Disengaged Responding in Educational Large-Scale Assessments

New Insights into its Identification, Modeling, and Consequences

Inaugural-Dissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität

Bamberg



vorgelegt von

Jana Welling

aus

München

Bamberg 2026

Tag der mündlichen Prüfung: 15.07.2026

Dekan: Prof. Dr. Claus H. Carstensen

Erstgutachter: Prof. Dr. Claus H. Carstensen

Zweitgutachter: Dr. Timo Gnambs

Dieses Werk ist als freie Onlineversion über das Forschungsinformationssystem (FIS; https://fis.uni-bamberg.de) der Universität Bamberg erreichbar.
Das Werk steht unter der CC-Lizenz CC BY.

Lizenzvertrag: Creative Commons Namensnennung 4.0
https://creativecommons.org/licenses/by/4.0/



URN: urn:nbn:de:bvb:473-irb-116761x

DOI: https://doi.org/10.20378/irb-116761

Acknowledgements

I would like to take the opportunity to thank all the people who have supported me throughout my time as PhD student, both professionally and personally. First and foremost, I would like to express my sincere gratitude to my three supervisors, Claus Carstensen, Timo Gnambs, and Esther Ulitzsch, for their continuous guidance and support. I feel very fortunate to have benefited from such a diverse supervisory team. I am especially grateful to Timo Gnambs for always taking the time to answer my countless questions, for engaging in stimulating discussions about my research, and for shielding me from work during particularly stressful periods. Your detailed and constructive feedback as well as your readiness to help were invaluable to the completion of this dissertation. A special thank you also goes to Esther Ulitzsch, who joined my supervisory team later in my PhD, but provided valuable and thoughtful feedback on parts of this dissertation. Your insights significantly improved this work and were greatly appreciated.

I further would like to thank all other co-authors, Eva Zink, Luis Fischer, and Kathrin Thums, for their constructive feedback and enriching contributions, which have substantially strengthened the articles included in this dissertation. I am grateful to my colleagues, both at LIfBi and at conferences, for the inspiring discussions and helpful suggestions that have further refined the work and shaped my thinking. Special thanks go to my former colleague Marie-Ann Sengewald, who not only provided valuable professional feedback but also fostered a warm and welcoming team environment. Your readiness to help and your sympathetic nature made a meaningful difference in helping me feel comfortable and supported. A huge thank you goes to my dear friend and co-PhD student Annette Malapally, who accompanied me through the ups and downs of the PhD journey both personally and academically. Our countless discussions and numerous phone calls as well as your thoughtful feedback and constant encouragement helped me not only with strengthening this thesis but also with cheering up even during difficult moments.

Moreover, this dissertation would not have been possible without numerous people at LIfBi, BAGSS, and the University of Bamberg, who supported me with their counseling, training, organization, IT services, and so much more. Thank you for always taking the time to quickly respond to any inquiries,

even on Friday afternoons (I am still sorry for the short-term travel request!). Thank you to my mentor Regina Becker, who made the beginning of my PhD journey so much easier. Special thanks go also to my fellow PhD students at LIfBi for creating such a welcoming community. Being part of the PhD representatives as well as our regular social events have made this journey both enjoyable and rewarding.

Finally, I would like to express my gratitude to those who have accompanied me throughout my PhD from outside academia. Especially during the past year, which has been challenging at times, they have supported me with their unwavering love and encouragement. My deepest thanks go to my partner, Daniel, for always being there for me, for cooking delicious meals while I was working on this dissertation, and for grounding me when I felt overwhelmed. A very special thank you also goes to my family and friends for their companionship, constant support, and for making this and all other journeys so much more enjoyable. Your belief in me has been a great source of strength.

Abstract

Educational large-scale assessments (LSAs) provide information on competencies across cohorts, populations, and test domains. However, when examinees lack test-taking motivation, they can respond to test items without investing their best effort, knowledge, and abilities, a behavior called disengaged responding. Because disengaged responses do not reflect the proficiency of the examinees, these responses can pose a threat to the validity of assessment outcomes. To ensure the interpretability of findings from LSAs, disengaged responding needs to be accurately identified and accounted for in the measurement of competencies. Moreover, it is important to understand in which contexts disengaged responding arises and which consequences it entails. Therefore, this dissertation includes three articles that addressed major, yet unresolved issues in the context of disengaged responding concerning its identification (Article 1), accounting for it in the measurement of competencies (Article 2), and the generalizability of established findings to child samples (Article 3).

Most approaches to identifying disengaged responding rely solely on item responses and response times, thereby risking the misclassification of fast engaged or slow disengaged responses. Article 1 aimed to improve the classification accuracy of response-time based identification methods by introducing and evaluating three novel process data indicators of disengaged responding in reading comprehension tasks: text reread, item revisit, and answer change. Findings from an empirical study suggest that the improvement in classification accuracy was marginal to non-existent, thereby confirming response times as an important indicator of disengagement.

Although many LSAs provide proficiency estimates as plausible values, no study has yet investigated how disengagement should be accounted for in their estimation. Article 2 addressed this gap by proposing and evaluating different approaches to (not) accounting for disengaged responding in the estimation of plausible values. A simulation study demonstrated that achievement gaps estimated with plausible values can be biased by disengaged responding. While models that accounted for disengagement in the measurement model effectively reduced these biases, models accounting for it solely in the population model did not. An additional empirical study found no difference between

the models, supporting recent studies which claim that the impact of disengaged responding on achievement gaps may be negligible in practice.

Most empirical findings on disengaged responding are based on adolescent or adult samples, ignoring the fact that many LSAs also include child samples. Article 3 challenged the generalizability of established findings on disengaged responding to child samples by providing theoretical considerations on why the test-taking behavior of children may differ from that of older examinees. Additionally, the findings from an empirical study supported recent research indicating that disengaged responding may be less common among children. Only proficiency could be replicated as predictor of disengaged responding, while gender and socio-economic status could not. Further results suggested that the impact of disengaged responding on achievement gaps may be limited in this age group.

Taken together, this dissertation combines theoretical, methodological, and empirical approaches to improve our understanding of disengaged responding and the validity of LSAs. The central findings bear important implications for the identification, modeling, and consequences of disengaged responding. First, they support response-time based measures as strong indicators of disengagement while suggesting that process data may not always provide sufficient information on test-taking engagement to inform new identification methods. Second, they highlight the need to accurately account for disengaged responding also in plausible value estimation. Third, this dissertation suggests that and discusses why the threat (rapid) disengagement poses to the validity of group-based assessment outcomes may be less pronounced than often assumed. Fourth, it challenges the generalizability of established findings to child samples. By laying conceptual and methodological groundwork, this dissertation provides a foundation for future research aimed at developing more robust, valid, and context-sensitive approaches to measuring competencies in educational LSAs.

Table of Contents

ACKNOWLEDGEMENTS ..... **I**

ABSTRACT ..... **III**

1 INTRODUCTION **1**

2 LITERATURE REVIEW **4**

2.1 Theoretical Framework 4

2.2 Identification of Disengaged Responding 6

2.3 Accounting for Disengaged Responding in Proficiency Estimation 10

2.4 Prevalence and Antecedents of Disengaged Responding 12

2.5 Consequences for the Validity of Proficiency Estimates 14

3 RATIONALE FOR DISSERTATION **17**

3.1 Misclassifications in Response-Time Based Identification of Disengaged Responding 17

3.2 Missing Approaches to Accounting for Disengaged Responding in Plausible Value Estimation 18

3.3 Generalizability of Findings on Disengaged Responding to Child Samples 19

4 SUMMARY OF ARTICLES **21**

4.1 Article 1: Identifying Disengaged Responding in Multiple-Choice Items: Extending a Latent Class Item Response Model with Novel Process Data Indicators 21

4.2 Article 2: Comparing Different Approaches of (Not) Accounting for Rapid Guessing in Plausible Values Estimation 23

4.3 Article 3: Prevalence, Predictors, and Consequences of Rapid Guessing among Children in Elementary School 25

5 DISCUSSION **28**

5.1 Summary of Key Findings 28

5.2 Integrating Findings Across Articles 29

5.3 Implications for Large-Scale Assessments 39

5.4 Strengths, Limitations, and Outlook 41

5.5 Conclusion 44

REFERENCES **47**

AUTHOR CONTRIBUTIONS **58**

APPENDIX **61**

1 Introduction

Educational large-scale assessments (LSAs) such as the *Programme for International Student Assessment* (PISA) or the *Programme for the International Assessment of Adult Competencies* (PIAAC) provide important information on competencies across different populations and cohorts, thereby enabling fundamental research on competence development, educational effectiveness, and social inequality in education. However, valid inferences require an accurate measurement of the competencies of interest. Because LSAs usually do not entail any performance-based consequences for the examinees, this prerequisite may be violated. The lack of consequences can impair participants' test-taking motivation, which in turn can lead to a behavior called *disengaged responding* – that is, the test takers respond to items without investing their best effort, knowledge, and abilities (Wise, 2017). When test takers are disengaged, their responses may not accurately reflect their true proficiency, thereby undermining a central assumption of competence testing and posing a risk to the validity of proficiency scores (Wise, 2017; Wise & Demars, 2005).

Because of the potentially severe implications for LSAs, there has been a growing body of research on disengaged responding and its consequences over the past three decades (Alahmadi & DeMars, 2025; Goldhammer et al., 2017; Rios et al., 2017, 2022; Schnipke & Scrams, 1997; Ulitzsch et al., 2020; Wise, 2017, 2020; Wise & Demars, 2005, 2010). Although previous studies have gained numerous insights into the causes, measurement, antecedents, prevalence, and impact of disengaged responding, important questions remain unanswered. Therefore, this cumulative dissertation addresses three of the most consequential research gaps concerning the identification of disengaged responses (Article 1), accounting for disengagement in the measurement of competencies (Article 2), and the generalizability of established findings to child samples (Article 3). Combining methodological, empirical, and theoretical approaches, it contributes to an improved understanding of disengaged responding and how to handle it in the context of educational LSAs. The following paragraphs present each of the three research gaps and shortly describe the unique contribution of each article included in this dissertation.

The prerequisite for any empirical research on disengaged responding and for accounting for it in proficiency estimation is to accurately assess it. Because disengaged responding is an unobservable construct, this poses a great challenge to the researcher. Most proposed identification methods focus on disengaged responses characterized by unusually short response times, since they are easily detectable (Nagy & Ulitzsch, 2022; Wang & Xu, 2015; Wise, 2019; Wise & Ma, 2012). However, response-time based approaches are prone to misclassifications when engaged examinees respond very fast or when disengaged examinees respond slowly (Wise, 2020). Therefore, there is a need for new methods that classify test-taking engagement more accurately. One possible approach is to incorporate additional information on the test-taking process into existing identification methods. Article 1 addressed this gap by proposing and evaluating novel potential indicators of test-taking engagement derived from process data, thereby contributing to an enhanced identification and classification of disengaged responding.

Once disengaged responses have been identified, they need to be accounted for in the estimation of competencies to ensure unbiased proficiency estimates and valid inferences. Although many approaches to accounting for disengaged responding have been proposed and evaluated (e.g., Ulitzsch et al., 2020; Wise & Demars, 2006), none exist for plausible value estimation. However, in many educational LSAs, proficiency estimates are provided as plausible values due to their considerable methodological and practical strengths (Wu, 2005). To close this gap, Article 2 proposed and evaluated different approaches to accounting for disengaged responding in the estimation of plausible values. In doing so, it advances our understanding of how disengaged responding can be appropriately handled in LSAs and supports researchers and practitioners in a more accurate measurement of competencies.

In addition to the methodological challenges posed by disengaged responding, the generalizability of established findings to child samples remains uncertain, since most studies on disengaged responding rely on adolescent and adult samples (e.g., Goldhammer et al., 2017; Kroehne et al., 2020; Michaelides & Ivanova, 2022). Because children are still undergoing cognitive development (Piaget, 1970/2022), their test-taking behavior may differ from that of older examinees. However, no study has yet examined whether findings on disengagement replicate in younger samples, although

initial evidence suggests that children may be more engaged than adolescents (Jensen et al., 2018; Kuhfeld & Soland, 2020; Wise & Kingsbury, 2022). The third study addresses this gap both by providing theoretical considerations on children's test-taking engagement and by investigating the antecedents, prevalence, and impact of disengaged responding in a sample of German elementary school students. It contributes to a more comprehensive picture of test-taking engagement by increasing our understanding of disengaged responding among children.

The remainder of the dissertation is structured as follows. First, a literature review provides an overview of the construct of disengaged responding, its identification and incorporation in proficiency estimation, as well as important empirical findings. The review begins with a section that defines key terms used in the disengagement literature and summarizes existing theories regarding the occurrence and potential consequences of disengaged responding. The subsequent two sections present and discuss different approaches to identifying disengagement and to accounting for it in the measurement of proficiency. The final two sections review important empirical findings on disengaged responding, beginning with its prevalence and antecedents and concluding with its impact on the validity of assessment outcomes. Following the literature overview, the rationale for the three articles included in this work is presented. Each article is then summarized, highlighting its unique contribution to the disengagement literature. Finally, an overarching discussion integrates the results of the articles by identifying shared findings and conclusions, discussing limitations and implications for LSAs, and outlining directions for future research.

2 Literature Review

2.1 Theoretical Framework

Educational LSAs measure competencies across different cohorts, populations, and domains to inform policy and practice. For example, evidence on achievement gaps between subpopulations can prompt decision makers to provide targeted support to disadvantaged groups. A fundamental assumption underlying competence assessment, however, is that test takers apply their best "knowledge, skills, and abilities when responding to each item" (Wise, 2017, p. 52) – that is, they are *engaged*. If, instead, examinees respond without investing their best effort (i.e., they are *disengaged*), the assessment may – contrary to the test developer's intention – measure not (only) proficiency but (also) motivation. As Cronbach put it already in 1970, "unless [the examinee] cares about the result he cannot be measured." (p. 60).

A growing body of research shows that the assumption of full engagement is often violated in LSAs (Avvisati et al., 2024; Goldhammer et al., 2016; Knekta & Eklöf, 2015; LaFave et al., 2022; Michaelides & Ivanova, 2022; Rios et al., 2022; Rios & Soland, 2022; Wise, 2017; Wise & Demars, 2005). While LSAs are high stakes for administrators, they are typically low stakes for test takers: although results may have important policy implications, they usually carry no direct consequences for participants. This lack of performance-based consequences can impair the examinee's *test-taking motivation* (Wise, 2017), defined as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001). Unmotivated test takers may not invest sufficient effort when responding to test items – a behavior referred to as *disengaged responding.* When the disengagement is accompanied by unusually short response times, it is termed *rapid guessing* (Wise, 2017). On the other hand, when the test takers are motivated and invest effort into their responses, they engage in *solution behavior* (Schnipke & Scrams, 1997).

The literature mainly draws on two theoretical frameworks for explaining the emergence of disengagement in competence assessments. Expectancy-value theory (e.g., Eccles & Wigfield, 2002; Wigfield & Eccles, 2000) focuses on *achievement motivation* and how it arises. According to this theory, achievement motivation is determined by two factors: (a) a person's *expectancy of success* and (b) their

perceived *value of the task*. While the expectancy component can be further divided into *ability beliefs* (i.e., beliefs about general competence) and *expectancies* (i.e., expectancy regarding the specific task), the task value component comprises four parts: *intrinsic value* (e.g., interest, enjoyment), *attainment value* (e.g., importance of doing well on the task), *utility value* (e.g., usefulness), and *cost* (e.g., test anxiety, worry). A person will be motivated in a task if they both expect to succeed and perceive that the task is valuable for them. Due to the lack of consequences in low-stakes assessments, test takers may not ascribe much value to the test and their performance in it (Eklöf, 2010; Wise & Demars, 2005). Empirical results support this assumption, with test-takers in high-stakes assessments reporting higher levels of expectancy, importance, interest, and test anxiety than those in low-stakes tests, while the largest differences were detected for importance and test anxiety (Knekta & Eklöf, 2015). Moreover, LaFave et al. (2022) found that less than two thirds of the participants in the *National Assessment of Educational Progress* reported that the test was important or very important to them.

In contrast to the expectancy-value theory, the demands-capacity model by Wise and Smith (2011) acknowledges that test-taking effort may not only be determined by a test-takers' motivation, but also by test and item characteristics. The authors assume that test takers decide separately for each item whether to respond with or without effort. This decision depends both on the *resource demands* of the item (e.g., difficulty, mental taxation) and on the *effort capacity* of the test taker (i.e., willingness or ability to invest effort). The test taker will invest effort into their response if the effort capacity exceeds the resource demands. Empirical findings support the assumptions of the demands-capacity model. Numerous test- and item-level predictors of test-taking engagement have already been identified (e.g., item type, difficulty and complexity, Avvisati et al., 2024; item position, Guo et al., 2022; test domain, Michaelides & Ivanova, 2022; item length, Setzer et al., 2013), indicating that test-taking engagement indeed does not only depend on person characteristics, but also on factors such as test design and item type. Other studies suggest that test-taking engagement manifests at the response (i.e., item-by-person) level, that is, the test takers decide for each item anew if they want to invest effort or not (Wise & Kingsbury, 2016).

While understanding why disengagement occurs is crucial, it is equally necessary to consider the challenges it poses for LSAs. Since disengaged responses do not reflect the true ability of the test takers, they introduce construct-irrelevant variance to proficiency estimates (see American Educational Research Association et al., 2018, or Messick, 1995, for an introduction to validity and potential threats in testing contexts). As a consequence, the validity of the test scores and interpretations based on them may be impaired (Wise, 2017; Wise & Demars, 2005). To understand the scope of the problem and to avoid the potentially negative consequences for the validity of assessment outcomes, it is essential to accurately identify disengaged responses and to correctly account for them in the measurement of competencies.

2.2 Identification of Disengaged Responding

Because disengaged responding is a motivational construct, it is not directly observable. This poses challenges for both its measurement and its interpretation: since we can never know if respondents *really* were disengaged on a given item or entire test, different approaches for its identification represent only approximations of the true construct. Due to these challenges, scientists have proposed numerous approaches to identifying disengaged responding (Guo et al., 2016; Meyer, 2010; Nagy & Ulitzsch, 2022; Pokropek, 2016; Schnipke & Scrams, 1997; Ulitzsch et al., 2020; Wise, 2019; Wise & Kong, 2005; Wise & Ma, 2012; Yamamoto, 1995).

While the most obvious choice for detecting disengagement is to simply ask test takers after an assessment if they were engaged during the test, this approach comes with the typical pitfalls associated with self-reports: social desirability, response style, memory bias, subjectivity, and potential disengagement also in post-assessment questions, to name a few (Silm et al., 2019; Wise, 2020; Wise & Demars, 2005; Wise & Kong, 2005). Therefore, most approaches rely on observable behavior for identifying disengaged responses.

For example, person-fit statistics from *item response theory* (IRT) models (see Chapter 2.3) can be used to flag test takers with aberrant response patterns (e.g., Bolt et al., 2002; Coniam et al., 2021; Yamamoto, 1995). However, these approaches assess disengaged responding only on the person level, assuming that test-taking engagement is constant across the entire test (Coniam et al., 2021) or

changes only once (Bolt et al., 2002; Yamamoto, 1995). Since both theory and empirical results suggest that disengaged responding manifests at the response level (Goldhammer et al., 2017; Michaelides & Ivanova, 2022; Rios & Soland, 2022; Ulitzsch et al., 2020; Wise, 2017; Wise & Kingsbury, 2016; Wise & Smith, 2011), a more fine-grained measurement of disengagement is required.

Numerous identification approaches leverage response time data to identify disengaged responding for each response separately. They can be broadly divided into two categories: (a) threshold-based and (b) model-based methods. Both approaches assume that disengaged responding is mainly characterized by unusually short response times, therefore focusing on rapid guessing as a common form of disengagement. While threshold-based methods identify disengagement deterministically by comparing the actual item response time to a pre-determined item-specific threshold (Guo et al., 2016; Lee & Jia, 2014; Wise, 2019; Wise & Ma, 2012), model-based methods estimate disengagement probabilistically by jointly modeling test-taking engagement and proficiency based on a test takers' item responses and response times (Nagy & Ulitzsch, 2022; Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015).

Due to their simplicity and low computational demands, many researchers and practitioners rely on threshold-based methods for the measurement of disengaged responding (e.g., Goldhammer et al., 2017; Michaelides & Ivanova, 2022; Rios et al., 2022). These approaches set a response-time threshold for each item that distinguishes between engaged and disengaged responses. If the item response time t_{ij} of person *i* on item *j* is shorter than the item-specific threshold ρ_j, the response engagement ϕ_{ij} is classified as disengaged (ϕ_{ij} = 0), otherwise as engaged (ϕ_{ij} = 1):

$$\phi_{ij} \in \begin{cases} 0, & if\ t_{ij} < \rho_j \\ 1, & if\ t_{ij} \geq \rho_{j.} \end{cases} \quad (1)$$

Threshold-based approaches differ in their procedure of selecting the appropriate thresholds. They either rely solely on response time information (e.g., Wise & Ma, 2012) or additionally include information gained from item responses or process data (e.g., Sahin & Colvin, 2020; Wise, 2019). The most simplistic is the common-*k* approach, which defines a uniform response-time threshold for all items based on theoretical considerations (e.g., no person can read an item in less than three seconds;

Wise et al., 2004). While this heuristic method is simple to apply, it does not account for item-specific differences in time demands. Other approaches incorporate item-specific response time information into threshold selection. The visual threshold method, for example, assumes that rapid guessing and solution behavior follow distinct response time distributions, typically resulting in a bimodal overall distribution (Wise et al., 2009). The threshold is set at the intersection of these two distributions. However, in small samples or when rapid guessing is infrequent, the bimodal structure may be unclear, rendering this approach potentially difficult to apply. Alternatively, the normative threshold method defines a threshold as a fixed percentage of the average item response time (e.g., 10%), independent of the observed distribution (Wise & Ma, 2012). Other approaches incorporate additional information into the threshold selection, for example by plotting the average accuracy (Guo et al., 2016; Lee & Jia, 2014) or the average correlation between item response and total score (Wise, 2019) per response time bin. The threshold is set at the response time at which the accuracy or correlation exceeds chance or a theoretically determined value. However, these approaches require huge sample sizes for being able to estimate a meaningful average accuracy or correlation per response time bin.

Threshold-based approaches are widely used in research and in practice, although the method of setting a threshold may vary. Their main advantage is that they are both easy to comprehend and to apply, while still offering plenty of flexibility to the researcher as they can choose from a multitude of methods. Furthermore, they do not require long computation times. However, setting thresholds can still be a cumbersome procedure. Non-automated methods of threshold selection require item-based analyses and decisions, which can become very time consuming in tests with large item sets. Moreover, some threshold-setting methods still require huge sample sizes (e.g., when plotting accuracy or item-total correlation by response time for each item) or specific response time distributions (e.g., bimodal distribution for the visual inspection method) – prerequisites that prevent the accurate selection of a threshold if not met (Wise, 2019). Lastly, the deterministic approach to classifying test-taking engagement does not include a measure of uncertainty, potentially leading researchers and practitioners to overestimate the accuracy of classification.

Model-based approaches to identifying disengaged responding overcome these issues by modeling test-taking engagement as a latent construct (Leventhal & Pastor, 2023; Nagy & Ulitzsch, 2022; Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015). The vast majority of these approaches are based on mixture models that define test-taking engagement as a latent class variable at the response level, consisting of the two latent classes *solution behavior* and *disengaged responding* (Nagy & Ulitzsch, 2022; Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015). In the solution behavior class, the probability of a correct response depends both on the examinee's proficiency and on item characteristics, whereas in the disengaged responding class, the probability corresponds to chance level. Thus, both the accuracy and the psychometric information (i.e., information on person and item parameters) are assumed to differ between engaged and disengaged responses. Item response times can be incorporated to improve classification accuracy either by defining distinct response time distributions for each latent class (Meyer, 2010; Ulitzsch et al., 2020; Wang & Xu, 2015) or by regressing test-taking engagement on response times (Pokropek, 2016).

The main strengths of using model-based approaches for measuring disengaged responding lie in the joint modeling of test-taking engagement and proficiency, thereby combining the identification of disengagement and its incorporation into proficiency estimation in one step. This approach allows a detailed understanding of the complex relationships between response time, proficiency, and test-taking engagement, while also accounting for the uncertainty in the classification of test-taking engagement by assigning each response a probability of disengagement. By integrating response times, accuracy, and psychometric information in the identification of disengagement, these approaches may additionally facilitate an improved classification compared to solely response-time based methods. Furthermore, model-based methods can account for person-level effects of disengagement and processing speed. However, these approaches are also computationally intensive, require large sample sizes, and rely on distributional assumptions for response times that may not always hold (Molenaar et al., 2018).

