examinees; b: item difficulty; β: time intensity; μ_b: mean item difficulty; μ_β: mean time intensity.

Table B.2. Coverage of person parameter means, variances, and covariances

K	N	% NR	Mechanisms	$var(\theta)$	$var(\tau)$	$var(\zeta)$	$cov(\theta,\tau)$	$cov(\theta,\zeta)$	$cov(\theta,\zeta)$	μ_ζ
20	350	2.5%	quitting	1.00	1.00	.98	1.00	.97	.96	.98
			speed & quitting	.99	1.00	.87	1.00	.97	.98	.93
		5%	quitting	.96	1.00	1.00	1.00	.99	.97	.95
			speed & quitting	.99	1.00	.97	1.00	.98	1.00	.98
		10%	quitting	1.00	1.00	.84	1.00	.97	1.00	.79
			speed & quitting	1.00	1.00	.96	1.00	1.00	.97	.98
	700	2.5%	quitting	.99	1.00	.97	1.00	.95	.94	.96
			speed & quitting	1.00	1.00	.88	1.00	.95	.97	.96
		5%	quitting	.98	1.00	.96	1.00	.93	.97	.92
			speed & quitting	.99	1.00	1.00	1.00	.97	.94	1.00
		10%	quitting	1.00	1.00	.92	1.00	1.00	.95	.76
			speed & quitting	.98	.96	.97	.99	1.00	.98	.91
40	350	2.5%	quitting	1.00	1.00	.91	1.00	.97	.96	.98
			speed & quitting	1.00	1.00	.80	1.00	.95	.91	.83
		5%	quitting	1.00	1.00	.97	1.00	1.00	.99	.98
			speed & quitting	1.00	1.00	.87	1.00	.93	.92	.91
		10%	quitting	1.00	1.00	.99	1.00	.96	1.00	.99
			speed & quitting	.99	1.00	.96	1.00	.99	1.00	.97
	700	2.5%	quitting	1.00	1.00	.95	1.00	.98	.99	.97
			speed & quitting	1.00	1.00	.67	1.00	.91	.90	.75
		5%	quitting	1.00	1.00	.97	1.00	.99	.99	.96
			speed & quitting	1.00	1.00	.86	1.00	.99	.98	.85
		10%	quitting	1.00	1.00	.98	1.00	.99	1.00	.93
			speed & quitting	1.00	1.00	.99	1.00	.96	.97	.99

Note: % NR: overall missingness rate; quitting and speed & quitting denote conditions under which all not-reached items go back to quitting and not-reached items occurred due to both lack of speed and quitting, respectively; K: number of items; N: number of examinees; θ: ability; τ: speed; ζ: test endurance; μ_ζ: mean test endurance.

B.3 Parameter Recovery

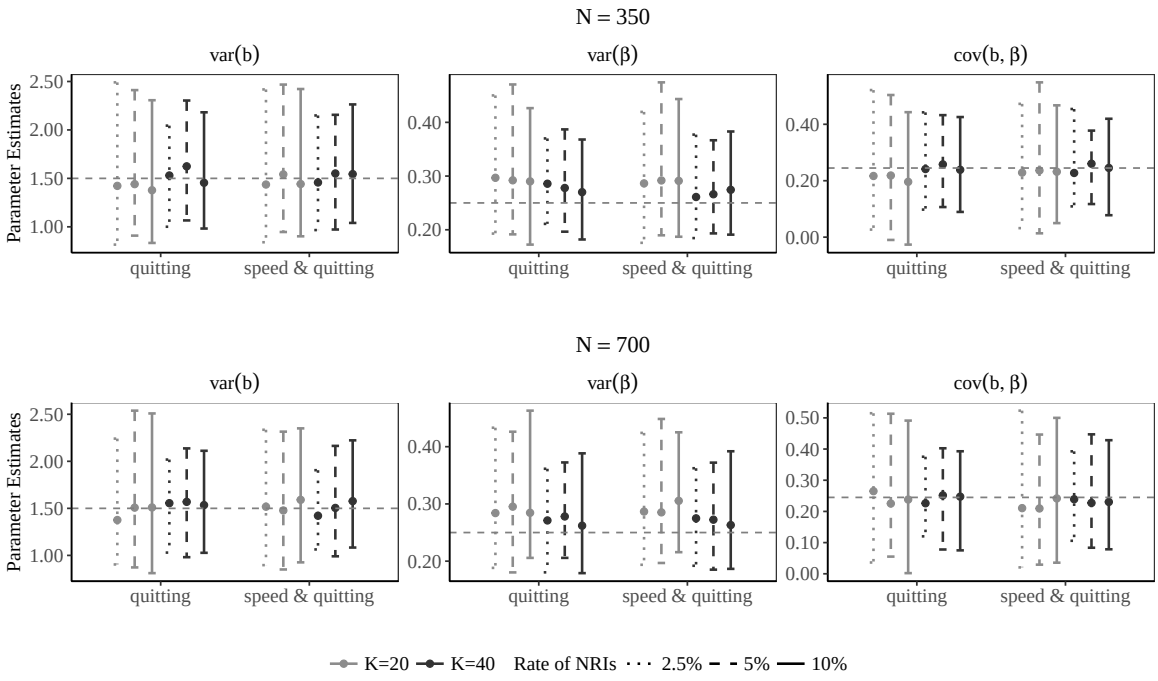

Figure B.2. Medians and 90% ranges of item parameter variance and covariance estimates over all 100 replications per condition. The dashed horizontal line indicates the respective true parameter. Note that y-axes differ in scale. b: item difficulty; β: time intensity; N: number of examinees; K: number of items; NRIs: not-reached items; quitting and speed & quitting denote conditions under which all not-reached items go back to quitting and not-reached items occurred due to both lack of speed and quitting, respectively.