A common drawback of the presented response-time based identification methods is that the binary classification in engaged or disengaged may be over-simplistic, with recent research suggesting

that not all responses classified as solution behavior may indeed be engaged (Nagy et al., 2022; Wise & Kuhfeld, 2021). By limiting the definition of disengagement to excessively fast responses, the problem posed by low test-taking engagement to competence assessments may be underestimated. However, identifying non-rapid or partial disengagement entails its own difficulties, with only few studies so far having investigated the possibilities of measuring disengagement beyond rapid guessing (Nagy et al., 2022; Wise & Kingsbury, 2022; Wise & Kuhfeld, 2021). Therefore, the majority of research on test-taking engagement relies on the presented threshold-based and model-based methods for identifying disengaged responses.

Since there is no perfect, one-size-fits-all approach to measuring disengaged responding, it is crucial to carefully consider and validate the choice of method. Moreover, empirical results should be interpreted with caution, as they are only estimates and heavily depend on the chosen identification approach.

2.3 Accounting for Disengaged Responding in Proficiency Estimation

Because disengaged responding introduces construct-irrelevant variance to proficiency estimates, it should be accounted for in the measurement of proficiencies. In LSAs, proficiency is typically estimated using IRT models, which assume that item responses depend on both a person's proficiency and item characteristics. For instance, the Rasch (1960) model defines the probability of a correct response, $P(Y_{ij} = 1)$, of person *i* on item *j* as a function of the person's proficiency θ_i and the item difficulty β_j:

$$P(Y_{ij} = 1) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \quad (2)$$

While the Rasch model is designed for dichotomous items and assumes that items differ only regarding their difficulties, other IRT models are also suitable for polytomous items (e.g., partial credit model, Masters, 1982) and/or include additional item parameters such as item discriminations (e.g., 2PL model, Birnbaum, 1968). Proficiency estimates can be provided either as manifest scores (e.g., weighted likelihood estimates, Warm, 1989) or as plausible values. Researchers and practitioners often rely on plausible values because they account for measurement error and complex sample designs

(Mislevy, 1991; Wu, 2005). Their estimation is based on both (a) the measurement (i.e., IRT) model and (b) a population model consisting of a latent regression of proficiency on background variables (Wu, 2005; see also Chapter 3.2 and Article 2).

Model-based approaches to identifying disengaged responding combine the identification of disengagement and the estimation of proficiency in one step (e.g., by defining two different latent variables, see Chapter 2.2), thereby accounting automatically for potential distorting effects of low test-taking engagement. Threshold-based approaches, on the other hand, require the selection of a scoring method once disengaged responses have been identified. One common approach is to apply listwise deletion to examinees who were disengaged on a non-negligible proportion of items (e.g., more than 10%; also referred to as *examinee-level filtering*). However, Rios and colleagues (2017) demonstrated in a simulation study that this approach led to an upward bias in competence estimates of up to 0.40 standard deviations (SD) when disengagement was related to proficiency. As a result, they recommended accounting for disengaged responding on the response level.

A recent simulation study by Rios and Deng (2024) compared different response-level scoring approaches and concluded that effort-moderated scoring, both with and without imputation, performed best across conditions with small error rates and minimal sensitivity to misclassifications. Because imputation adds computational complexity and provides similar parameter recovery to the non-imputed approach (Rios & Deng, 2024), the non-imputing model is preferable. Effort-moderated scoring involves recoding all disengaged responses as missing values (i.e., *response-level filtering*), following the effort-moderated model (EMM) by Wise and DeMars (2006). The EMM assumes that disengaged responses provide no meaningful information about proficiency or item difficulty. However, even when this assumption was violated, the EMM still yielded accurate results (Rios & Deng, 2024).

Since plausible values additionally include a latent regression in the estimation of proficiency (Wu, 2005), they provide more possibilities for scoring disengagement (see Chapter 3.2). However, no methods have yet been proposed and evaluated to account for disengagement specifically in plausible value estimation.

2.4 Prevalence and Antecedents of Disengaged Responding

Because most identification methods rely on flagging responses with unusually short response times, a large proportion of empirical research has focused on rapid guessing. Over the past two decades, numerous studies have identified non-negligible amounts of rapid guessing across different low-stakes assessments. Estimated rapid guessing rates vary significantly, with reported prevalences ranging between < 1% (Goldhammer et al., 2017; Kroehne et al., 2020; Michaelides & Ivanova, 2022) and nearly 20% (Ulitzsch et al., 2020). A meta-analysis by Rios et al. (2022) based on 25 studies applying threshold-based identification methods estimated that, on average, 7% of the responses were classified as rapid guesses, while 28% of the test takers engaged in rapid guessing on at least one item. Model-based approaches tend to identify even more rapid guessing, with estimated response-level prevalences ranging between 8% and 18% (Meyer, 2010; Nagy & Ulitzsch, 2022; Pokropek, 2016; Ulitzsch et al., 2020).

Why do the reported prevalences differ so much? First, the estimated rapid guessing rate depends heavily on the applied identification method – that is, different *estimated* prevalences do not necessarily mean that the *true* prevalence differs. For example, Goldhammer et al. (2017) report rapid guessing rates that range between nearly 0% and over 8% in the same problem solving assessment from PIAAC, solely depending on which threshold method was applied. In the meta-analysis by Rios et al. (2022), the reported rates differed significantly between identification methods, with normative thresholds classifying three times more rapid guessing than common-*k* thresholds. Furthermore, model-based approaches seem to generally estimate disengagement to be more prevalent than threshold-based approaches (e.g., Pokropek, 2016; Ulitzsch et al., 2020).

Second, the reported prevalence rates do not only depend on the identification method, but also on the unit of reference. Prevalences can be reported both as the proportion of disengaged responses per item or test (e.g., Goldhammer et al., 2016; Kuang & Sahin, 2023) or as the proportion of disengaged test takers – that is, the number of test takers who engage in rapid guessing on a non-negligible amount of items (e.g., Jensen et al., 2018; Kuhfeld & Soland, 2020). How this “non-negligible

amount" is defined differs between studies, with the majority using the criterion of 10% (Rios et al., 2022).

Third, findings from studies that use the same methods to estimate rapid guessing rates for different assessments and samples indicate that the actual prevalence can also vary widely. For example, estimated prevalence rates differ between countries, test domains, test settings, and assessment modes (Goldhammer et al., 2017; Kroehne et al., 2020; Michaelides & Ivanova, 2022; Rios & Soland, 2022; Wise et al., 2021).

While different methodological approaches and factors at the sample and test level account for differences in average disengagement rates, antecedents at the item and person level explain variation within samples and tests. Identifying item-level predictors is important for designing items that encourage examinees to invest effort into their responses. For example, rapid guessing has been shown to occur less often in interactive and open-response format items, whereas long and complex texts seem to induce disengagement (Avvisati et al., 2024; Michaelides & Ivanova, 2022; Rios & Soland, 2022) – a finding that provides a promising avenue for redesigning future assessments. In contrast, person-level predictors are critical for identifying potential threats to the validity of group-based assessment outcomes, because differential disengagement may lead to biased group comparisons. Previous research has shown that rapid guessing depends on individual characteristics such as gender, socio-economic status (SES), linguistic background, and ethnicity. Men, individuals with low SES, non-native speakers, and minorities tend to be less engaged than their peers (Goldhammer et al., 2016; Kroehne et al., 2020; Soland, 2018; Wise et al., 2021).

One particularly important finding is that less proficient individuals tend to be less engaged in competence assessments than their more proficient counterparts. In a meta-analysis by Silm et al. (2020), the average correlation between test performance and response-time based measures of test-taking engagement was $r = .72$. However, this association can be caused by both (a) high proficient individuals being more engaged or (b) highly engaged test takers achieving a better test score. Therefore, studies are more informative when they examine the relationship between test-taking engagement and competence measures that are either unaffected by low engagement or adjust for it.

While Silm et al. (2019) report a correlation of r = .23 between test-taking engagement and previous high-stakes test results, Rios et al. (2017) found differences of up to 1.50 SD in high-stakes test scores between examinees who were classified as disengaged versus engaged on a subsequent low-stakes assessment. In mixture models that account for disengagement, the estimated relationship between test-taking engagement and proficiency ranged from r = .59 to r = .72 (Nagy & Ulitzsch, 2022; Ulitzsch et al., 2020). The findings suggest that the relationship between test-taking engagement and proficiency may be moderate to strong, highlighting the importance of using methods that account for test-taking engagement without assuming a zero correlation between the two constructs.

To summarize, estimated prevalence rates vary greatly across assessments and identification methods, reflecting both real differences in disengagement and methodological factors. Identifying person-level antecedents of disengaged responding is essential for detecting potential threats to the validity of group-based assessment outcomes and, by extension, to inequality research in education. While some antecedents have been identified in adolescent and young adult samples (e.g., gender, SES, linguistic background, ethnicity, proficiency; Goldhammer et al., 2016; Kroehne et al., 2020; Soland, 2018; Wise et al., 2021), evidence from other age groups remains scarce.

2.5 Consequences for the Validity of Proficiency Estimates

When test takers are not engaged, their responses do not reflect their true proficiency, thereby violating a fundamental assumption of competence assessment (Wise, 2017; see also Chapter 2.1). Thus, disengaged responding may have severe consequences for the validity of proficiency estimates derived from LSAs. If the test scores do not only reflect proficiency but also motivation, how can we draw inferences from findings based on them? Can group differences in test scores (e.g., between genders) truly be attributed to differences in proficiency, or were the groups just differently motivated? Do predictors of competence estimates (e.g., SES) indeed affect proficiency, or rather test-taking motivation?

Due to the potentially grave consequences for the interpretability of assessment outcomes, numerous studies have investigated the effect of low test-taking motivation on proficiency estimates. Since a relationship between test-taking motivation and test performance can also be caused by an

underlying correlation between proficiency and motivation, it is crucial to draw upon other criteria when evaluating the impact of disengagement. One possible approach is to compare the test performance of high- versus low-motivated groups in (quasi-)experimental settings (e.g., by offering one group incentives for correct responses, but the other not). A meta-analysis from Wise and DeMars (2005) reported that across 25 effect sizes, the motivated group performed, on average, 0.59 SD (range: –0.04 to 1.49) better than the unmotivated group.

Another approach evaluates the convergent validity of proficiency estimates – that is, to what extent test scores correlate with external criteria that are not or barely affected by low test-taking motivation. The correlation between test performance and high-stakes test results was, across six studies, reported to increase by .08 to .12 (*Mdn* = .11) when unmotivated examinees (i.e., rapid guessing on more than 10% of the items) were filtered from the data (Wise, 2015). Lastly, it is possible to compare test scores estimated (a) when ignoring low test-taking engagement and (b) when accounting for it. A meta-analysis by Rios et al. (2022) reported that aggregated test scores increased, on average, by 0.13 SD when accounting for rapid guessing.

While these studies agree that test-taking engagement affects proficiency estimates, they mostly focus on *average* score distortion. However, educational LSAs usually provide information at the population level rather than at the individual level. Therefore, and because proficiency is defined as a latent construct with no meaningful absolute scores, I argue that average score distortion does not necessarily have any consequences for actual assessment outcomes such as ability comparisons, competence development across time, or predictors of proficiency. If *every* individual score was distorted to the same extent, analyses based on the estimates (e.g., correlations with other variables) would not be affected. Therefore, low test-taking engagement should bias assessment outcomes if proficiency estimates of different subpopulations (e.g., girls vs. boys) or at different measurement occasions are distorted to varying degrees – that is, when *differential* score distortion is present.

To understand the potential effects of disengagement on assessment outcomes, it is important to know which factors contribute to score distortion. The simulation study by Rios et al. (2017) demonstrated that the extent of distortion depends not only on the rapid guessing rate, but also on

test difficulty. Proficiency estimates were significantly underestimated (i.e., Cohen's $d > 0.20$) when the rapid guessing rate exceeded 6% on easy tests or 12% on moderate or difficult tests. This pattern can be explained by the increasing distortion of proficiency estimates as the probability of a correct response diverges between engaged and disengaged responding. Because correct-response probabilities under engagement are higher for easier items, the discrepancy between engaged and disengaged performance is larger in easier tests.

Studies investigating the impact of rapid guessing on actual assessment outcomes report inconclusive findings. When applying response-level filtering (see Chapter 2.3), the effects on achievement gap estimates (Kuhfeld & Soland, 2020; Soland, 2018), school effectiveness and summer learning loss (Kuhfeld & Soland, 2020), as well as school rankings (Rios & Guo, 2020; Wise, 2019; Wise et al., 2020, 2021) were minimal to non-existent, even when the disengagement rates varied greatly between subpopulations. Interestingly, the only studies that report noteworthy effects (e.g., on institutional rankings, teacher value added, or competence development) used examinee-level filtering (Jensen et al., 2018; Setzer et al., 2013; Wise & Demars, 2010), a scoring method that was previously shown by Rios et al. (2017) to distort test scores of up to 0.40 SD when proficiency and engagement are related. These findings highlight that the effect of *accounting for* low test-taking engagement reflects both the actual effect of disengagement *and* the effect of the scoring method itself. That is, estimated effects may not necessarily be caused by rapid guessing but may instead be caused by removing responses or examinees from the analyses. Nevertheless, other studies indicate that rapid guessing may still affect item parameters (Alahmadi & DeMars, 2025), model fit (Rios & Guo, 2020), metric measurement invariance (Rios, 2021), and score equating (Deng & Rios, 2022; Rios & Guo, 2020).

In summary, research on the impact of low test-taking engagement is still inconclusive, highlighting the need for additional studies investigating the consequences of disengaged responding in different contexts as well as determining its influencing factors in simulation studies.

3 Rationale for Dissertation

Despite the large body of research on disengaged responding, important questions still remain unanswered. This dissertation aims to contribute to the disengagement literature by addressing three of the most consequential research gaps: (1) misclassifications in response-time based identification of disengaged responding, (2) missing approaches of accounting for disengaged responding in plausible value estimation, and (3) generalizability of established findings on disengaged responding to child samples. The following paragraphs present each research gap and outline the rationale for each of the three articles included in this dissertation.

3.1 Misclassifications in Response-Time Based Identification of Disengaged Responding

Existing detection methods of disengaged responding focus either on person-level disengagement (e.g., self-report, person-fit statistics) or on identifying rapid guessing using item response times (Wise, 2020). While person-level identification ignores that test-taking engagement may not be constant throughout the test, response-time based approaches bear the risk of wrongly classifying fast engaged responses as disengaged or slow disengaged responses as engaged (Wise, 2020). This problem highlights the need for new approaches that identify disengagement at the response level using information beyond item responses and response times.

With the advent of computer-based assessments, process data become available to researchers and practitioners. Goldhammer and Zehner (2017) define process data “as the empirical data reflecting the course of working on a test item” (p. 128). Besides item response times, these data can include much more information on the test-taking behavior of an examinee (e.g., mouse clicks, navigation through the test, eye movements). Under the assumption that disengaged examinees differ in their test-taking behavior from engaged examinees in more than the time they need for responding, this additional information can be a rich source for improving the classification accuracy of disengagement.

As of now, only few studies have investigated the potential of additional process data for the identification of disengaged responding. Additionally, the few existing studies have mostly focused on interactive or constructed response items (Ivanova et al., 2020; Ivanova & Michaelides, 2023; Lundgren

& Eklöf, 2020; Sahin & Colvin, 2020), ignoring the fact that traditional item formats such as multiple-choice items still prevail in most competence assessments. Furthermore, only two studies (Ivanova & Michaelides, 2023; Sahin & Colvin, 2020) proposed approaches to incorporating the additional indicators in the identification of disengaged responding. These two studies showed that the number and type of actions in interactive and constructed response items may improve – in some cases – threshold-based detection methods.

In summary, process data beyond item response times may improve the classification accuracy of existing identification approaches to disengaged responding. However, research on this promising avenue remains scarce, particularly with regard to studies that propose and evaluate approaches to incorporating additional process data into the measurement of disengagement. Available studies focus on improving threshold-based approaches, thereby ignoring that model-based methods with their probabilistic approach to classification may be more suitable for accommodating additional information on the test-taking process. Article 1 addressed this gap by evaluating the incremental value of novel process data indicators for the identification of disengaged responding based on an established mixture modeling approach (Nagy & Ulitzsch, 2022; Pokropek, 2016).

3.2 Missing Approaches to Accounting for Disengaged Responding in Plausible Value Estimation

Once disengaged responses have been identified, they need to be accounted for in the estimation of competencies to ensure unbiased proficiency estimates. As discussed in Chapter 2.3, many approaches have been proposed and evaluated to incorporate (proxies of) test-taking engagement in item response modeling (e.g., EMM, Wise & Demars, 2006) or more complex latent mixture models (see Nagy & Ulitzsch, 2022, for an overview). However, many LSAs provide proficiency estimates as plausible values for the scientific community (e.g., PISA, Buchholz et al., 2022), because plausible values combine the strengths of accounting for measurement error and for complex sampling designs with avoiding the computational demands of latent modeling (Mislevy, 1991; Wu, 2005).

The plausible value approach argues that, as an unobservable latent construct, an examinee's proficiency is unknown to the researcher and should therefore be treated as missing (Mislevy, 1991; Wu, 2005). Consequently, the approach follows the concept of multiple imputation (Rubin, 1987) by

drawing a pre-defined number of plausible value sets from the posterior probability distribution of the latent proficiency (Mislevy, 1991). This distribution is a function of (a) a measurement model estimating proficiency based on the item responses, and (b) a population model consisting of a latent regression of proficiency on background variables (Wu, 2005; see also Article 2 for a formal introduction to plausible values). Therefore, plausible values allow for two different approaches to incorporating test-taking engagement in their estimation process: either by accounting for it in (a) or in (b).

While Buchholz et al. (2022) suggests that including disengagement in the latent regression suffices for accounting for its distorting effects in plausible value estimation (an approach already implemented in PISA, Buchholz et al., 2022), to my knowledge, no previous study has evaluated this claim. However, solely including disengagement in the latent regression may have two pitfalls. First, this approach assumes that the information gained from the item responses are correct, that is, that the responses in fact reflect proficiency. Second, existing approaches to handling rapid guessing have shown that accounting for it on the examinee level can itself bias competence estimation (Rios et al., 2017).

To summarize, no study has yet investigated to which extent plausible values can be distorted by low test-taking engagement and which methods may reduce these biases, although they are widely used in the LSA context and their estimation relies partly on different mechanics than common IRT models. Therefore, to ensure unbiased proficiency estimates in LSAs, it is crucial to develop approaches to accounting for disengaged responding in plausible value estimation. Article 2 addressed this gap by proposing and evaluating different plausible value models that do (not) account for rapid guessing in both a simulation and an empirical study.

3.3 Generalizability of Findings on Disengaged Responding to Child Samples

To understand early competence development and its determining factors, many LSAs include child samples (e.g., *Trends in International Mathematics and Science Study*, *National Educational Panel Study*). However, most findings on low test-taking engagement are based on adolescent or adult samples. Because children differ from older individuals in their cognitive development (Piaget, 1970/2022), they may behave differently in competence assessments. For example, children's

increased overconfidence in their abilities (van Loon et al., 2017; Xia et al., 2024) could lead to a stronger expectancy of success or perceived capacity. Moreover, a lack in strategic and prospective thinking (Lombardi et al., 2017) may lead to higher perceived importance of the test. Consequently, children's test-taking engagement could differ from those of older individuals, raising questions about the generalizability of findings based on other samples.

The few empirical studies that rely on child samples for investigating low test-taking engagement suggest that disengagement may indeed depend on the age of the examinees, with rapid guessing rates increasing throughout childhood and adolescence (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023; Wise & Kingsbury, 2022). However, these findings are limited to U.S. samples and classroom-based settings. To gain a more generalizable understanding of children's test-taking behavior, it is important to replicate these results in other regions and settings. Furthermore, no studies have explicitly tested whether the potential differences in test-taking engagement also affect other established findings from older samples such as the antecedents of disengagement or its impact on the validity of assessment outcomes.

Although theoretical considerations suggest that children's test-taking behavior may differ from that of older examinees, empirical research on disengaged responding in this age group remains scarce. Therefore, empirical studies are required that systematically evaluate the generalizability of established findings from adolescent or adult samples (e.g., predictors of test-taking engagement) to child samples. Article 3 addressed this gap both by outlining theoretical considerations why children may differ from older individuals in their test-taking engagement and by empirically investigating the antecedents, prevalence, and impact of rapid guessing in a sample of German elementary school students.

4 Summary of Articles

This dissertation contains three articles that each addressed a unique aspect of disengaged responding in educational LSAs. To facilitate understanding of the integrated discussion in Chapter 5, the following paragraphs briefly summarize each article's research questions, methods, findings, and main conclusions.

4.1 Article 1: Identifying Disengaged Responding in Multiple-Choice Items: Extending a Latent Class Item Response Model with Novel Process Data Indicators

Article 1 (Welling et al., 2024) sought to improve the identification of disengaged responding by leveraging the potential of process data beyond item response times. Thereby, it addressed the problem that test-taking engagement can be misclassified in response-time based approaches to identification. The research aims were (a) to derive novel process data indicators of test-taking engagement in multiple-choice items of reading comprehension tasks, (b) to incorporate these indicators into an existing model-based identification method, and (c) to evaluate the incremental value of the new indicators.

Three novel process data indicators of test-taking engagement in reading comprehension tests were derived based on theoretical considerations: *text reread*, *item revisit*, and *answer change*. A text reread was defined "as the navigation from an item page back to the corresponding text page" (p. 321), whereas an item revisit was "the repeated navigation to an item page" (p. 321). An answer change was defined "as changing the previously selected response option to another one" (p. 321). All three indicators were assumed to indicate high test-taking engagement, since they reflect a thorough interaction with the competence test.

To evaluate the incremental value of the novel indicators for accurately identifying disengaged responding, they were incorporated into an existing model-based identification method – the dependent latent class item response (DLC-IRT) model originally proposed by Pokropek (2016) and subsequently advanced by Nagy and Ulitzsch (2022). In this model, the probability of belonging to the solution behavior class is regressed on item response times (see Chapter 2.2). Three models were proposed and compared: (1) a *baseline model* that included response times as the only predictor of

engagement, (2) an *extended model* that included the novel indicators as additional predictors, and (3) an *interaction model* that included the three indicators and all six interaction terms between the new indicators and response time as predictors. All models were estimated based on data from a German sample of (former) university students (N = 1,932).

Descriptive results showed that text rereads (*Mdn* = 35.8%) occurred more often than answer changes (*Mdn* = 13.5%) and item revisits (*Mdn* = 5.3%). Comparing the three proposed models using information criteria and deviance tests revealed that the extended model with the novel indicators as additional predictors of engagement fit the data best. This model classified 13.3% of the responses to be disengaged and estimated the correlation between proficiency and test-taking engagement to be perfect (r = 1.00). Descriptively, more text rereads, item revisits, and answer changes occurred in the solution behavior class than in the disengaged responding class. However, only response time and text reread were significant predictors of test-taking engagement, with longer response times and *less* text rereads indicating higher engagement. Explorative analyses revealed that the new indicators seemed not to improve the classification accuracy of disengaged responding.

While the results partly indicate that additional process data indicators may indeed improve the identification of disengaged responding (i.e., the model including the new indicators fit the data best), other results were contrary to our assumptions. Although responses classified as engaged were associated with more test interactions than responses classified as disengaged, none of the novel indicators predicted test-taking engagement in the hypothesized direction. One possible explanation lies in suppression effects: the novel indicators may not predict engagement *beyond* the effect of response times. Furthermore, most test takers did not interact with the test interface apart from providing a response. The variance in the new predictors may therefore be too small to meaningfully establish an association with test-taking engagement. Additionally, we argued that test-taking engagement and proficiency may have been to some extent confounded in the DLC-IRT model.

Article 1 makes three key contributions to the disengagement literature. First, it demonstrates how process data from LSAs can be leveraged based on theoretical considerations. Second, it illustrates how such data can be processed and made available for methodological research. Third, it advances

our understanding of test-taking engagement beyond short response times, thereby contributing to the development of more accurate methods for identifying disengagement.

4.2 Article 2: Comparing Different Approaches of (Not) Accounting for Rapid Guessing in Plausible Values Estimation

Article 2 (Welling et al., 2026b) sought to determine (a) the threat rapid guessing poses to the validity of plausible value estimates and (b) to what extent different approaches to accounting for it can minimize potential biases. The article consists of three parts. First, it proposes different approaches to account for rapid guessing in plausible value estimation. Second, it evaluates both the bias introduced by rapid guessing and the performance of the proposed approaches in a simulation study. Third, it illustrates the practical applicability of these approaches in an empirical study. Because plausible values allow accurate inferences on proficiencies at the population level rather than at the individual level (Wu, 2005), and because differential score distortion is more important for the validity of assessment outcomes than average score distortion (see Chapter 2.5), the research questions were addressed in the context of ability group comparisons.

In contrast to other approaches to measuring proficiency, plausible value estimation relies both on a measurement and on a population model (see Chapter 3.2 or Article 2). Therefore, three different approaches to accounting for rapid guessing were proposed: (1) the *response-level model* that accounts for rapid guessing in the measurement model by recoding rapid guesses as missing (following the EMM by Wise and DeMars, 2006), (2) the *person-level model* that accounts for rapid guessing in the background model by including a proxy for disengagement as a predictor of proficiency (as suggested by Buchholz, 2022), and (3) a *combined model* that integrates both approaches in one model.

The simulation study generated data for two hypothetic groups based on the hierarchical mixture model proposed by Wang and Xu (2015). Three factors were systematically varied: (a) the overall rapid guessing rate, (b) the size of the ability difference between the two groups, and (c) the difference in rapid guessing rate between both groups. Rapid guessing was identified using two different methods of setting response-time thresholds: the common-*k* (i.e., fixed) and the visual thresholds (see Chapter 2.2).

Both threshold methods classified engaged responses almost perfectly, but the fixed thresholds more often misclassified rapid guesses as engaged than the visual thresholds. The distortion of ability group comparisons due to rapid guessing was generally small (bias < 0.16 SD), but increased under three conditions: when (i) the overall rapid guessing rate was high (15%), (ii) the achievement gap was large (d = 1.0), and (iii) the more competent group exhibited a higher rapid-guessing rate than the less competent group. The person-level model did not reduce the bias introduced by rapid guessing; its performance was barely distinguishable from the model ignoring rapid guessing. In contrast, the response-level and combined model effectively minimized the biases. However, their performance depended on the classification accuracy of the chosen threshold method when the overall rapid guessing rate was high: models based on the fixed threshold slightly underestimated the ability difference, whereas models based on the visual threshold slightly overestimated it.

In the empirical study, participants (N = 478) of a reading comprehension test were divided into two groups based on their study subject. One group exhibited more rapid guessing (3.5% using the fixed threshold, 4.0% using the visual thresholds) than the other (1.6% using the fixed threshold, 2.2% using the visual thresholds). However, the estimated difference in reading comprehension between both groups (0.59 ≤ Cohen's d ≤ 0.66) did not differ between any of the models accounting for rapid guessing or the model ignoring it. The results are consistent with the simulation study under the condition of a low overall rapid-guessing rate, a moderate ability gap, and higher rapid guessing in the more competent group.

These findings suggest that ignoring rapid guessing in the estimation of plausible values can compromise the validity of ability group comparisons (bias > 0.10 SD) when the overall rapid guessing rate is high and when the more competent group exhibits more rapid guessing than the less competent group. In contrast to Buchholz's (2022) claim, accounting for rapid guessing only in the background model does not seem to suffice to reduce potential biases. Instead, researchers should resolve to account for rapid guessing in the measurement model but should also be aware that the classification accuracy of the chosen threshold method may affect the performance of the models.

Article 2 provides three main contributions to the disengagement literature: First, it improves our understanding of the impact of low test-taking engagement on the validity of assessment outcomes by reporting the first simulation study to determine the bias introduced by rapid guessing (a) in ability group comparisons and (b) using plausible values. Second, it reports the first study to propose and evaluate different approaches to accounting for rapid guessing in plausible value estimation, challenging the assumption that rapid disengagement can be accounted for by incorporating background variables in the population model (Buchholz et al., 2022). Third, it supports the findings from a previous simulation study (Rios & Deng, 2024) that different approaches to accounting for rapid guessing are sensitive to the classification accuracy of the chosen threshold method.

4.3 Article 3: Prevalence, Predictors, and Consequences of Rapid Guessing among Children in Elementary School

Article 3 (Welling et al., 2026a) sought to investigate the generalizability of established findings on rapid guessing to child samples. Specifically, it addressed the following research questions: (a) How does the prevalence of rapid guessing change with increasing age? (b) Which person-level characteristics predict rapid guessing in child samples?, and (c) To what degree does rapid guessing bias ability group comparisons in child samples? Corresponding hypotheses were developed based on both the expectancy-value theory and the demands-capacity model (see Chapter 2.1).

Due to the decreasing overconfidence (van Loon et al., 2017; Xia et al., 2024) and increasing strategic and prospective thinking across childhood (Lombardi et al., 2017), we assumed that young children should exhibit a high expectancy of success, perceived capacity, and importance of the test, therefore being more engaged than older children. We further expected that boys, children with low SES, and less proficient children should be less engaged than their peers, because (i) gender differences in personality are already pronounced in early childhood (Slobodskaya & Kornienko, 2021), (ii) more proficient children should exhibit an increased expectancy of success, perceived capacity, and task enjoyment, and (iii) SES should positively influence expectancy of success and perceived capacity. Lastly, since the impact of rapid guessing depends – among other factors – on its prevalence, we assumed that the impact on group comparisons should increase across childhood.

The longitudinal two-wave study used data from German elementary school students (*N* = 1,877) who were administered a reading comprehension and mathematics test in Grades 2 and 4. The setting differed slightly between the two assessments because restrictions due to the COVID-19 pandemic were in place in Grade 2. Whereas the reading test was identical in both grades, the mathematics test was not, rendering longitudinal comparisons meaningful only for the reading domain. Rapid guessing was identified using two different threshold methods for each assessment (see Chapter 2.2). They were validated at test level based on (a) the average accuracy and score-total correlation for responses classified as engaged or disengaged, and (b) the average accuracy plotted by response time across all items. Proficiency was estimated twice: once when ignoring rapid guessing and once when scoring rapid guesses as missing (following the EMM by Wise and Demars, 2006; see also Chapter 2.3).

Based on the threshold methods with the best validation criteria, 3.3% of the responses were classified as rapid guesses in Grade 2 reading, compared to 0.5% in Grade 4 reading. Thus, the prevalence of rapid guessing did not increase between the two grades. In the mathematics tests, 0.8% of the responses were classified as rapid guesses in Grade 2, compared to 0.4% in Grade 4. Descriptive results showed that boys, children with a low SES, and less proficient children indeed exhibited rapid guessing more often than their peers. However, only engagement-adjusted proficiency was a significant predictor of test-taking engagement. Across all four assessments, rapid guessing had no impact on ability group comparisons between different genders or different levels of SES.

Overall, the rapid guessing rates were rather small compared to prevalences usually reported in adolescent or adult samples. However, they varied considerably between the different assessments, with no clear trend discernible (e.g., grade or domain-specific), suggesting that the disengagement rates depended on a combination of test setting, administration mode, sample, and the applied threshold method. We see two potential causes for the null effects regarding the predictors of engagement: either the effects were too small to be statistically detectable (especially in rare events data) or gender and SES do not strongly influence test-taking engagement in childhood. That rapid guessing had no impact on ability group comparisons aligns with previous findings on the non-impact of low test-taking engagement on group-based assessment outcomes.

Article 3 contributes to the disengagement literature in three major ways. First, it explicitly challenges the assumption that findings on rapid guessing are generalizable to child samples by outlining theoretical considerations why test-taking engagement in childhood may differ from adolescence and adulthood. Second, it provides first or further empirical evidence on the prevalence, predictors, and consequences of rapid guessing in child samples. Third, it provides further evidence on the impact of low test-taking engagement on the validity of assessment outcomes, thereby deepening our understanding of the consequences of disengaged responding for educational LSAs.

5 Discussion

Low test-taking engagement violates a fundamental assumption from the measurement of competencies: that the responses of the test takers reflect their proficiency in the assessed construct. Therefore, disengaged responding may have severe consequences for competence assessments in educational LSAs (Wise, 2017; Wise & Demars, 2005). This cumulative dissertation enhances our understanding of disengaged responding by addressing key research gaps regarding its identification, modeling, and consequences across three studies. Following a brief summary of the central findings of each article, this chapter provides an integrated discussion that highlights overarching findings, implications, limitations, and conclusions, as well as the main contributions of this dissertation.

5.1 Summary of Key Findings

Article 1 introduced three potential process data indicators of disengaged responding in multiple-choice items of reading comprehension tasks: text reread, item revisit, and answer change. An empirical study found that the additional indicators only scarcely improved the classification of test-taking engagement. Nonetheless, explorative results suggested that in the underlying mixture model, test-taking engagement may have been at least partly confounded with proficiency, since both constructs correlated perfectly.

The simulation study in Article 2 demonstrated that under certain circumstances, ability group comparisons based on plausible value estimates may be biased by rapid disengagement. Furthermore, it indicated that solely accounting for rapid guessing in the population model of plausible values may not suffice to reduce potential biases. An additional empirical study in Article 2 illustrated that the different approaches to (not) accounting for rapid guessing yielded comparable estimates of ability differences, supporting recent research (Kuhfeld & Soland, 2020; Soland, 2018) suggesting that its impact on group comparisons may be small in practice.

Article 3 provided theoretical considerations on test-taking engagement among children. Additionally, an empirical study replicated previous findings from U.S. samples (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023; Wise & Kingsbury, 2022) on low rapid guessing rates among elementary school students in a German sample. The study further

indicated that the influence of common predictors of rapid guessing (i.e., gender, SES) may not yet be pronounced in childhood, challenging the generalizability of findings from older samples. Lastly, the results suggested that also among children, ability group comparisons may not be distorted.

5.2 Integrating Findings Across Articles

The four studies included in the three articles of this dissertation addressed different research questions, used different identification and scoring methods, and were based on both simulated and empirical data. However, there are several findings that are consistent across multiple studies.

5.2.1 Response-Time Based Measurement of Disengaged Responding

Despite the disadvantages posed by response-time based approaches to identifying disengaged responding (Wise, 2020), this dissertation supports previous findings indicating that response times are indeed a strong indicator of rapid disengagement (Ivanova & Michaelides, 2023; Patel et al., 2021; Wise, 2020). Article 1 found that even in the presence of additional information on the test-taking process, response times remained an important predictor of disengaged responding. Although item revisits and answer changes were, descriptively, associated with higher test-taking engagement, they possessed no incremental predictive value over and above item response times. These results align with previous findings on additional process data indicators (Ivanova & Michaelides, 2023; Patel et al., 2021; Sahin & Colvin, 2020). Patel et al. (2021) applied a machine learning approach to evaluate the predictive value of different process data indicators for test-taking efficiency. Even though the authors were able to identify some additional indicators (e.g., answer change and navigation button use), most significant predictors were related to test or item response time. Moreover, Sahin and Colvin (2020) evaluated the incremental value of additional process data for threshold-based approaches and concluded that the improvements in the identification of disengaged responding were marginal. Ivanova and Michaelides (2023) evaluated two different threshold-setting methods in PISA data. While thresholds based on response times yielded satisfactory validation results across all items, thresholds based on the number of actions on the item did not. In summary, the incremental value of process data beyond response times for improving the classification of test-taking engagement may be limited.

Articles 2 and 3 relied solely on response-time based approaches for identifying disengaged responding. Results indicated that across samples, assessments, and most threshold methods, these approaches met pre-specified validation criteria (e.g., comparing average accuracy between responses classified as engaged or disengaged), suggesting that they successfully identified rapid guessing. Although these studies did not compare the performance of solely response-time based approaches to those incorporating additional process data, they provide further empirical evidence on the importance of response times for the identification of rapid disengagement. Furthermore, Article 3 suggested that response times can also be an indicator of rapid guessing in child samples.

However, response-time based methods are only able to identify *rapid* disengagement. Whereas existing validation methods (e.g., comparing average accuracy or psychometric information between responses classified as engaged or disengaged) can mainly determine whether all responses classified as disengaged indeed seem to be disengaged, they cannot specify whether all responses classified as engaged were indeed engaged (Wise, 2020; Wise & Kuhfeld, 2021). Although we know that the average accuracy for disengaged responses should be close to chance level and the psychometric information close to zero, we do not know the "true" average accuracy or psychometric information of engaged responses unless the same test was administered to a comparable sample in a high-stakes setting. We can only determine if the average accuracy and psychometric information are *higher* for engaged responses than for disengaged responses. Therefore, some of the responses classified as engaged may in reality – at least partially – be disengaged (Nagy et al., 2022; Pokropek & Khorramdel, 2025; Wise, 2020; Wise & Kuhfeld, 2021).

But how can *non-rapid* disengaged responses be identified? Some approaches identify unusually slow responses, claiming that these may represent another form of disengagement (Pokropek & Khorramdel, 2025; Yildirim-Erbasli & Bulut, 2020). However, this confines the definition of disengagement once more to outliers in item response times, assuming that (dis)engaged test takers mainly differ in the time they need to provide a response. The research on process data beyond response times, on the other hand, supposes that disengaged test-taking behavior differs in more than processing time from engaged test-taking behavior (Nagy et al., 2022; Pokropek & Khorramdel, 2025;

Wise & Kuhfeld, 2021; Wise & Kingsbury, 2022). Although Article 1 found, similar to Sahin and Colvin (2020), only small to negligible incremental effects of additional process data indicators for the identification of disengaged responding, they focused on improving the classification accuracy of *response-time based* approaches. That is, while additional indicators may indeed provide only minimal value in improving methods aimed at detecting *rapid* disengagement, they could still be able to identify other forms of disengagement that are not mainly characterized by unusually short response times. For example, data-intensive methods such as profile, sequence, or network analysis that include several process data indicators may be able to identify different forms (dis)engagement (e.g., Jerez et al., 2026; Lundgren & Eklöf, 2020; Zhang & Andersson, 2023).

However, these approaches require that process data provides enough information on the test-taking process to distinguish between engaged and (partially) disengaged responses. An important finding from Article 1 is that this information may be – at least under some circumstances – in fact limited, aligning with the finding by Ivanova and Michaelides (2023) that thresholds relying on the number of actions did not accurately identify disengagement on one third of the test items. Although the thought that additional data contains important information may be tempting, there needs to be enough variance in the data to draw inferences based on them. However, especially in traditional item formats such as multiple-choice questions, many test takers seem to respond to items without interacting with the test application, thereby creating no data besides their response times. In this case, process data can do little to inform on test-taking engagement. Even though disengaged responses may in fact be associated with less interactions, how can we differentiate between disengaged examinees that did not interact with the test and engaged examinees that did not interact with the test? A lack of interaction may in fact also be an indicator of high proficiency – when I know the answer right away, I don't need to invoke additional information, change my answer, or re-evaluate my response. These considerations highlight once more why response time may in fact be such a strong indicator of disengagement: it provides the most substantial information on the test-taking process. The argument further emphasizes the need for theoretical models that distinguish between test-taking behavior

expected from engaged and disengaged examinees, forming a theoretical foundation for research on partial or slow disengagement using process data beyond response times.

5.2.2 Conditional Performance of Identification and Scoring Methods

The findings of this dissertation align with previous studies indicating that the estimated prevalence of disengaged responding depends heavily on the chosen identification method (e.g., Goldhammer et al., 2017). The empirical studies in Articles 1 and 3 were partly based on the same data but used different identification methods. Whereas in Article 1, a mixture model identified 13.3% disengaged responses, the fixed and visual threshold method applied in Article 2 classified only around 2.8% to 3.3% of the responses to be disengaged. Although the samples overlapped only partly, it is unlikely that they differed enough to explain these stark differences.

Not only does the prevalence depend on the identification method, also the performance of the identification methods seems to depend on test and sample characteristics. For example, Article 1 was based on a mixture model for test-taking engagement that had previously been successfully applied in empirical settings (Nagy & Ulitzsch, 2022; Pokropek, 2016). However, explorative results indicated that the classification accuracy of the model may have been flawed in this particular empirical example, with proficiency and test-taking engagement being partly confounded as suggested by the perfect correlation between the two constructs. Potential reasons for this result include differing item formats, test composition, setting, and samples.

Similarly, Article 3 found that the performance of response-time thresholds may depend on a combination of sample and test. While the normative threshold method performed well in the Grade 2 reading assessment, it rendered rather short response-time thresholds in the same assessment in Grade 4. Explorative analyses revealed that this may be due to a different fit between sample and test. In Grade 2, the test's difficulty befitted the children's age, whereas in Grade 4, the test was rather easy for them. Consequently, the children responded to the items very fast, rendering the average item response times much shorter than in Grade 2, thereby potentially impeding accurate identification of rapid guesses.

Moreover, the simulation study in Article 2 suggested that the performance of scoring methods may also not be consistent but can depend on the classification accuracy of the chosen identification method, supporting findings from Rios and Deng (2024). These results highlight that there does not seem to be a "one-size-fits-all" approach to identifying or scoring disengaged responding. Each method combines different advantages and disadvantages, which may vary depending on the context. Because disengaged responding is a non-observable construct, with classifications of test-taking engagement being only estimates (see Chapter 2.2), results from empirical examples always depend on the chosen identification method. Therefore, this dissertation supports previous claims to evaluate the performance of identification methods on a study-to-study basis and to test the robustness of results by estimating them based on different identification approaches in empirical studies (Goldhammer et al., 2016; Lindner et al., 2019; Wise, 2019; Wise & Gao, 2017). Previously proposed criteria for assessing the performance of identification methods include (a) determinability across all items (i.e., can a threshold be set for all items or do the models converge; Wise, 2019), (b) visual inspection of bimodality (Lindner et al., 2019), and comparing the (c) accuracy (Goldhammer et al., 2016; Wise, 2019; Wise & Gao, 2017) or (d) score-total correlation (Wise, 2019; Wise & Gao, 2017) between responses classified as disengaged or engaged. In simulation studies, the performance of scoring methods should be evaluated based on different simulated classification accuracies (Rios & Deng, 2024).

5.2.3 Impact on Validity of Group Comparisons

Since the problem of low test-taking motivation is known to the research community, researchers have tried to determine its impact on the validity of proficiency estimates and other assessment outcomes (Rios et al., 2022; Wise, 2015; Wise & Demars, 2005, see also Chapter 2.5). To my knowledge, the simulation study in Article 2 is the first to systematically investigate the influence of rapid guessing (a) using plausible value estimates and (b) on ability group comparisons. Its results indicate that rapid guessing can pose a problem for the validity of proficiency estimates if the rapid guessing rate is high (i.e., 15%), with the impact of rapid guessing being positively associated with the proficiency of the test takers. These findings align with those of a previous simulation study (Rios et al., 2017) and can be easily explained based on the definition of the probability of a correct response in

IRT models (see Equation 2 in Chapter 2.3). The proficiency estimate of an examinee should be underestimated if (i) the probability of a correct *disengaged* response is lower than the probability of a correct *engaged* response, and (ii) they are indeed disengaged. The magnitude of the underestimation depends on the size of the difference from (i) and the extent of (ii).

While Rios et al. (2017) focused solely on the effect of rapid guessing on proficiency estimates, the simulation study in Article 2 demonstrated that the validity of group comparisons can also be impaired. In addition to the rapid guessing rate, the study identified two more factors that influence the validity of group comparisons: the impact of rapid guessing was higher when the achievement gap was large and when the more competent group engaged in rapid guessing more often than the less competent group. Both factors are a logical consequence of the association between an examinee's proficiency and the impact of disengagement on their proficiency estimates. The larger the achievement gap, the higher was the simulated proficiency of the more competent group, thereby leading to an increased bias in its proficiency estimates. Similarly, a high rapid guessing rate in the more competent group led to a stronger bias than a high rapid guessing rate in the less competent group.

Even though the simulation study demonstrated that group comparisons can *theoretically* be biased by rapid guessing, the empirical studies in Articles 2 and 3 could identify no impact of rapid guessing on group comparisons in either adults or children. These results align with findings from other empirical studies that consistently report negligible or no effect of rapid guessing on different group-based outcomes when applying response-level scoring methods (Kuhfeld & Soland, 2020; Soland, 2018; Wise, 2019; Wise et al., 2020). While some studies report a large impact of rapid guessing (Jensen et al., 2018; Setzer et al., 2013; Wise & Demars, 2010), they invariably rely on examinee-level filtering as a scoring approach. The apparently inconsistent findings illustrate an often-overlooked methodological issue when determining the impact of disengagement. In empirical studies, the estimated effect is always an effect of *accounting for* disengagement – that is, the true effect of rapid guessing is not directly determinable but can be approximated by comparing the results between methods ignoring rapid guessing and methods accounting for it (see Chapter 2.5). Therefore, the estimated effect can reflect both the true effect of rapid guessing *and* the effect of the scoring method. Because the

simulation study by Rios et al. (2017) demonstrated that examinee-level filtering can bias proficiency estimates considerably when proficiency and engagement are related, the large effects reported in studies relying on examinee-level filtering may in fact be an effect of removing disengaged – and thus less proficient – examinees from the data rather than of rapid guessing itself. Therefore, the impact of rapid guessing on group-based assessment outcomes based on low-stakes proficiency estimates could – at least in most cases – be in fact limited.

But how can this be? Given the theoretical and methodological considerations on low test-taking engagement and its consequences for the measurement of competencies (see Chapter 2.1), shouldn't there be a considerable impact on assessment outcomes? To explain this rather counterintuitive finding, I will distinguish in the following paragraphs between the impact of rapid guessing on individual proficiency estimates, on aggregated proficiency estimates, and on group-based outcomes based on proficiency estimates.

An individual's proficiency estimate is biased if the probability of a correct response under engagement differs from the probability of a correct response under disengagement (i.e., chance level), with the size of the bias depending on (a) the magnitude of this difference (which depends on the test-takers' proficiency, the item parameters, and the probability of guessing the item correctly), as well as (b) on which and how many items the examinee is disengaged (see also Wise, 2020, for a more extensive explanation). That is, the impact of disengagement on an individual's score is not only related to the amount of disengagement, but also to its interplay with the test-taker's proficiency and the item's difficulty. Thus, the same amount of disengagement will lead to higher biases in more proficient test takers and easier tests, an assumption supported by the simulation study in Article 2 and by Rios et al. (2017). Because empirical findings have repeatedly shown that disengaged responding does not occur randomly but is in fact related to both proficiency and item difficulty, with higher proficient individuals engaging in and easy items inducing less disengagement than less proficient test takers or harder items (Avvisati et al., 2024; Rios et al., 2017; Silm et al., 2019), the distortion is often smaller than expected (Wise, 2020). Therefore, the impact of disengaged responding can be rather small or – in some extreme cases – even positive (i.e., individual proficiency estimates can be *over*estimated;

Kuhfeld & Soland, 2020; Wise et al., 2020). Empirical studies focusing on the effect of low test-taking engagement on individual proficiency estimates are rare (for an example see Rios and Soland, 2021). However, studies comparing the relationship between low-stakes and high-stakes test results before and after filtering disengaged examinees consistently report higher correlations for the filtered data (Wise, 2015), indicating that individual proficiency estimates are – to some extent – distorted by low test-taking engagement.

Because educational LSAs usually aim to provide information on proficiencies at the population level rather than at the individual level, it is more informative to understand the potential impact of disengaged responding on aggregated (i.e., averaged across subpopulations) than on individual proficiency estimates. At the population level, the biases introduced by disengaged responding are averaged across examinees. Wise (2020) argues that this aggregation provides two more explanations for the rather small effects of low test-taking engagement. First, even though some test takers are disengaged in low-stakes assessments (see Chapter 2.4), most test takers are not. Second, following the arguments above, the proficiency estimates of disengaged examinees are not necessarily distorted, because test takers tend to be disengaged when their chances of responding to an item correctly are anyways low. Therefore, the majority of proficiency estimates should not be biased, and the ones that are biased may not be severely or in some cases even positively distorted, thereby reducing the *average* distorting effects of disengagement. Nevertheless, most empirical and simulation studies report a small to moderate bias of disengagement on aggregated proficiency scores (Rios et al., 2017, 2022; Welling et al., 2026b; Wise & Demars, 2005).

While these considerations explain why the impact of low test-taking engagement on proficiency estimates may be lower than expected, they do not fully explain why the impact on group-based outcomes such as ability comparisons or school rankings seems to be negligible (Kuhfeld & Soland, 2020; Soland, 2018; Welling et al., 2026a, 2026b; Wise et al., 2020). As already discussed in the introduction (see Chapter 2.5), *average* score distortion is not informative for determining the impact of disengagement on outcomes in LSAs. Instead, outcomes at the population level are only distorted if *differential* score distortion is present – that is, if the distortion depends on the variables included in

the analyses (e.g., subgroups for comparisons or predictors in regressions on proficiency). This argument goes beyond the usual assumption that differential disengagement (i.e., one group employs more rapid guessing than the other) should suffice for biased group-based outcomes (Rios, 2021; Wise et al., 2020). Based on the findings of this dissertation and of previous research, the following paragraphs shall disentangle determining factors of differential score distortion, thereby explaining recent null effects of low test-taking engagement on group-based outcomes.