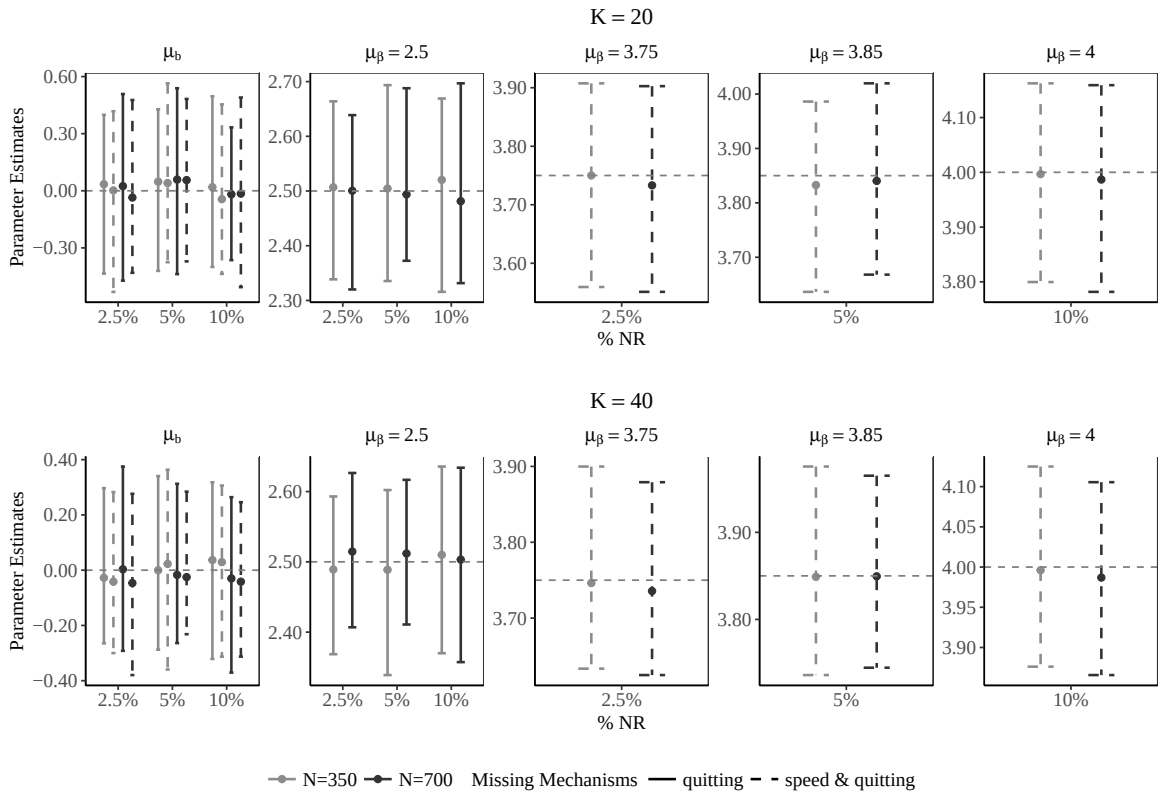

Figure B.3. Medians and 90% ranges of item parameter means over all 100 replications per condition. The dashed horizontal line indicates the respective true parameter. Plots are organized according to the data-generating values employed to achieve missingness rates due to lack of speed ranging from 0% (missingness due to quitting) to 5% (speed & quitting, 10%). Note that y-axes differ in scale. μ_b: mean item difficulty; μ_β: mean time intensity; N: number of examinees; K: number of items; % NR: overall missingness rate due to not-reached items; quitting and speed & quitting denote conditions under which all not-reached items go back to quitting and not-reached items occurred due to both lack of speed and quitting, respectively.

Contributions

This work was supported by the German Research Foundation (DFG) under Grant PO 1655/2-2.

The manuscripts (Chapters 2 to Chapter 5) included in this work are part of a project that was planned and fundraised by Steffi Pohl (SP), in which Matthias von Davier (MD) takes on the role of a cooperation partner, and for which SP takes responsibility.

Chapter 2

The core ideas of the manuscript were developed by SP and MD. The design of the simulation study and the analysis strategy for the empirical example were developed by SP, MD, and Esther Ulitzsch (EU). SP programmed the analysis model and obtained first results. EU programmed the simulation study and analyzed the empirical data. SP, MD, and EU discussed the results. SP wrote the first draft of the paper, MD and EU contributed to the final manuscript.

Chapters 3 and 4

The core ideas of the manuscript, the model, the design of the simulation study, and the analysis strategy for the empirical example were developed by EU, SP, and MD. EU programmed the simulation study and analyzed the data. SP supervised the work and gave feedback. EU, SP and MD discussed the findings of this work. EU wrote the first draft of the paper, SP and MD gave feedback on the results and the writing, and contributed to the final manuscript.

Chapter 5

The core ideas of the manuscript, the model, the design of the simulation study, and the analysis strategy for the empirical example were developed by EU, MD, and SP. EU programmed the simulation study and analyzed the data. MD and SP supervised the work and gave feedback. EU, SP and MD discussed the findings of this work. EU wrote the first draft of the paper. SP and MD gave feedback on the results and 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, den 26. September 2019 _____

Esther Ulitzsch