Group comparisons are only biased if the group scores are affected to different extents, because if the scores were distorted to equal extents, the estimated ability difference would remain identical. If both groups are equally proficient, differential score distortion depends on (i) differential test-taking engagement and (ii) the difficulty of the test. Group-based outcomes will be, in this case, significantly distorted when the difference in test-taking engagement is sufficiently large based on the test's difficulty. Analogous to the simulation study by Rios et al. (2017), who investigated the effect of rapid guessing on average proficiency estimates, this should be a difference of around 6 percentage points for easy tests, and 12 percentage points for moderate or difficult tests. That is, one group needs to exhibit 6 to 12 percentage points more disengagement than the other group for group comparisons to be biased. Because average disengagement rates range around 7% (Rios et al., 2022), this is rather unlikely to occur.

If the groups differ in their proficiency levels, the magnitude of the differential score distortion depends on the combination of the group's test-taking engagement, its average proficiency, and the test's difficulty. Because score distortion is positively associated with proficiency (as demonstrated in Rios et al., 2017, and in Article 2), the average score of the more proficient group will be more distorted than that of the less proficient group even at an equal disengagement rate. Therefore, comparisons between groups with different levels of ability should be biased when the more proficient group exhibits at least as much disengagement as the less proficient group, with the distortion being positively related to (i) the test's easiness, (ii) the overall disengagement rate, (iii) the differential disengagement in favor of the more proficient group, and (iv) the ability difference between the groups. The simulation study included in Article 2 demonstrated that when proficiency and test-taking

engagement are strongly related at the individual level, the disengagement rate needs to be as high as 15% *and* the more competent group needs to rapidly guess more often or be far more proficient than the less competent group in order to attain noteworthy differential score distortion. However, empirical findings indicate that rapid guessing rates are usually much lower (Rios et al., 2022) and that proficiency is negatively associated with engagement (Nagy & Ulitzsch, 2022; Rios et al., 2017; Silm et al., 2019, 2020; Ulitzsch et al., 2020), meaning that the more proficient group usually employs less rapid guessing than the other group (e.g., Welling et al., 2026a, 2026b). Whereas the strength of the relationship between test-taking engagement and proficiency at the individual level should influence the impact of disengagement on average scores, it should not substantially affect differential score distortion (because the correlation is equal for both groups).

These results explain why outcomes based on ability comparisons between different groups (e.g., achievement gaps, country or school rankings) or time points (e.g., competence development) may not be severely affected by rapid guessing. However, this does not necessarily signify that low test-taking engagement does not have any consequences for the validity of assessment outcomes. As discussed earlier in this chapter, all empirical findings on the impact of disengaged responding are based on comparisons between results obtained by ignoring disengagement vs. accounting for it. Moreover, all aforementioned studies focused on rapid guessing as an easily detectable but extreme form of disengaged responding. Because low test-taking engagement may not always be accompanied by short response times (Nagy et al., 2022; Pokropek & Khorramdel, 2025; Wise & Kuhfeld, 2021) and because most identification methods are rather conservative (i.e., focus on preventing wrongly classifying engaged responses as disengaged), many disengaged responses could go unnoticed (Kuhfeld & Soland, 2020; Wise & Kuhfeld, 2021). Therefore, the existing empirical findings mostly demonstrate that accounting for extremely fast responses does not have noteworthy effects on group-based assessment outcomes. A meta-analysis synthesizing (quasi-)experimental studies estimated the average difference in test performance between highly motivated groups and less motivated groups to be approximately 0.50 SD (Wise & Demars, 2005). This effect size can be considered as a rough indicator of the overall effect of low-test taking engagement on aggregated scores. Even though the impact on

actual assessment outcomes should be smaller (see argument above), the overall effect of disengagement may indeed be larger than previous studies assume.

Furthermore, low test-taking engagement may impact other aspects of measurement such as item parameters (Alahmadi & DeMars, 2025), model fit (Rios & Guo, 2020), metric measurement invariance (Rios, 2021), and score equating (Deng & Rios, 2022; Rios & Guo, 2020). No or negligible effects were shown for differential item functioning (Abulela & Rios, 2022), reliability (Rios et al., 2022), or other forms of measurement invariance (Rios, 2021). The inconclusive results highlight the need for more studies systematically investigating the consequences of low test-taking engagement for the validity of assessment outcomes. A combination of simulation and empirical studies could ensure that method effects (i.e., removing low proficient individuals from the sample) are not mistakenly interpreted as an effect of disengagement.

In summary, the combined results of the three articles included in this dissertation go beyond the key findings of each article, highlighting important consequences for response-time–based measurement of disengaged responding, the context-dependence of identification and scoring approaches, and the validity of group-based assessment outcomes in LSAs.

5.3 Implications for Large-Scale Assessments

This dissertation entails several important implications for educational LSAs concerning the identification of disengaged responding, the measurement of competencies, the impact of disengaged responding on the validity of assessment outcomes, and the generalizability of established findings to child samples.

The theoretical considerations and empirical findings in this dissertation support recent research indicating that the validity of group-based outcomes may be less affected by rapid disengagement than previously thought. Especially in samples with low disengagement rates (e.g., among children), the bias introduced by disengaged responding may be negligible to non-existent. This would be great news for LSAs, because it would mean that most of their findings are trustworthy even in the presence of rapid guessing. However, researchers and practitioners should still be cautious due to several reasons. First, the simulation study has demonstrated that under certain circumstances –

even though they may be rare in practice – rapid guessing indeed affects group comparisons. Second, even though group-based outcomes may not be strongly impaired, other factors such as the score equating may well be. Third, a potential non-impact of rapid guessing on group-based outcomes does not necessarily mean that disengaged responding as a whole has no effect. Disengagement not characterized by short response times may affect the validity of assessment outcomes, although its impact is difficult to determine since its measurement poses great challenges to researchers.

Because the impact of low test-taking engagement on the validity of assessment outcomes can only be reliably determined if we know which responses were engaged and which not, and because validity may still be affected by low test-taking engagement, it is essential to accurately identify it. This dissertation supports recent findings that even in the presence of additional information on the test-taking process, response times remain a strong indicator of low test-taking engagement and successfully classify fast responses that are characterized by low average accuracy as disengaged (Patel et al., 2021; Sahin & Colvin, 2020). Therefore, response-time based methods seem to be reliable approaches to identifying rapid disengagement. However, it is more difficult to determine if all disengaged responses are indeed rapid. While this dissertation suggests that additional process data indicators may not improve the classification of *rapid* disengagement substantially, they may still be able to identify different forms of disengagement.

While it is generally known that the performance of threshold-based approaches may vary depending on factors such as sample, test design, and test domain (Goldhammer et al., 2017; Kroehne et al., 2020; Wise, 2019), this dissertation suggests that the performance of model-based approaches may also be context-dependent. Researchers and practitioners should therefore – analogous to threshold-based approaches – evaluate the validity of model-based approaches on a study-to-study basis, using validation criteria such as comparing the proportion of correct responses or the score-total correlation between responses classified as solution behavior or disengaged responding (Goldhammer et al., 2016; Wise, 2019; Wise & Gao, 2017). Furthermore, results should be compared between different identification methods to ensure the robustness of results (Goldhammer et al., 2016; Kroehne et al., 2020).

Once disengaged responses have been identified, they need to be accounted for in the measurement of competencies. This dissertation demonstrates that also in plausible value estimation, researchers and practitioners should account for rapid guessing in the measurement model. If assessment outcomes are impaired by low test-taking engagement, accounting for it in the population model does not suffice to reduce the biases. Because educational LSAs often provide plausible value estimates in published data, this finding is an important step towards improving the data quality of proficiency estimates.

Lastly, this dissertation argues that empirical results on low test-taking engagement may in fact be age-dependent, thereby encouraging researchers and practitioners to consider the participants' age instead of generalizing the results across age groups when reporting or synthesizing empirical findings. However, not only the age of the examinees affects the occurrence and consequences of disengaged responding. Factors such as test design, assessment setting, and the fit between the test and the sample may also influence test-taking engagement. Future studies should disentangle these different effects to further our understanding of the contextual aspects of disengaged responding. Knowing which factors induce low test-taking engagement can help improve the design of competence assessments in LSAs.

5.4 Strengths, Limitations, and Outlook

This dissertation integrates theory with findings from simulation and empirical studies to advance the understanding of disengaged responding in LSAs. By outlining theoretical considerations both on test-taking engagement among children and on the potential non-impact of rapid guessing on group-based assessment outcomes, it strengthens the conceptual foundation for future research. It further advances methodological approaches to identifying and modeling disengaged responding by forming new indicators from raw process data and by conducting the first simulation study on rapid guessing in the context of plausible value estimation. Moreover, it provides first or further empirical evidence on the impact of rapid guessing on the validity of group comparisons as well as on the antecedents and consequences of rapid guessing among children using large samples from different cohorts in a German LSA. Lastly, it ensures the validity and robustness of the findings by not only

applying different approaches to identifying and accounting for disengaged responding, but also by evaluating the validity of these methods on a study-to-study basis and by comparing the results between different methods.

Despite its considerable strengths, this dissertation possesses some general limitations. First, all articles included in this work defined test-taking engagement as (a) mainly identifiable by response times, and (b) dichotomous in a sense that responses are classified as either engaged or disengaged. While Article 1 softened these assumptions by allowing additional indicators to improve classification and by using a probabilistic estimation approach, the model still postulates that test-taking engagement influences response times and can be distinguished in two latent classes: solution behavior and disengaged responding. However, not all disengaged responses may be fast and test-taking engagement may not always be classifiable as either fully disengaged or engaged. In recent years, research has emerged that investigates different forms of (dis)engagement. For example, some studies propose using measures of performance decline as indicators of partial test-taking engagement (Wise & Kingsbury, 2022; Wise & Kuhfeld, 2021), since the decline cannot be explained by rapid guessing alone (Nagy et al., 2022). Pokropek and Khorramdel (2025) have demonstrated that the relationship between response times and disengagement may be in fact quadratic, with both fast and slow responses indicating disengaged responding. Furthermore, data-driven methods such as cluster, sequence, profile, or network analysis have identified different forms of test-taking behavior based on indicators stemming from self-report, response times, and additional process data (Hofverberg et al., 2022; Jerez et al., 2026; Lundgren & Eklöf, 2020; Zhang & Andersson, 2023). Building on this new evidence, future studies should try to replicate findings from this dissertation based on broader definitions of disengaged responding to evaluate their generalizability to test-taking engagement in general. Furthermore, the present findings emphasize the need for theoretical models that distinguish between test-taking behavior expected from engaged and disengaged examinees, which could form a theoretical foundation for research on partial or slow disengagement using process data beyond response times.

Second, to keep the scope of the articles manageable, they each focused on one or two approaches to identifying disengaged responding and to accounting for it in the measurement of competencies. While the approaches were carefully selected based on their level of validation and prevalence in the disengagement literature, other methods may still yield slightly different results. Furthermore, this dissertation suggests that both test-taking engagement and the performance of identification methods can be context-dependent. To ensure the robustness of the findings, future studies should therefore replicate empirical studies included in this dissertation in other samples and contexts and by using different methods of measuring disengaged responding. Similarly, future simulation studies could evaluate the performance of additional scoring methods in the context of plausible value estimation and compare them to the EMM applied in Article 2. Such simulation studies should account for different (a) disengagement rates, (b) proficiency levels, (c) classification accuracies of the identification methods, and (d) in the case of group-based estimands also for the interplay between test-taking engagement and proficiency levels. To further improve our understanding of the effect of (d), future simulation studies should also systematically vary the correlation between disengagement and proficiency.

Lastly, all articles included in this dissertation were based on IRT models that include item difficulties and thresholds as the sole item parameters. The empirical studies were based on data from the *National Educational Panel Study* (Blossfeld & Roßbach, 2019), in which competence assessments are designed according to the theoretical framework of the partial credit (Masters, 1982) and Rasch (1960) models. However, some researchers and practitioners rely on IRT models that additionally include item discrimination parameters (e.g., 2PL model, Birnbaum, 1968). While most results should not be heavily affected, future studies could evaluate if the choice of IRT model has an impact on findings on the interplay between low test-taking engagement and the measurement of competencies.

Even beyond its limitations, this dissertation offers possibilities for future research. To further improve the understanding of the consequences of low test-taking engagement on the validity of assessment outcomes, future studies should synthesize past research using systematic reviews or meta-analyses. Moreover, additional simulation studies are necessary for explaining the circumstances

under which different assessment outcomes are affected by disengaged responding (i.e., rapid guessing or overall disengagement). This dissertation argues that these simulations should not only evaluate the effect of varying disengagement rates on score distortion but also of the interplay between proficiency and disengagement both on the person and on the group level. Lastly, new empirical studies could test the findings from the simulation studies in different samples and contexts, thereby allowing an inclusive perspective on the effects of test-taking engagement on the validity of assessment outcomes.

Furthermore, future research should include experimental studies that disentangle setting, test design, and age effects on disengaged responding. While Article 3 provided evidence on rapid guessing in child samples, results may also be influenced by a combination of setting, test design, and their interplay with identification methods. Understanding which factors induce disengaged responding can help to design assessments that minimize low test-taking engagement from the start – a far better alternative to post-hoc identification and scoring. While some research on setting effects exists (e.g., Kroehne et al., 2020), it includes only a limited number of settings and does not take their potential interaction with the age of the participants into account. Moreover, not only setting and design effects may differ between different age groups, but also the performance of prevalent identification methods. Because children seem to exhibit lower disengagement rates than older individuals, approaches that require a substantial amount of disengagement may not converge (e.g., mixture models) or render setting a threshold impossible (e.g., visual inspection methods) in child samples. Furthermore, children may express low test-taking engagement differently than older test takers (e.g., by exhibiting slow or partial disengagement rather than rapid disengagement). Future studies should therefore evaluate existing methods of identifying disengaged responding explicitly in different age groups.

5.5 Conclusion

Combining theoretical, methodological, and empirical approaches, this dissertation advances the literature on disengaged responding in the context of its measurement, scoring, impact, and the generalizability of established findings to different age groups. It contributes to an improved identification of disengaged responding by demonstrating how raw process data can be leveraged to form new potential indicators of test-taking engagement based on theoretical considerations. Although

the additional indicators barely improved classification accuracy, the findings emphasize the importance of response times as a predictor of rapid disengagement and challenge the assumption that process data generally provides sufficient information on test-taking engagement to inform new or improve existing identification methods (Article 1). Additionally, this dissertation contributes to a more accurate measurement of competencies by proposing and evaluating different approaches to accounting for rapid guessing in plausible value estimation. The results indicate that incorporating test-taking engagement in the population model does not suffice to reduce potential biases, highlighting the need for improving existing estimation procedures in LSAs (Article 2). Findings from both simulation and empirical studies further suggest that the impact of rapid guessing on the validity of group-based outcomes may be smaller than often assumed (Articles 2 and 3). Furthermore, this dissertation claims that test-taking behavior may be age-dependent, outlining theoretical considerations and (further) empirical evidence on the antecedents, prevalence, and impact of rapid guessing among elementary school students (Article 3).

The general discussion builds on previous research and the findings included in this dissertation to discuss the strengths and limitations of response-time based measures of disengagement, to show that the performance of identification and scoring methods may be context-dependent, and to provide potential explanations for the negligible impact of disengagement on the validity of group-based assessment outcomes. Potential avenues for future research include investigating disengagement beyond rapid guessing, developing theoretical models for the usage of process data in the context of test-taking engagement, systematically evaluating the consequences of rapid and non-rapid disengagement on the validity of assessment outcomes, and conducting experimental studies to disentangle setting, test design, and age effects of disengaged responding.

Taken together, this dissertation addressed three important yet unresolved issues in the context of low test-taking engagement, thereby providing a more differentiated understanding of the identification, modeling, and consequences of disengaged responding. By laying conceptual and methodological groundwork, this dissertation provides a foundation for future research aimed at

developing more robust, valid, and context-sensitive approaches to measuring competencies in educational LSAs.

References

Abulela, M. A. A., & Rios, J. A. (2022). Comparing the robustness of three nonparametric DIF procedures to differential rapid guessing. *Applied Measurement in Education*, *35*(2), 81–94. https://doi.org/10.1080/08957347.2022.2067542

Alahmadi, S., & DeMars, C. E. (2025). From item estimates to test operations: The cascading effect of rapid guessing. *Journal of Educational Measurement*, *62*(4), 740–762. https://doi.org/10.1111/jedm.70010

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Standards for educational and psychological testing*. American Educational Research Association. https://doi.org/10.4135/9781506326139.n662

Avvisati, F., Buchholz, J., Piacentini, M., & Vargas Madriz, L. F. (2024). *Item characteristics and test-taker disengagement in PISA* (OECD Education Working Papers No. 312). OECD Publishing. https://doi.org/10.1787/7abea67b-en

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. https://doi.org/10.1007/BF03173192

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-23162-0

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Buchholz, J., Cignetti, M., & Piacentini, M. (2022). *Developing measures of engagement in PISA* (OECD Education Working Papers No. 279). OECD Publishing. https://doi.org/10.1787/2d9a73ca-en

Coniam, D., Lee, T., & Lampropoulou, L. (2021). Identifying guessing in English language tests via Rasch fit statistics: An exploratory study. *English Language Teaching*, *14*(5), 23–29. https://doi.org/10.5539/elt.v14n5p23

Cronbach, L. J. (1970). *Essentials of Psychological Testing* (3rd ed.). Harper & Row.

Deng, J., & Rios, J. A. (2022). Investigating the effect of differential rapid guessing on population invariance in equating. *Applied Psychological Measurement*, *46*(7), 589–604. https://doi.org/10.1177/01466216221108991

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers No. 133). OECD Publishing. https://doi.org/10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education*, *5*, 18.

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

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. https://doi.org/10.1080/08957347.2016.1171766

Guo, H., Rios, J. A., Ling, G., Wang, Z., Gu, L., Yang, Z., & Liu, L. O. (2022). Influence of selected-response format variants on test characteristics and test-taking effort: An empirical study. *ETS Research Report Series*. https://doi.org/10.1002/ets2.12345

Hofverberg, A., Eklöf, H., & Lindfors, M. (2022). Who makes an effort? A person-centered examination of motivation and beliefs as predictors of students' effort and performance on the PISA 2015 science assessment. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.791599

Ivanova, M., & Michaelides, M. (2023). Measuring test-taking effort on constructed-response items with item response time and number of actions. *Practical Assessment, Research, and Evaluation*, *28*(15). https://doi.org/10.7275/PARE.1921

Ivanova, M., Michaelides, M., & Eklöf, H. (2020). How does the number of actions on constructed-response items relate to test-taking effort and performance? *Educational Research and Evaluation*, *26*(5–6), 252–274. https://doi.org/10.1080/13803611.2021.1963939

Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *40*(2), 267–284. https://doi.org/10.3102/0162373718759600

Jerez, D., Mazzullo, E., & Bulut, O. (2026). Exploring slow responses in international large-scale assessments using sequential process analysis. *Computers*, *15*. https://doi.org/10.3390/computers15010064

Knekta, E., & Eklöf, H. (2015). Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value-based questionnaire. *Journal of Psychoeducational Assessment*, *33*(7), 662–673. https://doi.org/10.1177/0734282914551956

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, *62*(2), 147–177. https://doi.org/10.25656/01:23630

Kuang, H., & Sahin, F. (2023). Comparison of disengagement levels and the impact of disengagement on item parameters between PISA 2015 and PISA 2018 in the United States. *Large-Scale Assessments in Education*, *11*, 4. https://doi.org/10.1186/s40536-023-00152-0

Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness*, *13*(1), 147–175. https://doi.org/10.1080/19345747.2019.1636437

LaFave, A. J., Taylor, J. A., Barter, A. M., & Jacobs, A. S. (2022). Student engagement on the National Assessment of Educational Progress (NAEP): A systematic review and meta-analysis of extant research. *Educational Assessment*, *27*(3), 205–228. https://doi.org/10.1080/10627197.2022.2043151

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*, 8. https://doi.org/10.1186/s40536-014-0008-1

Leventhal, B. C., & Pastor, D. (2023). An illustration of an IRTree model for disengagement. *Educational and Psychological Measurement*, *84*(4), 810–834. https://doi.org/10.1177/00131644231185533

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01533

Lombardi, E., Di Dio, C., Castelli, I., Massaro, D., & Marchetti, A. (2017). Prospective thinking and decision making in primary school age children. *Heliyon*, *3*(6). https://doi.org/10.1016/j.heliyon.2017.e00323

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation*, *26*(5–6), 275–301. https://doi.org/10.1080/13803611.2021.1963940

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

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

Michaelides, M., & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: Country and item-type differences. *Psychological Test and Assessment Modeling*, *64*(3), 304–338.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457

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. https://doi.org/10.1111/bmsp.12117

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*, *82*(5), 845–879. https://doi.org/10.1177/00131644211045351

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*, *39*(3), 751–766. https://doi.org/10.1111/jcal.12719

Patel, N., Sharma, A., Shah, T., & Lomas, D. (2021). *Modeling NAEP test-taking behavior using educational process analysis*. *13*(2), 16–54. https://doi.org/10.5281/zenodo.5275312

Piaget, J. (2022). *Meine Theorie der geistigen Entwicklung [My theory of cognitive development]* (R. Fatke, Ed.; H. Kober, Trans.; 4th edition). Julius Beltz GmbH & Co. KG. (Original work published 1970)

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

Pokropek, A., & Khorramdel, L. (2025). Analyzing test-taking behaviors through process data using the IRT explanatory model for guessing. In L. Khorramdel, M. Von Davier, & K. Yamamoto (Eds.), *Innovative Digital-Based International Large-Scale Assessments* (pp. 559–584). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-90951-1_23

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rios, J. A. (2021). Is differential noneffortful responding associated with type I error in measurement invariance testing? *Educational and Psychological Measurement*, *81*(5), 957–979. https://doi.org/10.1177/0013164421990429

Rios, J. A., & Deng, J. (2024). A comparison of response time threshold scoring procedures in mitigating bias from rapid guessing behavior. *Educational and Psychological Measurement*, *84*(2), 387–420. https://doi.org/10.1177/00131644231168398

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, *27*(4), 356–373. https://doi.org/10.1080/10627197.2022.2110465

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, *33*(4), 263–279. https://doi.org/10.1080/08957347.2020.1789141

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. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2021). Investigating the impact of noneffortful responses on individual-level scores: Can the effort-moderated IRT model serve as a solution? *Applied Psychological Measurement*, *45*(6), 391–406. https://doi.org/10.1177/01466216211013896

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, *22*(2), 154–184. https://doi.org/10.1080/15305058.2022.2036161

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (1st ed.). Wiley. https://doi.org/10.1002/9780470316696

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education*, *8*, 5. https://doi.org/10.1186/s40536-020-00082-1

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Silm, G., Must, O., & Täht, K. (2019). Predicting performance in a low-stakes test using self-reported and time-based measures of effort. *Trames: Journal of the Humanities and Social Sciences*, *23*(3), 353–376. https://doi.org/10.3176/tr.2019.3.06

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*. https://doi.org/10.1016/j.edurev.2020.100335

Slobodskaya, H. R., & Kornienko, O. S. (2021). Age and gender differences in personality traits from early childhood through adolescence. *Journal of Personality*, *89*(5), 933–950. https://doi.org/10.1111/jopy.12624

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment*, *24*(4), 327–342. https://doi.org/10.1080/10627197.2019.1645592

Svetina Valdivia, D., Rutkowski, L., Rutkowski, D., Canbolat, Y., & Underhill, S. (2023). Test engagement and rapid guessing: Evidence from a large-scale state assessment. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1127644

Ulitzsch, E., Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *The British Journal of Mathematical and Statistical Psychology*, *73*(1), 83–112. https://doi.org/10.1111/bmsp.12188

van Loon, M., De Bruin, A., Leppink, J., & Roebers, C. (2017). Why are children overconfident? Developmental differences in the implementation of accessibility cues when judging concept learning. *Journal of Experimental Child Psychology*, *158*, 77–94. https://doi.org/10.1016/j.jecp.2017.01.008

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *The British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Welling, J., Fischer, L., Thums, K., & Gnambs, T. (2026). *Prevalence, predictors, and consequences of rapid guessing among children in elementary school* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/tsvyf_v1

Welling, J., Gnambs, T., & Carstensen, C. H. (2024). Identifying disengaged responding in multiple-choice items: Extending a latent class item response model with novel process data indicators. *Educational and Psychological Measurement*, *84*(2), 314–339. https://doi.org/10.1177/00131644231169211

Welling, J., Zink, E., & Gnambs, T. (2026). Comparing different approaches of (not) accounting for rapid guessing in plausible values estimation. *Educational and Psychological Measurement*, Advance Online Publication. https://doi.org/10.1177/00131644251395590

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252. https://doi.org/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. https://doi.org/10.1111/emip.12165

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, *32*(4), 325–336. https://doi.org/10.1080/08957347.2019.1660350

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5–6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

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

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

Wise, S. L., & Demars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment*, *26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86–105. https://doi.org/10.1111/jedm.12102

Wise, S. L., & Kingsbury, G. G. (2022). Performance decline as an indicator of generalized test-taking disengagement. *Applied Measurement in Education*, *35*(4), 272–286. https://doi.org/10.1080/08957347.2022.2155651

Wise, S. L., Kingsbury, G. G., Kong, X. J., & Thomason, J. (2004, April 15). *An investigation of motivation filtering in a statewide achievement testing program*. National Council on Measurement in Education 67th Annual Meeting, San Diego, CA, United States.

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. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. (2021). A method for identifying partial test-taking engagement. *Applied Measurement in Education*, *34*(2), 150–161. https://doi.org/10.1080/08957347.2021.1890745

Wise, S. L., & Ma, L. (2012, April 13). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper Presentation]. 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. https://doi.org/10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 139–153). American Psychological Association. https://doi.org/10.1037/12330-009

Wise, S. L., Soland, J., & Bo, Y. (2020). The (non)impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57–77. https://doi.org/10.1080/15305058.2019.1605999

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Xia, M., Poorthuis, A. M. G., & Thomaes, S. (2024). Children's overestimation of performance across age, task, and historical time: A meta-analysis. *Child Development*, *95*(3), 1001–1022. https://doi.org/10.1111/cdev.14042

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. *ETS Research Report Series*, *1995*(1). https://doi.org/10.1002/j.2333-8504.1995.tb01637.x

Yildirim-Erbasli, S. N., & Bulut, O. (2020). The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. *Educational Research and Evaluation*, *26*(7–8), 368–386. https://doi.org/10.1080/13803611.2021.1977152

Zhang, M., & Andersson, B. (2023). Identifying problem-solving solution patterns using network analysis of operation sequences and response times. *Educational Assessment*, *28*(3), 172–189. https://doi.org/10.1080/10627197.2023.2222585

Author Contributions

Article 1

Welling, J., Gnambs, T., & Carstensen, C. H. (2024). Identifying disengaged responding in multiple-choice items: Extending a latent class item response model with novel process data indicators. *Educational and Psychological Measurement*, *84*(2), 314–339. https://doi.org/10.1177/00131644231169211

CRediT Role	**Author**		
	Jana Welling	Timo Gnambs	Claus Carstensen
Conceptualization	X	X	X
Data Curation	X		
Formal Analysis	X		
Methodology	X	X	X
Supervision		X	X
Visualization	X		
Writing – original draft	X		
Writing – review & editing	X	X	

Article 2

Welling, J. [1], Zink, E. [1], & Gnambs, T. (2026). Comparing different approaches of (not) accounting for rapid guessing in plausible values estimation. *Educational and Psychological Measurement*, Advance Online Publication. https://doi.org/10.1177/00131644251395590

[1] These authors contributed equally

CRediT Role	**Author**		
	Jana Welling	Eva Zink	Timo Gnambs
Conceptualization	X	X	X
Data Curation	X	X	
Formal Analysis	X	X	
Methodology	X	X	
Supervision			X
Visualization	X		
Writing – original draft	X	X	
Writing – review & editing	X	X	X

Article 3

Welling, J., Fischer, L., Thums, K., & Gnambs, T. (2026). *Prevalence, predictors, and consequences of rapid guessing among children in elementary school* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/tsvyf_v1

CRediT Role	**Author**			
	Jana Welling	Luis Fischer	Kathrin Thums	Timo Gnambs
Conceptualization	X	X	X	
Data Curation	X			
Formal Analysis	X			X
Investigation	X	X		
Methodology	X	X		X
Supervision				X
Visualization	X			
Writing – original draft	X			
Writing – review & editing	X	X		X

Appendix

Appendix A

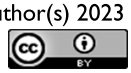
Welling, J., Gnambs, T., & Carstensen, C. H. (2024). Identifying disengaged responding in multiple-choice items: Extending a latent class item response model with novel process data indicators. *Educational and Psychological Measurement*, *84*(2), 314–339. https://doi.org/10.1177/00131644231169211

Supplemental Material: https://osf.io/tuwmq

Check for updates

Original Research Article

Identifying Disengaged Responding in Multiple-Choice Items: Extending a Latent Class Item Response Model With Novel Process Data Indicators

Educational and Psychological Measurement
2024, Vol. 84(2) 314–339
© The Author(s) 2023

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644231169211
journals.sagepub.com/home/epm


Jana Welling[1], Timo Gnambs[1] and Claus H. Carstensen[2]

Abstract

Disengaged responding poses a severe threat to the validity of educational large-scale assessments, because item responses from unmotivated test-takers do not reflect their actual ability. Existing identification approaches rely primarily on item response times, which bears the risk of misclassifying fast engaged or slow disengaged responses. Process data with its rich pool of additional information on the test-taking process could thus be used to improve existing identification approaches. In this study, three process data variables—text reread, item revisit, and answer change—were introduced as potential indicators of response engagement for multiple-choice items in a reading comprehension test. An extended latent class item response model for disengaged responding was developed by including the three new indicators as additional predictors of response engagement. In a sample of 1,932 German university students, the extended model indicated a better model fit than the baseline model, which included item response time as only indicator of response engagement. In the extended model, both item response time and text reread were significant predictors of response engagement. However, graphical analyses revealed no systematic differences in the item and person parameter estimation or item response classification between the models. These results suggest only a marginal improvement of the identification of disengaged responding by the new indicators.

[1]Leibniz Institute for Educational Trajectories, Bamberg, Germany
[2]Otto-Friedrich University Bamberg, Germany

Corresponding Author:
Jana Welling, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, Bamberg 96047, Germany.
Email: jana.welling@lifbi.de

Implications of these results for future research on disengaged responding with process data are discussed.

Keywords

disengaged responding, rapid guessing, process data, item response theory, computer-based assessments, multiple-choice items

Educational large-scale assessments such as the *Program for International Student Assessment* (PISA; Schleicher, 2019) or the German *National Educational Panel Study* (NEPS; Blossfeld & Roßbach, 2019) administer cognitive tests to provide high-quality data on various domain-specific competences and their development over time. Still, a potential weakness of these studies is that they are low-stakes assessments in which the test results do not have direct consequences for the test-takers. Test-takers might thus lack the motivation to engage properly in the test and, consequently, respond to the items without investing their best effort, knowledge, and ability, a behavior defined as disengaged responding (Wise & DeMars, 2005). As disengaged responses do not reflect the ability of the test-takers, they pose a severe threat to the validity of the assessment (Wise, 2017).

Although various methods exist to identify disengaged responding, these typically rely either on item responses, item response times, or a combination of both (e.g., Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015). However, this practice bears the risk of misclassifications if fast engaged responses are mistaken for disengagement or slow disengaged responses are erroneously classified as engaged responses. Such misclassifications can in turn introduce bias into the estimation of item and person parameters. Therefore, additional indicators of response engagement might improve existing approaches by classifying item responses more accurately as engaged or disengaged. One potential source of information on test-taking behavior that has been increasingly available with the advent of computer-based assessments is process data, which provides a rich array of easily accessible information on the test-taking process. Few studies have yet investigated the potential of process data beyond item response times for the identification of disengaged responding (Ivanova et al., 2020; Lundgren & Eklöf, 2020; Patel et al., 2021; Sahin & Colvin, 2020). So far, available studies either concentrated on response engagement in interactive items or were rather exploratory. However, most competence tests in educational large-scale assessments are still based on traditional item formats such as multiple-choice questions. Furthermore, to evaluate the actual benefit of new indicators, it is essential to compare novel identification methods with already existing ones.

The aims of this study are thus (a) to derive novel indicators of response engagement from process data in a reading comprehension test, (b) to incorporate these indicators into an existing latent class item response model based solely on item response times, and (c) to evaluate the incremental value of the new indicators. After

providing a brief overview of existing methods for the identification of disengaged responding, we will introduce the dependent latent class item response model (DLC-IRT) originally proposed by Pokropek (2016) and advanced by Nagy and Ulitzsch (2022). Subsequently, we will present the novel process data indicators of response engagement and derive an extended version of the DLC-IRT model. In the end, we will evaluate the new indicators by implementing both the original and the extended model in a sample of German university students as part of the NEPS.

Types of Test-Taking Behavior

Schnipke and Scrams (1997) assumed that test-taking behavior can be distinguished into solution behavior and rapid guessing behavior. While solution behavior describes a state of engagement in which test-takers invest their best effort, knowledge, and ability into the response of an item, rapid guessing describes a state of disengagement in which test-takers respond to an item fast and without regard for its actual content. Although most previous research focused on rapid guessing, it is not the only possible form of disengaged responding. Slow disengaged responses, perfunctory answers to open format items, and item omissions can also be a consequence of disengagement in the test-taking process (Wise et al., 2020).

Numerous studies demonstrated that disengaged responding is a serious concern in educational large-scale assessments, with estimated prevalence rates up to 28%, depending on various aspects such as the content domain, setting, assessment mode, test-taker characteristics, country, or detection method (Goldhammer et al., 2017; Kroehne et al., 2020; Nagy & Ulitzsch, 2022; Rios et al., 2022). Disengaged responding manifests itself on the item-by-person level with engagement rates varying both between examinees and items (Ulitzsch et al., 2020). Consequently, disengaged responding is related to both person (e.g., ability, Wise et al., 2020) and item characteristics (e.g., item position, Lee & Jia, 2014). However, even though disengaged responding might be related to the ability of the test-taker or the item difficulty, disengaged responses do not reflect these properties in the same way as responses stemming from solution behavior. As a consequence, simply ignoring disengaged responding can lead to biased item and person parameter estimates (Jin et al., 2022; Rios et al., 2017). Furthermore, it can also influence conclusions and decisions on other aspects such as differential item functioning (DeMars & Wise, 2010) or the speed-ability relation (Deribo et al., 2022). These findings emphasize the necessity to identify and account for disengaged responding when using data from large-scale assessments to address substantial research questions.

Identification of Disengaged Responding

Previous research has mainly focused on the identification of rapid guessing, a form of disengaged responding characterized by unrealistic short item response times (e.g., Wise, 2017). Identification methods can be generally divided into three

different kinds of approaches: misfit-statistic-based, threshold-based, and model-based approaches. Misfit-statistic-based methods detect aberrant behavior by analyzing person-misfit statistics to identify extreme response patterns on the person level (e.g., Marianti et al., 2014). However, these methods fall short of identifying disengaged responding also on the item level. In contrast, threshold-based methods classify each item response as either engaged or disengaged (e.g., Goldhammer et al., 2017). If the respective item response time is shorter than a previously specified threshold, it is classified as disengaged. Disengaged responses are consequently filtered from the data set (e.g., Lee & Jia, 2014; Wise & DeMars, 2006). However, this approach requires setting either a common threshold for all items, which does not reflect item-specific time demands on the test-takers, or separate thresholds for each item, which may result in a cumbersome procedure especially for long tests. Since filtering responses classified as disengaged is based on the assumption that disengaged responding and ability are not related, ability estimates might be further biased if this assumption is violated.

Model-based methods simultaneously estimate ability and response engagement in a joint model (Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015). These mixture item response theory (IRT) models define response engagement as a latent class variable, assuming a mixed distribution for the probability of a correct response. In the solution behavior class, the probability follows a customary IRT model, whereas in the disengaged responding class, the probability of a correct response is independent of the test-taker's ability and item difficulty. Model-based approaches differ in the way they incorporate item response times into the model (Nagy & Ulitzsch, 2022). Independent latent class IRT models assume a mixed distribution for item response time dependent on latent class membership and thus treat response engagement as an independent variable determining item response times (Ulitzsch et al., 2020; Wang & Xu, 2015). In contrast, DLC-IRT models incorporate item response times as a predictor of latent class membership into the model and thus treat response engagement as a dependent variable (Pokropek, 2016). Nagy and Ulitzsch (2022) provided a common multilevel framework for both approaches and illustrated with empirical data that both model types can indeed improve traditional IRT models ignoring disengaged responding.

A great advantage of the DLC-IRT model is the flexibility to extend the model by adding new predictors in the regression of latent class membership. Furthermore, it is less computationally extensive and thus rather forgiving when adding new parameters to the model. Due to these reasons, the DLC-IRT model is used in this study as the baseline model.

A Dependent Latent Class Item Response Model for Disengaged Responding

The DLC-IRT model was originally proposed by Pokropek (2016) and is based on the grade of membership model of Erosheva (2002), which allows mixing proportions on the item-by-person level. The DLC-IRT model defines a latent class variable

C_{ij} for person i and item j with the two latent classes *solution behavior* ($C_{ij} = 1$) and *disengaged responding* ($C_{ij} = 0$). The latent class is defined on the item-by-person level, allowing the proportions of the latent classes to vary between all item-by-person encounters. The probability of a correct response $P(Y_{ij} = 1)$ for person i on item j follows a mixed distribution and depends on the latent class. In the solution behavior class, the probability of a correct answer is defined by any IRT model. In this study, we define the probability as following a Rasch model (Rasch, 1960):

$$P\left(Y_{ij} = 1 | C_{ij} = 1\right) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}, \tag{1}$$

where Y_{ij} denotes the item response for person i on item j. The parameter θ_i defines the ability of person i, and β_j the item difficulty of item j. The probability of a correct response thus depends both on the ability of the examinee and the difficulty of the item. In the disengaged responding class, however, the probability of a correct response is assumed to be independent of both person and item characteristics and is hence defined by an item-specific constant z_j:

$$P\left(Y_{ij} = 1 | C_{ij} = 0\right) = z_j, \tag{2}$$

The size of z_j depends on the item type. In constructed response items, the probability of a correct response for disengaged responding approximates zero. However, in multiple-choice questions, the test-takers can guess the items correctly even without investing any effort into the response. The probability of a correct response in the disengaged responding class should thus equal the probability of a random guess, which depends on the number of response options M_j for item j. For multiple-choice questions with only one correct response option, we define z_j as:

$$z_j = \frac{1}{M_j}. \tag{3}$$

Overall, the probability of a correct response for person i on item j is thus defined as:

$$P\left(Y_{ij} = 1\right) = \pi_{ij} \frac{\exp\left(\theta_i - \beta_j\right)}{1 + \exp\left(\theta_i - \beta_j\right)} + (1 - \pi_{ij})\, z_j, \tag{4}$$

where π_{ij} defines the probability of belonging to the solution behavior class.

The probability of latent class membership for each item-by-person encounter is regressed on the response engagement indicators. In the model by Pokropek (2016), the only predictor in the logistic regression is the log-transformed item response time l_{ij}:

$$P\left(C_{ij} = 1\right) = \frac{\exp(\alpha + \gamma l_{ij})}{1 + \exp(\alpha + \gamma l_{ij})}, \tag{5}$$

where α defines the general intercept and γ the regression coefficient for the predictor log-transformed item response time. This model assumes that item response times indicate disengaged responding equally well for all items, represented by the general intercept that can be understood as a general item response time threshold. However, the average time test-takers need to spend on the completion of an item might differ between items. Nagy and Ulitzsch (2022) thus proposed to replace the general intercept in Equation 5 with item-specific intercepts for J items, incorporating the varying time demands into the model:

$$P\left(C_{ij}=1\right)=\frac{\exp\left(\sum_{j=1}^{J}\alpha_j d_j+\ \gamma l_{ij}\right)}{1+\exp\left(\sum_{j=1}^{J}\alpha_j d_j+\ \gamma l_{ij}\right)}, \tag{6}$$

where α_j defines the item-specific intercept and d_j a dummy variable indicating whether the current item is item j. As the model is defined on one level only, the authors call it the dependent latent class single level item response theory (DLC-SL-IRT) model. As Equations 5 and 6 assume that persons do not differ regarding their response speed, it is also assumed that individual test-taking speed and ability are not correlated, contradicting several studies that suggest the opposite (e.g., Goldhammer et al., 2016; Lee & Jia, 2014; Nagy & Ulitzsch, 2022). Therefore, Nagy and Ulitzsch (2022) proposed a second extension of Pokropek's model by adding the random effect ψ_i to the logistic regression of latent class membership:

$$P\left(C_{ij}=1\right)=\frac{\exp\left(\sum_{j=1}^{J}\alpha_j d_j+\ \gamma l_{ij}+\psi_i\right)}{1+\exp\left(\sum_{j=1}^{J}\alpha_j d_j+\ \gamma l_{ij}+\psi_i\right)}. \tag{7}$$

The random effect ψ_i reflects between-individual differences in response time thresholds. It is normally distributed with zero mean and the variance σ^2_ψ. In this model, ψ_i is allowed to covary with the individual ability θ_i. As the random effect introduces a second level to the logistic regression, the authors call it a dependent latent class two-level IRT (DLC-TL-IRT) model. In the empirical sample studied by Nagy and Ulitzsch (2022), both the DLC-SL-IRT and the DLC-TL-IRT model fitted the data better than a simple IRT model ignoring disengaged responding. Moreover, the DLC-TL-IRT model exhibited a better model fit than the DLC-SL-IRT model. However, the authors did not compare the extended models to the original DLC-IRT model proposed by Pokropek (2016).

Enhancing the Identification of Disengaged Responses with Process Data

Similar to other identification approaches, the DLC-IRT model solely relies on item responses and item response times as indicators of response engagement, risking misclassifications of slow disengaged or fast engaged responses. Additional indicators of response engagement might thus improve the DLC-IRT model. As process data offers a rich pool of additional information on the test-taking process, it can be a major source of novel indicators.

As of yet, only few studies have investigated the potential of process data for the identification of disengaged responding. In Lundgren and Eklöf (2020), clusters of process data representing test-taking strategy and effort predicted self-reported effort and test performance in an interactive problem-solving task. Sahin and Colvin (2020) showed that the number and type of actions could in some cases slightly improve the response time thresholds in interactive tasks. Ivanova and colleagues (2020) found that the number of actions in constructed response items was related to test performance, self-reported test-taking effort, and item position. Finally, Patel and colleagues (2021) used a machine learning approach to predict test-taking efficiency in a test of mixed item types. Besides measures related to test or item response time, the number of certain actions (e.g., answer changes and navigation button use) were identified as important features of test efficiency.

These findings indicate that process data beyond response times might indeed possess valuable information for the identification of disengaged responding. However, the aforementioned studies focused mostly on interactive or constructed response tasks. Although these item types are becoming increasingly popular in educational large-scale assessments, most competence tests are still based on traditional item formats such as multiple-choice questions. Furthermore, most existing studies were rather exploratory. Therefore, this study further explores the relationship of process data from multiple-choice items of a reading competence test with test-taking engagement. We aim to derive novel indicators of response engagement and incorporate them into a model-based approach of disengagement. In doing so, we hope to refine existing response-time based approaches to facilitate interpretations of disengagement and its correlates.

The Current Study

Development of Novel Indicators of Response Engagement

The novel indicators of response engagement were derived based on theoretical considerations. Each indicator is defined on the item-by-person level, so they can differ between all item-by-person encounters. Initially, all possible kinds of interactions with multiple-choice items in reading tasks were identified. A reading test usually consists of several tasks with a text at the beginning and subsequent items referring to the text. The text and each item are each presented on single pages of the test. Therefore, the test-taker can interact with the task by navigating through the pages,

either between a text and an item page or between two item pages. In multiple-choice items, test-takers can choose between several response options with usually only one response option being correct. The test-taker thus can additionally interact with the items by selecting and changing response options. Based on these considerations, we derived three potential indicators of response engagement.

First, the test-taker can choose to navigate back to the text while working on an item. Thus, we define the indicator *text reread* as the navigation from an item page back to the corresponding text page. When the test-taker navigates back to the text, it can be assumed that she or he has actually read the content of the item and now wants to collect more information to be able to respond to the item. The test-taker hence allocates resources to the item response, contradicting the definition of disengaged responding and thus indicating solution behavior.

Second, the test-taker might navigate to an item page she or he has already visited before. We thus define the indicator *item revisit* as the repeated navigation to an item page. The revisit of an item might occur for different reasons. The test-taker might (a) first scan the item and return to it later for the actual response, (b) return to a difficult item she or he was not able to solve before, (c) review an item she or he has already responded to, or (d) navigate aimlessly through the item pages. In the first three cases, the test-taker shows her or his will to properly engage with the item. However, in the last case, no conclusions regarding the test-taking behavior on single items can be drawn, as a random navigation behavior might occur either in between engaged item responses or due to a state of disengagement. As we do not expect this behavior to occur frequently, we assume that the revisit of an item nevertheless indicates solution behavior.

Third, a test-taker might change an already provided response to an item. Thus, we define the indicator *answer change* as changing the previously selected response option to another one. An answer change might occur for three reasons. The test-taker might (a) review the item and subsequently change their belief about the correct response option, (b) correct an accidentally provided response, or (c) select accidentally another response option. In the first two cases, the test-taker has apparently read the item and allocated effort and knowledge into the response, indicating solution behavior. In the third case, however, no strong assumptions regarding the test-taking behavior can be made and thus can be seen as noise in the data.

In total, we identified three novel indicators that can be derived from process data in multiple-choice items of a reading test: text reread, item revisit, and answer change. All three indicators are expected to indicate solution behavior and thus contradict disengaged responding.

Research Questions and Hypotheses

To evaluate the new indicators of response engagement, they were incorporated into an extended version of the DLC-IRT model and compared with a baseline model that only acknowledged item response times. As a potential baseline model, we

considered the original DLC-IRT (Pokropek, 2016), the DLC-SL-IRT, and the DLC-TL-IRT model (Nagy & Ulitzsch, 2022). The extended DLC-IRT model included the new indicators as additional predictors in the logistic regression of the latent class membership as defined in Equations 5 to 7. Note that the interpretation of the item-specific intercepts d_j and the random effect ψ_i alter when additional engagement indicators are added to the regression. In this case, these parameters rather represent item-specific engagement difficulties and an individual propensity to show solution behavior, respectively.

We expected the new indicators to enhance the baseline DLC-IRT models by improving the classification of response engagement. In the new model, we assumed item response time to be a significant predictor of latent class membership, replicating the findings of Pokropek (2016) and Nagy and Ulitzsch (2022). However, we did not expect all auxiliary indicators to be equally informative. Specifically, we expected the indicator *item revisit* to be at least partly redundant with item response times, as revisiting the item page simultaneously extends the item response time by the time of the additional visit. Instead, we expected the indicators *answer change* and *text reread* to be more informative and thus the only significant new predictors of latent class membership. We hence propose the following hypotheses:

1. The extended DLC-IRT model describes the observed responses better than the best baseline model.
2. The indicators *item response time*, *answer change*, and *text reread* predict latent class membership, so that responses are more likely to be classified as solution behavior when test-takers need more time for the response, change their answer, or reread the text.
3. The indicator *item revisit* does not predict latent class membership over and above item response time.

The relationship between response engagement and its indicators might not solely be linear. For example, the indicators might only improve the classification of response engagement when response times are long enough that the test-takers could have read and processed the item. To allow also for nonlinear effects, we also propose a second extension of the DLC-IRT model by further introducing interactions between the indicators as additional predictors in the logistic regression. As this extension is exploratory, we include all six possible interaction terms in the regression but have no specific a priori hypotheses.

Method

Sample

This study used a subsample from the student cohort in the NEPS, which assesses competence development and educational trajectories across the life span (Blossfeld & Roßbach, 2019). In Wave 12, reading competence was assessed in a quasi-experimental

design, with the sample being split between a computer-based, proctored and a web-based, unproctored test setting. To avoid setting or mode effects, only the data of the web-based test was used. The 1,932 participants (59.4% female) were aged from 23 to 59 years (M = 28.06, SD = 3.59). The majority identified German as their native language (94.9%) and was employed at the time of assessment (85.8%).

Instruments and Measures

Test of Reading Comprehension. The test of reading comprehension in German was developed specifically for the NEPS (Gehrer et al., 2012) and included five reading tasks. Each task consisted of a short text and a respective item set referring to this text. The test comprised a total of 21 items with different response formats. However, for the present analyses, only the 14 simple multiple-choice items were used. Each of these items included four response options with one correct solution and three distractors.

After a standardized instruction, the main part of the test allowed participants to navigate within each task, but not between different tasks. Each text and each item was presented on different pages (only very long texts were presented on two consecutive pages). Navigation between different pages within a task was possible by either selecting the forward or backward button on each page or by clicking on a specific page name in the navigation panel. Thus, all text and item pages could be visited several times. The tasks could be finished by selecting the forward button on the last page of the task. However, participants were then asked if they definitely wanted to leave the task, as they were unable to return to it later. Response options of the multiple-choice items could be selected by a mouse click on the respective radio button.

Indicators From Process Data. Process data were continuously logged while the participants took the test. Mouse clicks, the selection of buttons, and the use of the navigation panel were each logged with a corresponding time stamp. This study includes only navigation and response option selection events (N = 85,848). Preliminary analyses revealed that several participants navigated through the test using only the forward and backward buttons. This resulted in numerous navigation events that did not reflect the participants' intention to *stay* on the selected test page but rather occurred as intermediate navigation steps. These irrelevant navigation events prevent an accurate assessment of some of the response engagement indicators and were thus filtered from the data, resulting in a new sample of 71,070 log events. More information on the implementation of the navigation filter and the exclusion criteria can be found in the Supplementary Material.

The four disengaged responding indicators were each defined on the item level. *Item response time* was defined as the time the participant spent in total on the respective item page. To determine the item response times, the time interval for each page visit was calculated. The start of the time interval was defined as the click on a navigation button that led to the respective item page and the end was defined as the click

Table 1. Frequency Distributions for New Indicators of Response Engagement.

Indicator	Frequency				
	0	1	2	3	>3
Answer change	21,642	2,224	1,017	182	63
Text reread	16,875	4,933	2,024	784	512
Item revisit	23,440	1,521	145	17	9

on a navigation button to leave the respective item page. If a participant visited the page several times, the item response time was defined as the sum of all page visit time intervals. As the distribution was heavily right-skewed, the log response time was used for the analyses.

An *answer change* was registered each time when a participant had already selected one response option before and, subsequently, selected another response option within the same item. A *text reread* was defined as navigating from the respective item page to a text page. An *item revisit* was registered when three conditions were met. First, the participant navigated to the respective item page. Second, it was not the first visit to this item page. Third, between the last and the current visit of the item, at least one other item was visited (to distinguish text rereads from item revisits). The frequency distributions of the three new indicators are given in Table 1. As all indicators were severely skewed, they were dichotomized for the analyses (1 = occurred at least once, 0 = did not occur).

Statistical Analyses

Data preparations, descriptive analyses, and visualizations were conducted in R (Version 4.0.3/4.2.2; R Core Team, 2021), whereas the IRT models were estimated in *Mplus* (Version 8.4; Muthén & Muthén, 1998-2017). Although the scored test data are provided at NEPS Network (2022), the process data cannot be made publicly available due to legal restrictions. However, the computer code and analysis results are provided in the Supplemental Material.

We fitted six different IRT models to the data, beginning with a Rasch model that ignored disengaged responding. Then, the original DLC-IRT model, the DLC-SL-IRT model, and the DLC-TL-IRT model as defined in Equations 1 to 7 with item response time as the only indicator of response engagement were estimated as potential baseline models. The models were compared using Akaike's Information Criteria (AIC; Akaike, 1974), Bayesian Information Criteria (BIC; Schwarz, 1978), and sample size adjusted BIC (aBIC), as well as log-likelihood ratio tests. The best fitting model was used as the baseline model for further analyses.

As fifth model, the baseline model was extended by adding the three new dichotomized response engagement indicators as additional predictors in the regression of latent class membership. In the sixth model, all possible interaction terms of the four

indicators were added as further predictors to the same regression. Both extended models were then compared with the Rasch and baseline model using information criteria and deviance tests.

All models were implemented as defined in Equations 1 to 7 by adapting the Mplus syntax provided in the Supplementary Material of Nagy and Ulitzsch (2022). The probability of a correct response followed a Rasch model in the solution behavior class, but was set to chance level (25%) for all items in the disengaged responding class. When the item response time didn't exceed zero seconds (n = 1,934), which only occurred in combination with a missing item response, all four process data indicators were defined as missing. As small numbers of missing responses do not introduce severe parameter bias in IRT models (Rose et al., 2010), missing data were ignored in the analyses. Starting values were set to 80/8 and increased to 250/25 when the best loglikelihood had not been replicated. All models were estimated using the marginal maximum likelihood procedure. Entropy as an index of classification accuracy (Asparouhov & Muthén, 2018) is reported for each model. Because previous research suggests that the probability of correct guesses may indeed exceed what should be expected by chance alone (e.g., Ulitzsch et al., 2021), we re-estimated in a sensitivity analysis the best fitting model with equal, but not fixed item difficulties in the disengaged responding class, and report the results in the Supplementary Material.

Results

Descriptive Results

The mean item response times (ranging from 17.35 to 36.24 seconds, Mdn = 27.77) and the proportions of participants who changed their answer at least once (ranging from 8.5% to 20.7%, Mdn = 13.5%) differed only slightly between items (see Table 2). We observed more pronounced differences for the proportion of correct answers (ranging from 29.8% to 83.8%, Mdn = 52.9%) and the percentage of participants who reread the text (ranging from 5.7% to 68.2%, Mdn = 35.8%) or revisited the item at least once (ranging from 1.0% to 14.0%, Mdn = 5.3%). Although all four indicators correlated moderately with each other (see Table 3), we did not observe substantial multicollinearity. The largest correlation resulted between the log item response times and the dichotomized text rereads (r = .30, $p < .001$).

Baseline Models for Disengaged Responding

In the first step, the baseline model for disengaged responding was identified. To this end, we compared the ordinary Rasch model that does not acknowledge disengaged responding, the DLC-IRT with a common intercept for all items, the DLC-SL-IRT with item-specific intercepts, and the DLC-TL-IRT with an individual threshold for response engagement (see Table 4). Importantly, the three DLC models only included item response time as a predictor of class membership. Because the DLC-SL-IRT model was empirically underidentified when considering unique intercepts for all

Table 2. Descriptive Results Per Item.

Item	Item-by-person encounters	Response time	Item responses	Answer change	Text reread	Page revisit	Probability to belong to the solution behavior class
	N	*M* (*SD*)	Percentage correct	At least once	At least once	At least once	
1	1,930	30.17 (28.09)	70.7%	15.1%	68.2%	14.0%	89.8%
2	1,928	36.24 (18.58)	64.8%	18.6%	49.5%	11.9%	86.7%
3	1,928	27.87 (15.77)	32.9%	12.4%	55.9%	5.0%	43.5%
4	1,898	26.06 (18.27)	83.8%	20.7%	11.0%	4.4%	97.8%
5	1,900	21.67 (12.85)	49.9%	11.7%	5.7%	2.9%	95.9%
6	1,872	27.67 (14.62)	54.1%	14.6%	40.3%	10.4%	96.9%
7	1,867	21.03 (12.74)	50.0%	15.4%	7.3%	9.6%	67.3%
8	1,856	29.85 (17.64)	38.2%	16.9%	40.0%	11.7%	96.4%
9	1,849	33.03 (17.53)	71.2%	8.5%	8.3%	9.3%	96.5%
10	1,808	27.36 (13.68)	39.5%	12.1%	36.1%	1.4%	96.1%
11	1,733	32.55 (27.07)	51.6%	15.1%	35.5%	1.2%	68.9%
12	1,617	32.44 (17.80)	63.6%	9.7%	18.7%	1.0%	94.9%
13	1,464	22.34 (12.32)	29.8%	11.7%	28.8%	5.6%	93.7%
14	1,413	17.35 (12.23)	69.9%	9.8%	57.1%	3.3%	94.8%

Table 3. Correlations Between Indicators of Response Engagement and Item Responses.

Variable	*M*	*SD*	1	2	3	4
1. Item response	0.55	0.50				
2. Response time	3.16	0.59	.08***			
3. Answer change	0.14	0.35	–.01	.23***		
4. Text reread	0.33	0.47	.01*	.31***	.12***	
5. Page revisit	0.07	0.25	.00	.18***	.16***	.10**

Note. Presented are Pearson correlation coefficients. Response time refers to logarithmized response time.
*$p < .05$. **$p < .01$. ***$p < .001$.

items, item-specific intercepts were identified using a sequential process. First, a model was estimated for each item that included a general intercept and the respective item-specific intercept. Then, this model was compared with the DLC-IRT model without the item-specific intercept. Only those item-specific intercepts were retained for which model comparisons provided a better fit to the data. In the end, the DLC-SL-IRT model comprised a general intercept and item-specific intercepts for Items 1, 2, 3, 7, 11, and 14. The model comparisons showed the best fit for the DLC-TL-IRT

Table 4. Model Comparisons for Item Response Models.

Number	Model	n_{par}	Information criteria			Entropy	χ^2-test		
			AIC	BIC	aBIC		ΔDev	*df*	*p*
1.	Rasch	15	31,106	31,106	31,180				
	Baseline models								
2.	DLC-IRT	17	30,944	31,082	31,028	0.98	166.09	2	<.001
3.	DLC-SL-IRT	23	30,730	30,917	30,844	0.93	225.81	6	<.001
4.	DLC-TL-IRT	25	30,697	30,900	30,821	0.87	37.35	2	<.001
	Extended models								
5.	DLC-TL-IRT + new indicators	28	30,678	30,906	30,817	0.87	24.68	3	<.001
6.	DLC-TL-IRT + new indicators + interactions	34	30,659	30,935	30,827	0.86	31.68	6	<.001

Note. aBIC = sample-size adjusted BIC; ΔDev = difference in deviance; DLC-IRT = dependent latent class item response theory; DLC-SL-IRT = single-level DLC-IRT; DLC-TL-IRT = two-level DLC-IRT.

n_{par} = number of parameters; AIC = Akaike's information criterion (Akaike, 1974); BIC = Bayesian information criterion (Schwarz, 1978).

Table 5. Results of Hierarchical Linear Modeling.

Parameter	Estimate	*SE*	*t*	*P*
Within level				
General intercept	–21.52	4.65	–4.63	<.001
Intercept Item 1	–6.34	1.57	–4.04	<.001
Intercept Item 2	–9.83	2.16	–4.56	<.001
Intercept Item 3	–13.93	2.74	–5.08	<.001
Intercept Item 7	–8.43	1.78	–4.74	<.001
Intercept Item 11	–13.17	2.76	–4.77	<.001
Intercept Item 14	3.92	2.21	1.78	.076
Item response time	11.26	2.28	4.94	<.001
Answer change	–0.76	0.79	–0.96	.339
Text reread	–2.60	0.79	–3.30	.001
Item revisit	–1.43	1.01	–1.42	.156
Between level[a]				
Variance of ψ_i	7.26	3.69	1.97	.049
Covariance between ψ_i and θ_i	2.69	0.68	3.94	<.001
Correlation between ψ_i and θ_i	1.00	0.01	68.94	<.001

Note. ψ_i defines the individual threshold for response engagement and θ_i the individual reading proficiency of person *i*.
[a]The variance of θ_i was fixed to 1 for identification.

model with six item-specific intercepts and a random effect for the individual speed. This model was therefore used as the baseline model for the subsequent analyses.

Extended Models for Disengaged Responding

As compared with the baseline model, the loglikelihood ratio test and the AIC indicated a better fit of the extended DLC-IRT model that included the main effects of the new indicators of response engagement and also the extended DLC-IRT model that additionally included the respective interactions (see Table 4). Moreover, both criteria slightly favored the most complex model also including interactions. In contrast, the aBIC slightly favored the extended model without interactions, while the BIC slightly favored the baseline model over both extended models. Given the mixed results, we selected the simpler model without interactions for the subsequent analyses (=*extended model*), thus supporting Hypothesis 1.

The entropy of the extended model was 0.87 and the average probability for item responses to be classified as solution behavior based on the posterior probability distribution was 86.7%. However, as can be seen in Table 2, this probability differed strongly between items. Item 3 exhibited an average probability of 43.5% and Item 4 exhibited an average probability of 97.8% for item responses to belong to the solution behavior class. More items were responded correctly (60.2%) and participants needed longer to respond to an item (M = 29.67 seconds, SD = 17.93) in the solution behavior

class than in the disengaged responding class (19.4%, $M = 14.15$ seconds, $SD = 8.07$). Furthermore, participants more often changed their answer (14.8%), reread the text (33.7%) and revisited the item (7.2%) in the solution behavior class than in the disengaged responding class (7.8%, 26.0%, and 3.2%, respectively).

The item-specific intercepts for Items 1, 2, 3, 7, and 11 were significant ($p < .05$), with a smaller probability to be classified into the solution behavior class for these items (see Table 5). The variance of the random effect for the individual speed was significant ($Var = 7.26$, $p = .049$) and correlated perfectly with the latent reading ability ($r = –1.00$, $p < .001$). However, this result should be interpreted only carefully as the standard error of the variance of the random effect was rather large ($SE = 3.69$). Item response time and text reread were significant predictors of latent class membership. The longer a participant needed for an item response, the higher was the probability that this item response belonged to the solution behavior class ($B = 11.26$, $p < .001$). In contrast to our expectations, rereading the text while working on an item was associated with a higher probability of being *disengaged* ($B = -2.60$, $p = .001$). The other indicators were no significant predictors of latent class membership. Hypothesis 2 could thus only be partly confirmed for the predictor item response time, while Hypothesis 3 was supported by the results.

One of the assumptions of this study was that the new indicators of response engagement can improve the DLC-IRT model by reducing the error rates of the (dis)-engagement classification, as slow disengaged or fast engaged responses might be misclassified when item response time is the only predictor of latent class. Figure 1 displays the distribution of item response times by latent class membership for both the baseline and the extended model. As the item response time distributions look basically the same for both models, the new indicators do not seem to substantially reduce the error rates of engagement classification.

A second assumption of this study was that the new indicators of response engagement can reduce the bias in person and item parameters by improving the prediction of latent class membership. As can be seen in Figure 2, the competence estimates were systematically underestimated in the Rasch model when participants tended to respond in a disengaged manner. However, the figure does not show any systematic difference in competence estimates between the baseline and the extended model. The item difficulties are displayed in the Supplementary Material. Figure 3 indicates that the Rasch model overestimated item difficulties for some of the items (notably the items for which item-specific intercepts were included in the model). However, no major changes in item difficulties between the baseline and the extended model could be observed. The new indicators of response engagement thus do not seem to substantially reduce bias in the person or item parameters.

Discussion

Because disengaged responding poses a serious threat to the validity of educational large-scale assessments, various methods have been proposed to identify them and correct for their distorting influence. Most of these approaches rely solely on item

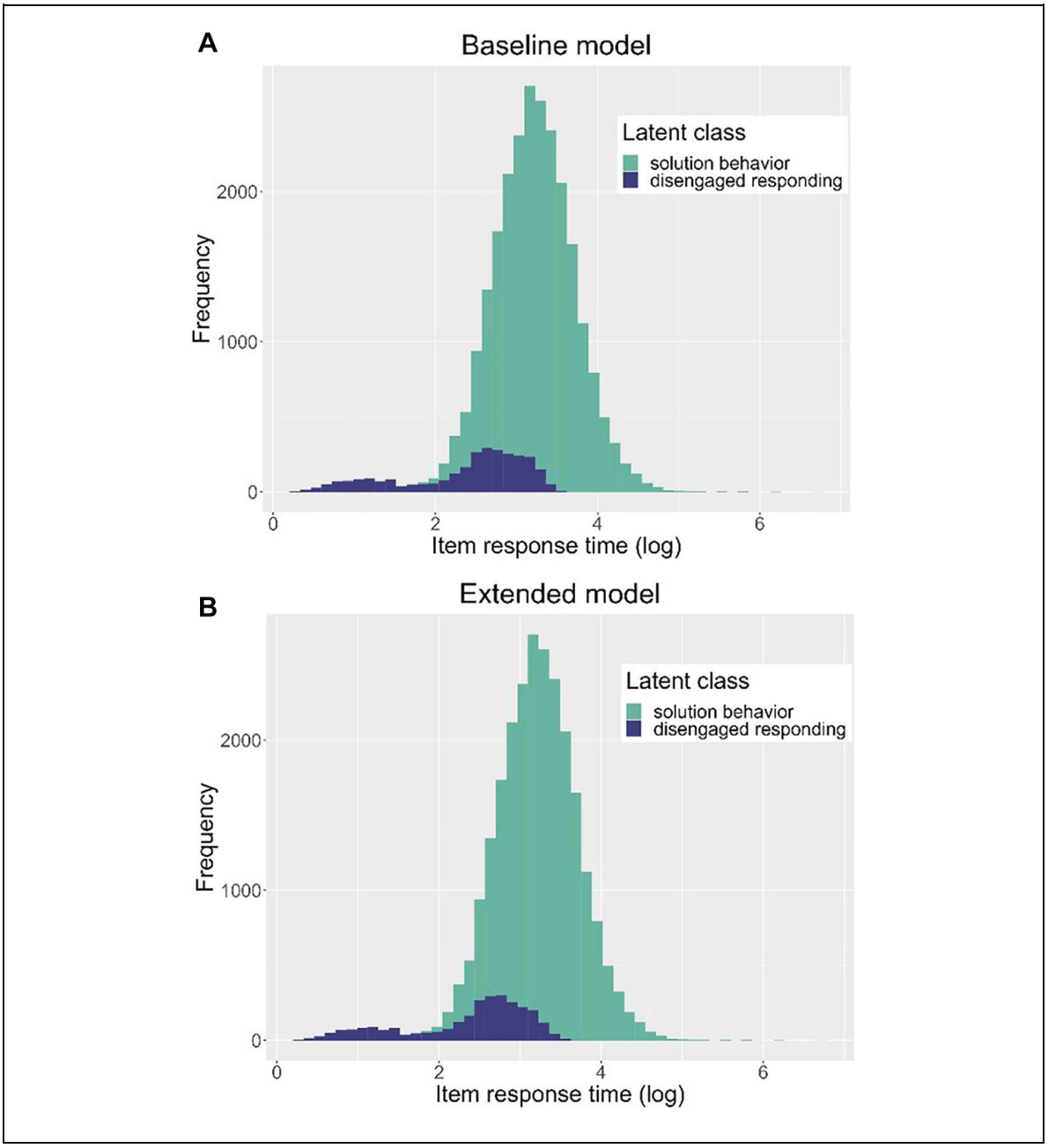


Figure 1. Distribution of Item Response Times by Latent Class and Response Engagement Model.

responses and item response times (e.g., Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015), while ignoring potentially useful information on the test-taking process provided by additional process data. To further enhance our understanding of test-taking behavior, this study derived novel indicators of response engagement from process data and evaluated whether they improve model-based identification approaches of disengaged responding. In a sample of German university students, an application of the extended latent class item response model including the three new

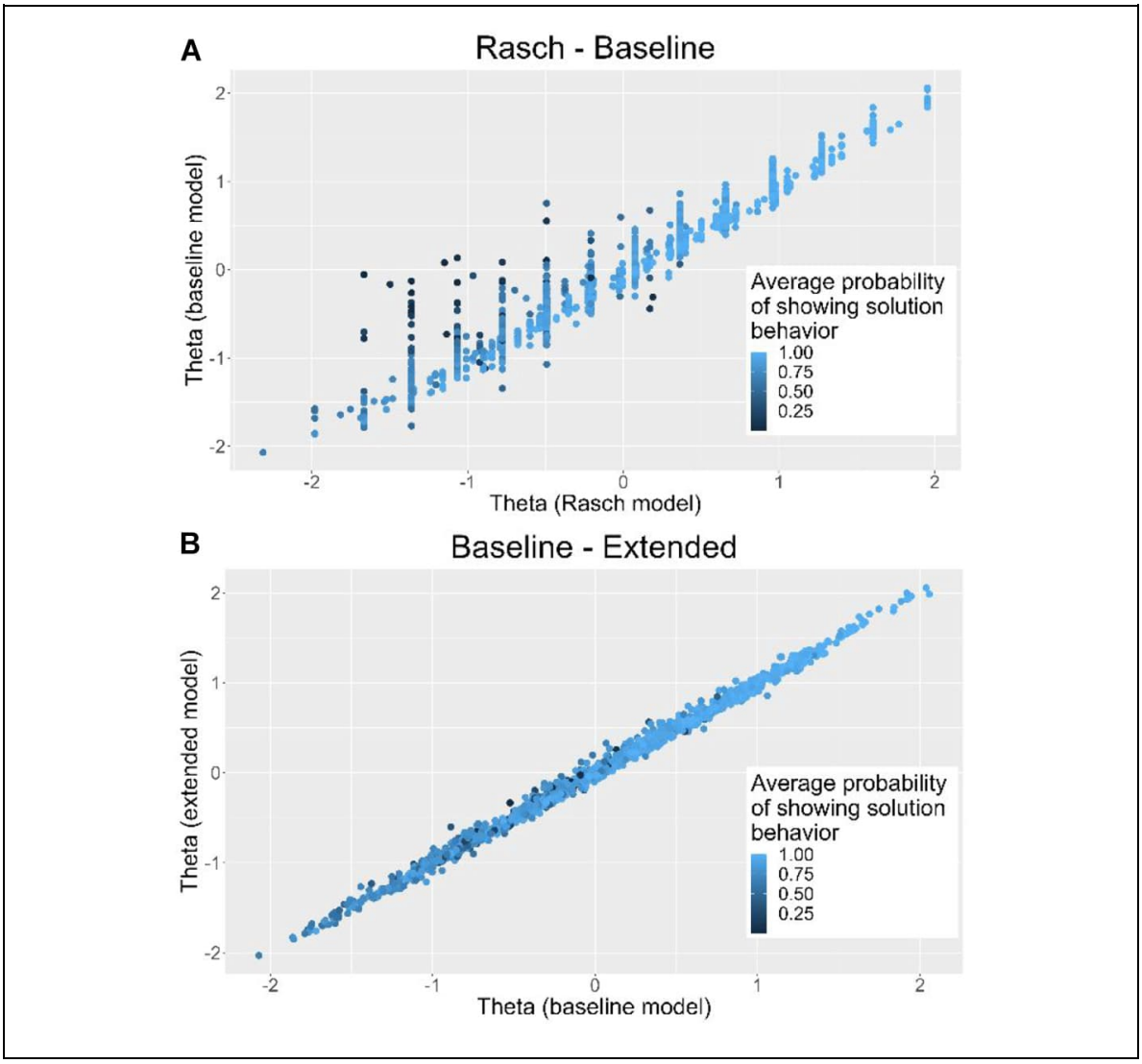


Figure 2. Comparison of Proficiency Estimates Between the Rasch, Baseline, and Extended Model.

indicators as additional predictors of response engagement led to three central findings.

In line with Hypothesis 1, the extended DLC-TL IRT model that included the new indicators showed a better fit as compared with a model with response times alone, supporting the finding of previous studies that process data beyond response time might indeed provide valuable information on response engagement (Ivanova et al., 2020; Lundgren & Eklöf, 2020; Patel et al., 2021; Sahin & Colvin, 2020). However, a closer inspection of the results revealed that only item response times and text reread predicted response engagement while answer change and item revisit did not. Supporting Hypothesis 2, shorter item response times were associated with a greater likelihood of disengaged responding, thus replicating numerous previous studies on disengaged responding (e.g., Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015).). Similar to the findings of Patel and colleagues (2021), the result

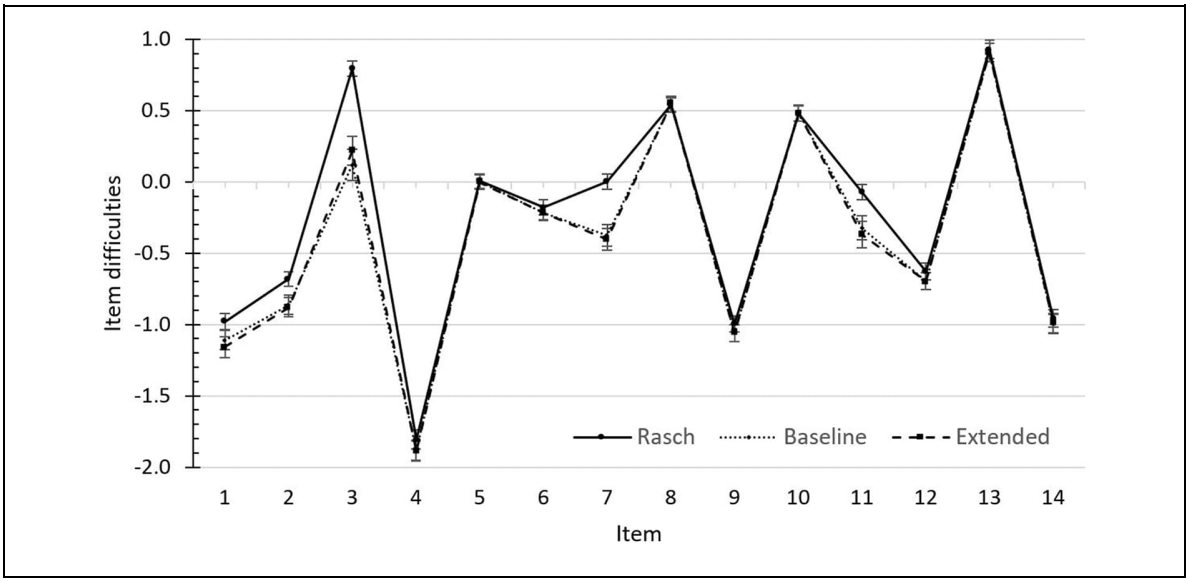


Figure 3. Comparison of Item Difficulties Between the Rasch, Baseline, and Extended Model.

further indicates that also in the presence of other process data indicators, item response time stays a principal indicator of response engagement. Moreover, it suggests that item response time is a valid indicator of response engagement also in unproctored settings when test-takers can engage in other activities while working on the test or even choose to temporarily leave the test.

Over and above this effect, rereading the text while working on an item predicted the probability of disengaged responding. Although the direction of the effect was in contrast to our initial hypotheses, this might reflect a suppression effect. Text rereads were moderately correlated with item response times and thus the information gained from this indicator is partially redundant with the information gained from item response time. The descriptive results support this assumption, since the percentage of text rereads was lower for disengaged item responses than for engaged item responses. However, text rereads seem to further distinguish between disengaged responding and solution behavior after conditioning on the item response times. In this case, text rereads might help to identify disengaged test-takers that navigate back to the text page rather aimlessly instead of directly guessing the item at hand. Finally, the two remaining indicators, answer change and item revisit, showed no incremental contribution to the prediction of latent class membership. This is likely a consequence of the restricted range of these indicators because both behaviors were rarely observed and, thus, provided little information about the test-taking process. Furthermore, these indicators were once more redundant with the information gained from the item response times. Taken together, the results highlight that indicators derived from process data can inform about response engagement. But whether the information they provide is relevant beyond response times strongly depends on the specific indicator.

Despite the relevance of text rereads for the prediction of response engagement, we found no systematic bias in students' reading proficiencies or item difficulty when not accounting for the additional indicators of response engagement. Furthermore, histograms of item response times by latent class revealed no visible classification differences between the models for both slow and fast item responses. These results contradict our assumption that the novel indicators might be able to improve the classification of (dis)engaged responses in cases where item response times do not suffice as the only indicator (e.g., fast engaged and slow disengaged responses).

Accuracy of Response Engagement Classification

In the extended model, the probability of showing solution behavior varied drastically between items, with one item even exhibiting a probability of less than 50%. In other model-based approaches, however, prevalence rates per item do usually not fall below 76% (e.g., Nagy & Ulitzsch, 2022; Ulitzsch et al., 2021). Moreover, response options in the disengaged responding class were selected with differing frequency (rates ranged between 1% and 67%), instead of being equally distributed, as should be expected for random guesses. Frequently selected response options were mostly wrong response options, indicating rather strong distractors than random guessing. This assumption is supported by the fact that, averaged over all items, the probability of a correct response in the disengaged responding class was estimated to be less than chance level. Freeing the item difficulties of the disengaged responding class in the sensitivity analysis resulted in an even lower probability of a correct response for disengaged responses. Furthermore, graphical analyses revealed that also numerous slow responses were classified as being disengaged, both in the baseline and extended model. Finally, the individual threshold for response engagement exhibited a nearly perfect correlation with reading proficiency. Although response engagement and ability were repeatedly shown to be related, the size of the correlation is usually estimated to range between .25 and .70 (e.g., Goldhammer et al., 2016; Ulitzsch et al., 2021). This suggests rather limited evidence for individual differences in the threshold for response engagement in the present sample. After all, the respective random variance was only borderline significant. Consequently, the respective correlation should be interpreted only carefully.

These findings indicate that in the present response engagement model, more responses were classified as disengaged than should be expected when the assumption that disengaged responses are equal to random guesses were true. Thus, also responses that were provided with little or decreasing effort or engaged responses of low-ability test-takers might have been classified as disengaged. As rereading the text predicted disengaged responding, ''disengaged'' test-takers seem to have started solving an item and reread the text to find the correct answer, but might not have perceived the correct response in the end. The strong correlation between reading proficiency and response engagement further suggests that these two constructs were to some degree confounded in the dependent latent class item response model of this study.

Since these results were also found (though sometimes a bit extenuated) in the baseline model, they are not explainable by the model extension alone. In contrast to Nagy and Ulitzsch (2022), the present study implemented the response engagement model in multiple-choice questions instead of open-end questions. The DLC-TL-IRT model might thus have difficulties distinguishing between disengaged wrong guesses and engaged wrong answers when the success probability of disengaged responses does not approximate zero. However, also the setting, the test, and the sample differed between the two studies. Whether one of these factors might have influenced the outcome as well, can only be answered by future studies testing the applicability of the DLC-TL-IRT model in different conditions.

Implications for Educational Large-Scale Assessments

The results emphasize the importance of item response time as an indicator of response engagement while estimating the incremental value of the selected additional process data indicators as rather weak. One possible explanation for this finding is that the vast majority of disengaged responding might indeed be rapid guessing (as suggested by Wise et al., 2020), rendering item response times a sufficient single indicator of response engagement. However, even if disengaged responding beyond rapid guessing did exist, the additional process data indicators might simply not provide enough information on the test-taking process to improve the classification of response engagement. Combining or rearranging the selected indicators (e.g., by clustering) or including other process data (e.g., mouse or eye movements) as potential indicators might improve the explanatory power. However, the results could also indicate that process data might generally possess only weak explanatory power for response engagement in multiple-choice items, as only little interaction between the test and the test-takers takes place. Future studies should thus systematically investigate the predictive value of different process data indicators for response engagement in multiple-choice items, while also taking into account other measures of disengagement (e.g., self-reports).

Our findings suggest that the DLC-TL-IRT model did not only classify random guesses as disengaged. We see two different explanations for this finding. First, item responses might not simply be classified as either engaged or disengaged, but some intermediate steps or continuum might exist which are confounded with the ability in the test domain. A. Pokropek (personal communication, July 14, 2022) therefore suggested adding a *slow guessing* latent class to the DLC-IRT model. This latent class includes all responses of test-takers who first try to solve the item but then give up, resulting in long item response times and interactions with the item (e.g., rereading the text). Second, the result might also indicate that the DLC-IRT model does not function in multiple-choice items as well as in open-end item formats. Future studies should thus further investigate the applicability of the different DLC-IRT models to tests with multiple-choice items and compare it to other existing identification methods of disengaged responding. In addition, it might be insightful to include a third

response engagement class in the DLC-IRT models and compare it to models with only two latent classes.

Strengths and Limitations

This study contributed to the literature on disengaged responding in several ways. On one hand, it is one of rather few attempts to evaluate the potential of process data beyond item response times for the identification of disengaged responding and, thus, deepens our understanding of influential test-taking behavior affecting outcomes of reading comprehension tasks. On the other hand, the study not only investigated the validity of a model-based approach for the identification and handling of disengaged responding, but also built on these prior developments to improve existing methods. Therefore, the results of this study can help to further minimize the threat to test validity introduced by disengaged responding.

Besides its considerable strengths, the study also possesses some limitations. First, for reasons of practicability, we chose a study with a rather homogeneous sample and a rather short test. Because previous studies showed that age, educational attainment, and country can be associated with disengaged responding (e.g., Goldhammer et al., 2016; Lindner et al., 2019), more representative samples might identify more disengaged responses and, thus, stronger effects for the novel indicators. Second, the extended model included only one random effect for all four response engagement indicators. By not accounting separately for individual differences in the indicators, the model ignores potential relationships between the indicators on the person-level (which do exist; e.g., Bezirhan et al., 2021). Future studies could handle this problem by including random slopes in the model. However, random slopes are computationally intensive and thus could lead to non-identified models. Third, item properties were not included in the present study as potential indicators of response engagement. Previous studies have shown that disengaged responding is indeed related to item parameters such as item position (Goldhammer et al., 2016) and item content (Rios & Soland, 2022). Fourth, the indicator item revisit might not have been very reliable. As outlined in the Supplementary Material, we filtered irrelevant navigation events to enable the determination of reliable indicators. However, the filter might not be perfect and thus result in falsely identified item revisits, obscuring potential relationships with other variables.

Future studies might thus include additional parameters to the DLC-IRT models, further explaining the complex relationships between the indicators and adding new explanatory variables. Other promising future directions of research could be to test the generalizability of the current findings by applying them to other item formats or other test domains. All three process data indicators can also be defined for other item formats such as complex multiple-choice tasks and text enrichment tasks, although the indicator answer change might have to be slightly altered. As test-takers interact more with the test items of other formats, the additional indicators might convey more information on the test-taking process in these cases. Similarly, the process data

indicators can mostly be transferred to other test domains, such as mathematics or foreign language tests. Naturally, the indicator text reread only applies to tests with an initial text or task description and respective item sets.

Conclusion

The advent of computer-based assessments offers with process data a wide array of new information on the test-taking process. Nonetheless, studies investigating the potential of process data for the identification of disengaged responding are still rare. To fill this gap, this study sought to identify new process data indicators of response engagement in multiple-choice items and determine their predictive value. Three new potential indicators were introduced: answer change, text reread, and item revisit. In an empirical study, a DLC-IRT model with item response time as the only predictor of response engagement was compared with a DLC-IRT model additionally including the new indicators. The results suggest that the incremental predictive value of the new indicators is weak, while short item response times stay an important indicator of disengaged responding. Counterintuitively, text rereads seem to be associated with disengaged responding for equal item response times. However, further findings also indicate that in the present study, the DLC-IRT model does not only identify random guesses as disengaged but also presumably engaged but wrong answers. Future studies shall investigate whether this generally applies to tests based on multiple-choice items. Overall, this study contributes to the literature on disengaged responding by enhancing our understanding of test-taking behavior and further improving existing identification methods, thus minimizing the threat to test validity introduced by disengaged responding in the first place.

Authors' Note

This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network. The study was not preregistered. The scored test data are provided at NEPS Network (2022), whereas the process data cannot be made publicly available due to legal restrictions. The computer code and analysis results are provided in the Supplemental Material.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jana Welling https://orcid.org/0000-0002-6377-8472
Timo Gnambs https://orcid.org/0000-0002-6984-1276

Supplemental Material

Supplemental material for this article is available online.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Asparouhov, T., & Muthén, B. (2018). Technical appendix: Variable-specific entropy contribution. In L. K. Muthén & B. O. Muthén (Eds.), *Mplus user's guide* (7th ed.).

Bezirhan, U., Von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: The hierarchical speed-accuracy-revisits model. *Educational and Psychological Measurement*, *81*(2), 363–387. https://doi.org/10.1177/0013164420950556

Blossfeld H.-P. & Roßbach H.-G (Ed.). (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (ZfE 2nd ed.). Springer VS.

Demars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing*, *10*(3), 207–229. https://doi.org/10.1080/15305058.2010.496347

Deribo, T., Goldhammer, F., & Kroehne, U. (2022). Changes in the speed-ability relation through different treatments of rapid guessing. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/00131644221109490

Erosheva, E. A. (2002). *Grade of membership and latent structure models with application to disability survey data* [Doctoral dissertation, Princeton University].

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for grade 5 and 9)* (Scientific Use File 2012, Version 1.0.0). University of Bamberg, National Educational Panel Study.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. In OECD (Ed.), *OECD education working papers* (No. 133). OECD. https://doi.org/10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education*, *5*, Article 18. https://doi.org/10.1186/s40536-017-0051-9

Ivanova, M., Michaelides, M., & Eklöf, H. (2020). How does the number of actions on constructed-response items relate to test-taking effort and performance? *Educational Research and Evaluation*, *26*(5–6), 252–274. https://doi.org/10.1080/13803611.2021.1963939

Jin, K.-Y., Siu, W.-L., & Huang, X. (2022). Exploring the impact of random guessing in distractor analysis. *Journal of Educational Measurement*, *59*(1), 43–61. https://doi.org/10.1111/jedm.12310

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, *62*(2), 147–177.

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), Article 8. https://doi.org/10.1186/s40536-014-0008-1

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, *10*, Article 1533. https://doi.org/10.3389/fpsyg.2019.01533

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation*, *26*(5–6), 275–301. https://doi.org/10.1080/13803611.2021.1963940

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451. https://doi.org/10.3102/1076998614559412

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.).

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*, *82*(5), 845–879. https://doi.org/10.1177/00131644211045351

NEPS-Network. (2022). *National educational panel study, scientific use file of starting cohort first-year students*. Leibniz-Institute for Educational Trajectories (LIfBi). https://doi.org/10.5157/NEPS:SC5:17.0.0

Patel, N., Sharma, A., Shah, T., & Lomas, D. (2021). Modeling NAEP test-taking behavior using educational process analysis. *Journal of Educational Data-mining*, *13*(2), 16–54. https://doi.org/10.5281/zenodo.5275312

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.0.3/4.2.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, *27*(4), 356–373. https://doi.org/10.1080/10627197.2022.2110465

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. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, *22*(2), 154–184. https://doi.org/10.1080/15305058.2022.2036161

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), i–53. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education*, *8*(1), Article 5. https://doi.org/10.1186/s40536-020-00082-1

Schleicher, A. (2019). *PISA 2018 insights and interpretations*. OECD.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Ulitzsch, E., Penk, C., ,Von Davier, M., & Pohl, S. (2021). Model meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment*, *26*(2), 104–124. https://doi.org/10.1080/10627197.2020.1858786

Ulitzsch, E., Von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(Suppl.1), 83–112. https://doi.org/10.1111/bmsp.12188

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

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

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

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

Wise, S. L., Soland, J., & Bo, Y. (2020). The (non)impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57–77. https://doi.org/10.1080/15305058.2019.1605999

Appendix B

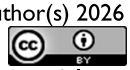
Welling, J. [1], Zink, E. [1], & Gnambs, T. (2026). Comparing different approaches of (not) accounting for rapid guessing in plausible values estimation. *Educational and Psychological Measurement*, Advance Online Publication. https://doi.org/10.1177/00131644251395590

[1] These authors contributed equally

Supplemental Material: https://osf.io/gdnse

Check for updates

Original Research Article

Educational and Psychological Measurement
1–30
© The Author(s) 2026

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644251395590
journals.sagepub.com/home/epm


Comparing Different Approaches of (Not) Accounting for Rapid Guessing in Plausible Values Estimation

Jana Welling[1]*, Eva Zink[1]* and Timo Gnambs[1]

Abstract

Educational large-scale assessments provide information on ability differences between groups, informing policies and shaping educational decisions. However, some of these differences might partly reflect variations in test-taking motivation rather than in actual abilities. Existing approaches for mitigating the distorting effects of rapid guessing focus mainly on point estimates of abilities, although research questions often refer to latent variables. The present study seeks to (a) determine the bias introduced by rapid guessing in group comparisons based on plausible value estimates and (b) introduce and evaluate different approaches of handling rapid guessing in the estimation of plausible values. In a simulation study, four models were compared: (1) a baseline model did not account for rapid guessing, (2) a person-level model incorporated rapid guessing as a respondent characteristic in the background model, (3) a response-level model filtered responses with item response times lower than a predetermined threshold, and (4) a combined model merged the person- and response-level approaches. Results show that the response-level and combined model performed best while accounting for rapid guessing on the person level did not suffice. An empirical example using data from a German large-scale assessment (N = 478) demonstrates the applicability of all approaches in practice. Recommendations for future research are given to improve ability estimation.

[1]Leibniz Institute for Educational Trajectories, Bamberg, Germany

*These authors contributed equally to this work.

Corresponding Author:
Jana Welling, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, Bamberg 96047, Germany.
Email: jana.welling@lifbi.de

Keywords

rapid guessing, plausible values, large-scale assessments, test-taking engagement, engagement gap, ability comparisons

Educational large-scale assessments (LSAs) are designed to evaluate domain-specific competencies and their differences across relevant groups such as countries, educational institutions, or gender. Prominent LSAs like the *Programme for International Student Assessment* (PISA), *Programme for the International Assessment of Adult Competencies* (PIAAC), and *Trends in International Mathematics and Science Study* (TIMSS) typically use plausible values (PVs) to account for measurement error and correct for sampling design in the estimation of abilities to provide unbiased estimates of population parameters (e.g., Mislevy, 1991; Wu, 2005). Thus, they enable valid inferences about populations and subgroups, rather than providing precise estimates of individual abilities, making them a cornerstone of LSAs.

LSAs are typically low stakes because test results do not have direct consequences for the participants. Consequently, some test-takers may lack sufficient motivation or exert minimal effort when completing these tests, leading to rapid guessing (RG)—a response behavior characterized by superficially quick item responses that do not reflect actual knowledge or abilities (Wise & DeMars, 2005). Because responses influenced by RG do not reflect the actual proficiency of the test-takers, they pose a threat to the validity of ability estimates (Wise, 2017). Simulation studies have shown that RG introduces substantial biases in point estimates of proficiency and has a nonnegligible impact across diverse testing contexts (e.g., DeMars et al., 2013; Osborne & Blanchard, 2011; Wise & DeMars, 2010). However, while PVs are commonly used in educational LSAs, the extent to which RG affects PV estimates has not been systematically explored. Given that PVs reflect uncertainty at the population level rather than for individual respondents, RG may primarily affect analyses of aggregated data, for example, on differences between groups.

This study therefore aimed to investigate the effect of RG on PV estimates and to evaluate methods for addressing it in the context of ability comparisons between groups (hereinafter referred to as *group comparisons*). Specifically, we seek to (a) determine the bias introduced by RG in group comparisons based on PV estimates, and (b) introduce and evaluate different approaches of handling RG in the estimation of PVs. To address these research aims, we first present different methods of accounting for RG in PV estimation. In a simulation study and an illustrative application using data from university students participating in a German LSA, we determine to what extent RG biases group comparisons based on PVs and evaluate the effectiveness of the different approaches in mitigating these biases.

Rapid Guessing in Large-Scale Assessments

Defining Rapid Guessing

Schnipke and Scrams (1997) classify test-taking behavior into two distinct categories: solution behavior (SB) and RG behavior. SB involves carefully analyzing a question while employing one's abilities and effort after thoroughly reading the item. In contrast, RG behavior is characterized by quickly skimming the question and selecting an answer without meaningful engagement. As a result, responses produced with RG do not accurately reflect test-taker's true abilities in the same way as those given under SB (Wise, 2017).

However, many existing IRT models are designed for SB by assuming that the probability of a correct response follows a monotonically increasing function depending on the respondent's proficiency, meaning that higher proficiency levels correspond to higher probabilities of correct responses (C. Wang & Xu, 2015). For example, under the one-parametric logistic test model (Rasch, 1960), the probability of a correct response is given as

$$P(Y_{ij} = 1 | \theta_i, \xi_j) = \frac{\exp(\theta_i - \xi_j)}{1 + \exp(\theta_i - \xi_j)} \quad (1)$$

with the binary response Y_{ij} given as a function of the ability parameter θ_i of person i and the difficulty parameter ξ_j of item j. In contrast, under RG, the assumptions of the Rasch model do not hold as test-takers respond without considering the response options, relying purely on guessing. The observed responses are assumed to be independent of the test-taker's true ability and thus provide no meaningful information about it (Schnipke & Scrams, 1997; C. Wang & Xu, 2015).

Reported RG rates in LSAs vary widely, ranging from under 1% (Kroehne et al., 2020) to nearly 20% (Ulitzsch et al., 2020), depending on factors such as test domain, testing conditions, and the identification method used (e.g., Goldhammer et al., 2016; Kroehne et al., 2020; Michaelides & Ivanova, 2022; Rios et al., 2017). Because RG occurs at the response level (Ulitzsch et al., 2020), it is influenced by both test-taker characteristics (e.g., ability; Rios & Soland, 2022) and item characteristics (e.g., placement, difficulty, and formatting; Lee & Jia, 2014). Typically, test-takers with higher ability levels exhibit lower rates of RG, and more difficult items tend to lead to higher RG rates (e.g., Goldhammer, Martens, & Lüdtke, 2017; Rios et al., 2022; Soland, 2018), highlighting the complex relationship between test-takers' ability, RG, and item characteristics.

Biases Introduced by Rapid Guessing

Failing to account for RG biases both item and person parameter estimates in models that assume SB (Jin et al., 2022; Rios et al., 2017; Rios & Soland, 2021a). The extent of this bias increases with higher RG rates (Rios & Soland, 2021a; Wise et al., 2021) and a greater difference between the probability of a correct response when engaging

in SB and the probability of a correct guess when engaging in RG (which is mostly notable in easy tests, where the probability of a correct response for SB is generally high; Rios et al., 2017).

In addition, RG can distort ability comparisons between groups (Soland, 2018). For example, Anaya and Zamarro (2023) found that boys exhibit higher RG rates than girls in PISA assessments, independent of their ability level. Accounting for these differences substantially altered estimates of the gender achievement gap, shifting mathematics and science scores by up to 36% and 40% of a standard deviation, respectively, in favor of boys. These findings underscore the substantial threat RG poses to the validity of scores in general (e.g., DeMars, 2007; Setzer et al., 2013; Wise et al., 2009) and to group comparisons in particular (Anaya & Zamarro, 2023; Soland, 2018; Wise, 2017). They highlight the importance of robust methods to identify and adjust for RG when interpreting data from LSAs to ensure accurate interpretation of results.

Measuring Rapid Guessing

RG can be measured at multiple analytical levels. At the examinee level, it is measured across all responses for each test-taker to assess individual differences in RG behavior (Wise & Kong, 2005). At the item level, RG patterns are aggregated across all test-takers for a given item because some items may provoke higher rates of RG than others (Schnipke & Scrams, 1997). Finally, at the response level, each interaction between a test-taker and an item is analyzed to identify specific responses from a test-taker as a rapid guess.

Although RG behavior cannot be directly observed, three primary proxies have been proposed for its measurement: (a) self-reported effort, (b) aberrant response patterns, and (c) response times (Rios & Deng, 2021). Of these, response time-based methods have gained particular popularity in recent years (e.g., Goldhammer, Naumann, et al., 2017; Silm et al., 2020; Wise & Kuhfeld, 2021) due to their effectiveness in classifying responses below a specified threshold as RG. By leveraging process data automatically collected in computerized assessments, this approach minimizes concerns about observer effects (Rios & Deng, 2021). Furthermore, it enables the evaluation of RG on the response level, facilitating detailed analyses of shifts in test-taking behavior across items (Wise & Kingsbury, 2016).

Methods of detecting and accounting for RG using response times can be broadly categorized into two groups: model-based and threshold-based approaches. Model-based approaches distinguish between SB and RG using mixture modeling techniques (Nagy & Ulitzsch, 2022; C. Wang & Xu, 2015). In these mixture models, the probability of a correct response depends on the response behavior. For responses identified as SB, the probability of a correct response is modeled using traditional IRT models, for example, the Rasch model as depicted in Equation 1. For responses identified as RG, however, the probability is set to chance level, assuming that these

responses do not provide meaningful information about a test-takers' ability (Schnipke & Scrams, 1997; C. Wang & Xu, 2015).

Research shows that model-based approaches effectively reduce RG bias by identifying and accounting for responses classified as RG (Rios et al., 2022). However, their practical application is limited by high computational demands, large sample size requirements, challenges with model convergence, and assumptions about response time distributions (Molenaar et al., 2018). Therefore, in practice, threshold-based approaches are most commonly used (Rios & Deng, 2021; Silm et al., 2020). These assume distinct response time distributions for SB and RG. By comparing an individual's item response time t_{ij} against a predefined threshold ρ_j of item j, the engagement ϕ_{ij} for each item-person encounter is determined (Goldhammer, Martens, & Lüdtke, 2017). If the response time is shorter than the threshold, the response is flagged as RG, meaning

$$\phi_{ij} \in \begin{cases} 0, & if \ t_{ij} < \rho_j \\ 1, & if \ t_{ij} \geq \rho_{j.} \end{cases} \tag{2}$$

Therefore, the engagement ϕ_{ij} takes the value 0 if the response is flagged as a rapid guess and 1 otherwise. However, this requires an accurate threshold determination, as inaccurate thresholds may also bias item and parameter estimates (Rios, 2022).

Selection of a Threshold

Threshold-based approaches of accounting for RG require the selection of a response time threshold to classify responses as either RG or SB based on the assumption that the two behaviors correspond to distinct, nonoverlapping response time distributions (Rios & Deng, 2023). Thresholds can be determined through heuristic rules, examination of response time distributions, or the combination of response time with accuracy information (Rios & Deng, 2021). In the latter case, unusually short responses that are also incorrect are taken as strong indicators of RG, whereas longer response times accompanied by higher accuracy are more consistent with solution behavior (Rios & Deng, 2021). The following paragraphs introduce two well-established and easily applicable threshold techniques (Buchholz et al., 2022; Goldhammer et al., 2016; Kroehne et al., 2020; Michaelides & Ivanova, 2022): the common-*k* and the visual-inspection method.

The common-*k* or fixed threshold method applies a uniform response time threshold across all items (Wise et al., 2004) and is widely used due to its simplicity, including in LSAs like PIAAC (Goldhammer et al., 2016). Its main advantage is that it does not require item-specific data, making it suitable for large item pools with minimal effort. However, a key limitation is its failure to account for item variability. Applying a uniform threshold to items with, for example, different reading demands may overlook meaningful differences in response behavior. Although some research suggested that response time distributions remain nearly uniform under RG due to

disengagement (C. Wang & Xu, 2015), the response time distributions under SB can vary between items and populations. This method may thus fail to account for heterogeneity in items (e.g., item difficulties) and population characteristics (e.g., person abilities), potentially undermining its validity (Wise, 2017).

The visual-inspection method proposed by Schnipke (1995) identifies response time thresholds based on the bimodal distribution of response times, where the first mode represents RG and the second reflects SB. The threshold is set at the lowest intersection point between these distributions. While this method is effective, it is also time-consuming and may lead to different interpretations among observers (Rios & Deng, 2021). The assumption that response times follow a bimodal distribution for every item does not always hold in practice, particularly for items that can be solved quickly even under SB (Wise, 2017). In addition, this method has the tendency to set excessively high thresholds, which may lead to misclassification of engaged responses as RG (e.g., Wise & Kuhfeld, 2020, 2021).

Plausible Value Technique

PVs account for measurement error in the estimation of competencies while also correcting for sampling design to allow for unbiased population estimates (e.g., Mislevy, 1991; Wu, 2005). They were developed in the context of LSAs to enable valid inferences about populations and subgroups rather than to allow a precise measurement of an individual's ability (e.g., von Davier et al., 2009; Wu, 2005). PVs are generated by drawing multiple estimates of an individual's latent ability, thereby reflecting the range of abilities that are plausible for that person given their item responses (Wu, 2005). The PVs approach applies the concept of multiple imputation (Rubin, 1987; see also Jewsbury et al., 2024), wherein multiple values for each individual i are drawn from the posterior predictive distribution of their latent ability, θ_i (Mislevy, 1991). The posterior predictive distribution can be described as

$$P(\theta_i|Y_i, X_i; \xi, \psi) \propto P(Y_i|\theta_i; \xi) \cdot P(\theta_i|X_i; \psi), \tag{3}$$

with Y_i representing the participant's item responses, ξ the parameters of the measurement model, and X_i the background variables. The posterior predictive distribution integrates two key components: the measurement model $P(Y_i|\theta_i; \xi)$, which establishes the relationship between the latent ability θ_i and the item responses Y_i, and the background model $P(\theta_i|X_i; \psi)$, which describes how background variables X_i relate to the ability θ_i (Mislevy, 1991; Mislevy et al., 1992). The measurement model is typically defined using an IRT model, which specifies the probability of a response pattern $Y_i \in \{y_{i1}, \ldots, y_{iJ}\}$ for a person $i \in \{1, \ldots, I\}$ to a set of items $j \in \{1, \ldots, J\}$ as follows,

$$P(Y_i|\theta_i; \xi) = \prod_{j=1}^{J} P(y_{ij}|\theta_i; \xi_j). \tag{4}$$

A commonly used IRT model is the Rasch model (Equation 1). The background model is typically parameterized as a linear regression as shown in Equation (4), where ψ_0 denotes the intercept, and ψ_X represents a vector of regression coefficients for the predictors included in the background model

$$P(\theta_i|X_i;\psi) \propto N\left(\psi_0 + \psi_X X_i; \sigma^2_{\theta_i|X_i}\right) \tag{5}$$

An important challenge in PV estimation is ensuring that all relevant variables X_i are included in the background model in their correct functional form. Omitting key variables can lead to biased estimates and compromise the accuracy of PVs (Bondarenko & Raghunathan, 2016; Meng, 1994). Therefore, it is recommended to include all variables in the background model that are used in subsequent analyses to provide unbiased population estimates (von Davier et al., 2009; Wu, 2005). When the background model is correctly specified, Sengewald and Mayer (2024) showed that estimates obtained by PVs are comparable to those from latent variable models.

Accounting for Rapid Guessing in Plausible Value Estimation

While several approaches are available to account for RG in analyses of point estimates, the optimal strategy for handling RG in the context of PVs remains underexplored. Because relevant person-level variables can be incorporated into the background model, PVs offer a unique framework that can be leveraged to account for RG in ability estimation. The three threshold-based approaches of accounting for RG in PV estimation presented below aim to mitigate biases induced by RG by assuming that RG manifests at the response level. However, each of these models addresses RG at a distinct analytical level: either at the person level, at the response level, or both.

Rapid Guessing as a Person-Level Characteristic

The *person-level model* conceptualizes RG as an individual characteristic measurable at the examinee level. This model builds upon the principle of PV estimation by incorporating a proxy for RG into the background model, facilitating adjustments for RG and its correlation with the estimated ability at the person level. RG at the person level can be operationalized using the response time effort (RTE; Wise & Kong, 2005), which describes an individual's average engagement across the test. The idea is that each test consists of multiple person-item encounters. For each encounter, the test-taker decides whether to engage in SB or RG, and this decision is reflected in the response time. The overall RTE of individual *i* is calculated as

$$RTE_i = \frac{\sum_{j=1}^{J} \phi_{ij}}{J}, \tag{6}$$

with J describing the total number of items in the test and ϕ_{ij} being the engagement indicator given in Equation 2. As the RTE is usually heavily skewed, it is common practice (e.g., Wise et al., 2021) to further dichotomize the RTE. Wise and Gao (2017) argue that, based on previous research, it can be assumed that ability estimates are considerably distorted when RG exceeds 10% of all provided responses. Therefore, test-takers can be defined as being engaged when they use SB on more than 90% of the items:

$$RTE_i^* = \begin{cases} 0, & if\ RTE_i \leq .9 \\ 1, & if\ RTE_i > .9. \end{cases} \tag{7}$$

Buchholz and colleagues (2022) described the inclusion of a person-level RG operationalization as the simplest method for addressing this aberrant test-taking behavior in PV estimation. However, to our knowledge, no systematic analyses have been conducted to evaluate whether this approach can effectively attenuate biases induced by RG.

In summary, the person-level model accounts for RG in PV estimation by including the RTE as a proxy of RG in the background model. Thus, PVs are generated conditional on examinees' overall test-taking effort, thereby adjusting population estimates for systematic differences in engagement.

Rapid Guessing as a Response-Level Characteristic

The measurement model in the *response-level model* corresponds to the effort-moderated model (EMM) by Wise and DeMars (2006). The EMM assumes that, for each person-item encounter, examinees decide whether to engage in SB or RG. Under SB, the probability of a correct response increases with ability and can be modeled using a standard IRT model. Under RG, by contrast, the probability of a correct response remains at the chance level, independent of ability. Thus, the functional relationship between ability and accuracy fundamentally differs depending on the response process. The EMM combines the two submodels representing SB and RG as

$$P_{ij}(\theta_i) = \left(\phi_{ij}\right)(SB\ model) + \left(1 - \phi_{ij}\right)(RG\,model) \tag{8}$$

Depending on the engagement of examinee i for item j, which is determined by comparing the item response time to a threshold, either one of the two submodels is used to model the examinee's response. Because the model assumes that responses given under RG are not psychometrically informative, rapid guesses are coded as missing in the IRT model, thereby reducing parameter bias compared with scoring them as incorrect (Rios & Deng, 2023). EMM scoring effectively handles RG but assumes RG occurs at random; if this assumption is violated, the accuracy of parameter estimates will be affected (DeMars, 2024; Rios & Soland, 2021a). Moreover, as RG rates increase, the standard errors of ability estimates also increase when using EMM

scoring (Rios & Soland, 2021b). Despite these limitations, the EMM remains a practical and efficient method for handling RG (Liu et al., 2019; Rios & Deng, 2023).

In contrast to the person-level approach, the response-level model integrates RG directly into the measurement model. By coding rapid guesses as missing within the IRT framework, PVs are estimated conditional on effort at the response level, which allows for a fine-grained adjustment that accounts for response-level variation in engagement.

Rapid Guessing as Person- and Response-Level Characteristic

The *combined model* addresses RG at both the person and response levels by combining the two previous approaches. It incorporates RG at the person level as a covariate into the background model for PV estimation while scoring the rapid guesses at the response level based on the EMM. Therefore, this approach utilizes the EMM as an established RG scoring method while also accounting for the correlation between RG and ability in PV estimation. Consequently, it combines the strengths of both the person-level and response-level models.

Aims of the Study

The study aimed to investigate the impact of RG on group comparisons based on PVs and to evaluate different methods of accounting for RG in PV estimation. While PVs are widely used to adjust for measurement error and estimate population-level effects, the influence of RG on PVs has not been systematically explored. Therefore, this study seeks to (a) determine the bias introduced by RG in group comparisons based on PV estimates and (b) introduce and evaluate different approaches of handling RG in the estimation of PVs. To address these research aims, we present a simulation study that evaluates the performance of three approaches designed to account for RG in PV estimation on different analytical levels and compares them with an approach that does not account for RG. Furthermore, we demonstrate the application of these approaches using data from university students in a German LSA, providing insights into how these methods can improve the validity of group comparisons in LSAs.

Simulation Study

Design

The description of the simulation design follows the ADEMP (aim, data generation, estimands, methods, performance measures) structure suggested by Morris et al. (2019). The simulation was conducted in *R* (R Core Team, 2024). The analysis code, results, and supplemental material are provided at https://osf.io/gdnse/.

Aims. The simulation study aimed to evaluate the parameter recovery of group comparisons for different approaches of (not) accounting for RG in PV estimation.

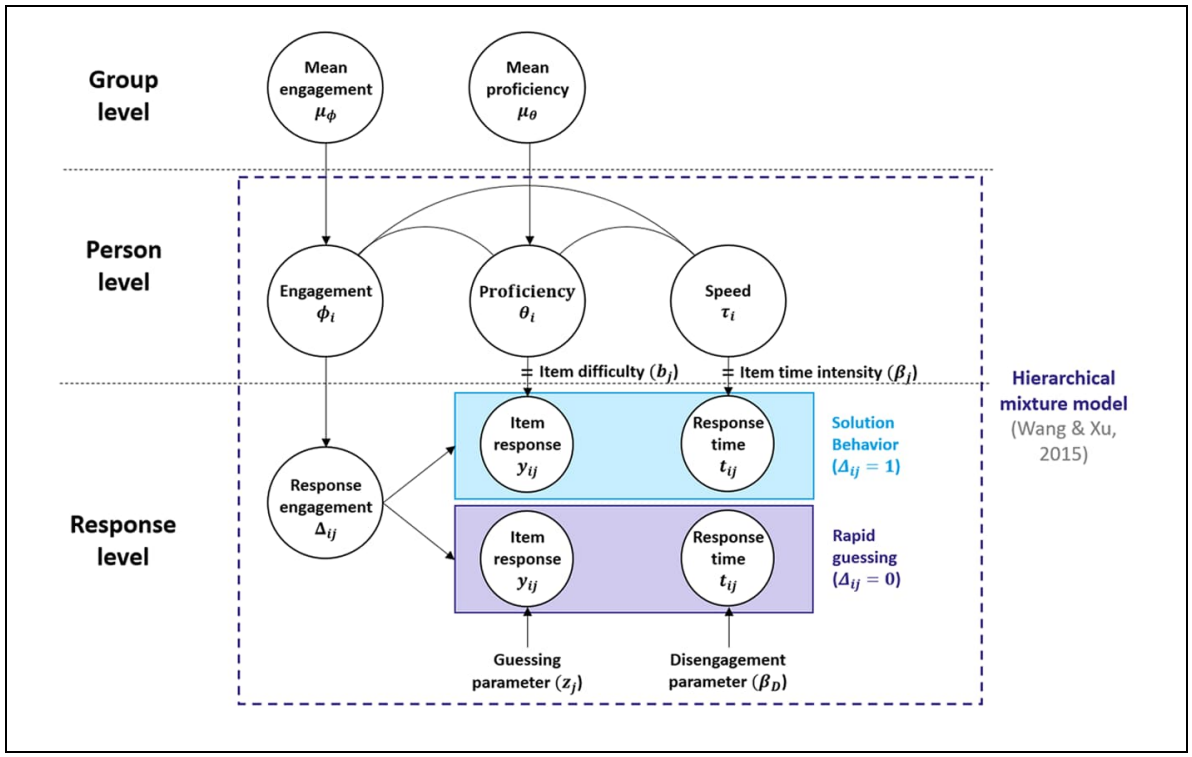


Figure 1. Hierarchical Mixture Model Used for Data Generation.

Note. The figure depicts the hierarchical mixture model used for data generation in the simulation study. The model is adapted from C. Wang and Xu (2015; marked by dashed rectangle), who proposed a mixture modeling approach accounting for rapid guessing on the response level while still incorporating person-level characteristics and their relationships. It distinguishes between two latent response behavior classes: rapid guessing and solution behavior. Each class is characterized by different distributions of correct item responses and item response times. The current adaptation additionally incorporates a group level with aggregated engagement and proficiency values. The figure is adapted from Nagy and Ulitzsch (2022).

Data Generation. Data were generated based on the mixture hierarchical model by C. Wang and Xu (2015), as this model allows for individual differences not only in proficiency but also in test-taking engagement and speed while accounting for the relationship between these variables. The model assumes that test-takers decide for each item to engage either in SB or RG, represented by two latent classes on the response level. The probability of a correct response and the item response times each follow a mixed distribution, depending on the latent class. To accommodate the group comparison setting, the model was expanded in the current study at a group level (see Figure 1). The parameters were adapted from a previous simulation study on RG (Ulitzsch et al., 2020) to mirror empirical data of reading comprehension tests used in educational LSAs (e.g., Goldhammer et al., 2014; Welling et al., 2024).

Two groups were generated, with each group g consisting of $N_g = 1{,}000$ participants. Within each group, the test-taking engagement ϕ_i, ability θ_i, and test-taking-speed τ_i of person i originated from a multivariate normal distribution with a mean vector $\mu = (\mu_{\phi_g}, \mu_{\theta_g}, \mu_\tau)$. Both μ_{θ_g} and μ_{ϕ_g} differed between the groups and were

Table 1. Parameters Across Simulation Conditions.

Condition	Factor			Value			
	Ability difference	RGR overall	RGR difference	μ_{θ_1}	μ_{θ_2}	RGR in group 1[a]	RGR in group 2[a]
1	0.50	5%	–5%	–0.25	0.25	7.5%	2.5%
2	0.50	5%	0%	–0.25	0.25	5.0%	5.0%
3	0.50	5%	5%	–0.25	0.25	2.5%	7.5%
4	0.50	15%	–5%	–0.25	0.25	17.5%	12.5%
5	0.50	15%	0%	–0.25	0.25	15.0%	15.0%
6	0.50	15%	5%	–0.25	0.25	12.5%	17.5%
7	1.00	5%	–5%	–0.50	0.50	7.5%	2.5%
8	1.00	5%	0%	–0.50	0.50	5.0%	5.0%
9	1.00	5%	5%	–0.50	0.50	2.5%	7.5%
10	1.00	15%	–5%	–0.50	0.50	17.5%	12.5%
11	1.00	15%	0%	–0.50	0.50	15.0%	15.0%
12	1.00	15%	5%	–0.50	0.50	12.5%	17.5%

Note. RGR = rapid guessing rate on response level, μ_{θ_1}= mean ability in Group 1, μ_{θ_2}= mean ability in Group 2.
[a]The RGRs were converted to the group mean engagement parameter μ_{ϕ_g} as $1 - RGR_g = \frac{\exp(\mu_{\phi_g})}{1+\exp(\mu_{\phi_g})}$.

systematically varied between simulation conditions (see Table 1). In contrast, μ_τ was zero for both groups across all conditions. The variance-covariance matrix for ϕ_i, θ_i, and τ_i was invariant across groups and defined as

$$\Sigma = \begin{pmatrix} 1 & & \\ .70 & 3.5 & \\ -.20 & -.20 & 0.05 \end{pmatrix} \tag{9}$$

The competence test was simulated to comprise 20 items. In the SB class E, the probability of a correct response followed a Rasch model (Rasch, 1960), thus depending on θ_i and the item difficulty ξ_j of item j, with $\xi_j \in \{-1.0, -0.5, 0.0, 0.5, 1.0\}$. The log-transformed item response times followed a normal distribution with standard deviation $\sigma_E = 0.30$, mean $\beta_j - \tau_i$, and item time intensities $\beta_j \in \{2.9, 3.1, 3.3, 3.5, 3.7\}$ corresponding to average item response times of 15 to 40 seconds. The five values of the item parameters ξ_j and β_j were each replicated four times to simulate the 20 items included in the test. In the RG class D, the probability of a correct response was defined for each item as $z_j = 0.25$, which resembles the guessing probability in multiple-choice questions with four response options. The log-transformed item response times followed a normal distribution with mean $\beta_D = 1.25$ ($\approx$3.5 seconds) and standard deviation $\sigma_D = 0.60$.

To evaluate the performance of the models in diverse conditions, three factors were systematically varied. First, the overall RG rate on the response level was varied between a lower (5%) and a higher (15%) rate. RG rates reported in the literature vary

from under 1% to nearly 20%, depending on factors such as sample, test domain, and item or test characteristics (e.g., Michaelides & Ivanova, 2022; Ulitzsch et al., 2020). The RG rates in the present study were chosen to cover different scenarios while still being plausible for applied settings. Second, the size of the ability difference between the two groups was varied to correspond to a medium (0.5 *SD*) and a large (1.0 *SD*) gap. Similar values have been repeatedly found in educational studies (e.g., Bloom et al., 2008). Third, the difference in the RG rates between the two groups was varied, with (a) Group 1 having a lower RG rate than Group 2 ($\Delta_{\%} = -5$), (b) both groups having the same RG rate ($\Delta_{\%} = 0$), and (c) Group 1 having a higher RG rate than Group 2 ($\Delta_{\%} = 5$). Since the impact of RG depends both on its frequency and the ability level of the (sub)sample, different RG rates in the two groups can bias the estimated aggregated abilities to varying degrees and thus also affect the estimation of ability differences. In total, the study comprised 2 (overall RG rate) × 2 (ability difference) × 3 (difference in RG rates) = 12 conditions. The specific parameters for each condition (i.e., $\mu_{\theta_g}, \mu_{\phi_g}$) are displayed in Table 1.

Estimands. The main estimand is the ability difference between the two groups, measured by Cohen's *d* (Goulet-Pelletier & Cousineau, 2018). To enhance the interpretation of the results, the mean ability μ_{θ_g} and standard deviation of ability σ_{θ_g} of each group are included as supportive estimands.

Methods. For each simulated dataset, we estimated (a) the baseline model ignoring RG as well as (b) the person-level model, (c) the response-level model, and (d) the combined model that accounted for RG on the person level, response level, or both, respectively. As different approaches of accounting for RG vary in their sensitivity to distinct levels of misclassifications (Rios & Deng, 2023), models (b) to (d) were evaluated using two different methods to set the response-time thresholds: (a) the common-*k* method with a fixed threshold of 5 seconds for all items (*fixed threshold*) and (b) the visual-inspection method with an individual threshold for each item (*visual threshold*). Because it was not possible to visually determine a threshold for 20 items in all 12,000 simulated datasets, for each item parameter set (i.e., same item difficulty and time intensity) in each condition, one common threshold was determined by computing the minimum of the response time distribution between 5 and 15 seconds. The resulting 60 thresholds ranged between 5.54 and 12.86 seconds and were visually verified. The RTE was calculated and dichotomized using Equations 6 and 7, respectively.

For each dataset and method, the model with corresponding PVs was estimated in three steps using the package *TAM* (Robitzsch et al., 2024). First, a Rasch model (see Equation 1) was computed to obtain unbiased item difficulties. Then, the full model including the fixed item difficulties from step one and all covariates was estimated. As it is recommended to include all variables in the background model that are used in subsequent analyses (von Davier et al., 2009; Wu, 2005), the grouping variable was defined as a covariate for all models. In the person-level and combined models,

the dichotomized RTE was included as an additional covariate to account for RG on the person level. Finally, 25 sets of PVs were drawn from the posterior distribution of the full model. Although the number of PVs provided in some educational LSAs is limited to 5 or 10, the use of more PVs can result in more efficient parameter estimates (see Laukaityte & Wiberg, 2017) and is common in LSAs such as the *National Assessment of Educational Progress*.

Performance Measures. As main performance measures, the rate of converged models, the absolute bias (as the relative bias is dependent on the true value and thus differs between conditions), and the empirical standard error were computed. In addition, the mean squared error (MSE) and the coverage are reported in the supplemental material. For each performance measure except the convergence criteria, the Monte Carlo standard error (MCSE) was computed (see Morris et al., 2019, p. 2086, for an overview of the performance measures with equations). The standard error of Cohen's d was calculated using Equation 4 in Goulet-Pelletier and Cousineau (2018). All parameters were pooled over all PV sets using Rubin's rules (Rubin, 1987). To obtain precise performance measures with small MCSEs, we chose a large number of simulated datasets, $N_{sim} = 1{,}000$.

Results

Threshold Diagnostics. There was a high agreement between the engagement estimated by the thresholds and the true engagement ($> 93\%$ overlap for both thresholds and both RTEs in all conditions, see Tables S1 and S2 in the Supplemental Material). The visual threshold and the corresponding RTE exhibited a slightly better overlap than the fixed threshold and the corresponding RTE, especially in conditions with a high RG rate. The sensitivity (i.e., correctly classified engaged responses) was extremely high for both thresholds and RTEs ($> 98\%$), but the specificity (i.e., correctly classified rapid guesses) was higher for the visual threshold and corresponding RTE ($> 89\%$) than the fixed threshold and corresponding RTE ($> 72\%$) throughout all conditions.

Performance Measures. All models converged in all conditions and simulated datasets. For all performance measures, estimands, models and conditions, the MCSE was estimated to be below 0.01 (see Tables 2 and 3 as well as Tables S3–S20 in the Supplemental Material). Overall, the bias in the estimated ability difference was rather low for most conditions and models (see Figure 2 and Table 2). The baseline model and the two person-level models tended to underestimate the ability difference, especially in the conditions with a high RG rate, a large true ability difference, and/or a higher RG rate in the high ability group than in the low ability group ($bias \geq -0.16$). The biases of the response-level and combined model were very small ($bias \leq |0.04|$), independent of the threshold and in all conditions and models. The most pronounced bias occurred when the RG rate was high and the true ability difference large, with the models based on the fixed threshold slightly

Table 2. Absolute Bias of the Estimated Ability Difference With Monte Carlo Standard Errors.

		Person-level model		Response-level model		Combined model	
Con.	Baseline model	Fixed	Visual	Fixed	Visual	Fixed	Visual
1	0.012	0.013	0.013	–0.001	–0.004	0.004	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
2	–0.015	–0.016	–0.015	–0.001	0.002	–0.005	–0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
3	–0.039	–0.041	–0.040	0.002	0.010	–0.010	–0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
4	–0.014	–0.013	–0.014	–0.006	–0.002	–0.005	–0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
5	–0.052	–0.050	–0.049	–0.008	0.012	–0.010	0.008
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
6	–0.093	–0.090	–0.090	–0.015	0.018	–0.020	0.011
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
7	0.001	0.002	0.002	–0.002	–0.003	–0.001	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
8	–0.025	–0.025	–0.024	0.004	0.011	–0.003	0.003
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
9	–0.059	–0.059	–0.059	0.002	0.014	–0.014	–0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
10	–0.069	–0.065	–0.065	–0.017	0.004	–0.018	0.003
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
11	–0.112	–0.108	–0.107	–0.022	0.017	–0.027	0.011
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
12	–0.154	–0.149	–0.148	–0.027	0.029	–0.034	0.019
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Note. Con. = simulation condition. The terms "fixed" and "visual" refer to the threshold method used to flag rapid guesses. Displayed is the absolute bias and its Monte Carlo standard error in parentheses as described in Morris and colleagues (2019).

underestimating ($bias \geq -0.04$), and the models based on the visual threshold slightly overestimate the ability difference ($bias \leq 0.03$). The empirical standard error of Cohen's d ranged between 0.05 and 0.06 for all models and conditions (see Table 3).

The bias in the supportive estimands is displayed in Supplemental Figures S1 to S4 and Tables S3 to S6 in the supplemental material. The response-level models slightly overestimated the mean ability in the low ability group (μ_{θ_1}) when the RG rate was high ($bias \leq 0.06$), while the baseline, person-level, and combined models barely showed any bias ($|bias| \leq 0.04$). The baseline and person-level models underestimated the mean competence in the high ability group (μ_{θ_2}), especially when the RG rate was high, the true ability difference large, and/or when the high ability group exhibited a higher RG rate than the low ability group ($bias \geq -0.20$). The response-

Table 3. Empirical Standard Error of the Estimated Ability Difference With Monte Carlo Standard Errors.

		Person-level model		Response-level model		Combined model	
Con.	Baseline model	Fixed	Visual	Fixed	Visual	Fixed	Visual
1	0.053	0.052	0.052	0.052	0.052	0.052	0.052
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
2	0.055	0.055	0.055	0.055	0.055	0.055	0.055
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
3	0.055	0.055	0.055	0.056	0.056	0.055	0.056
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
4	0.053	0.052	0.053	0.053	0.054	0.053	0.054
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
5	0.052	0.053	0.053	0.053	0.053	0.054	0.053
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
6	0.052	0.052	0.052	0.053	0.054	0.053	0.054
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
7	0.058	0.058	0.058	0.057	0.057	0.057	0.057
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
8	0.054	0.054	0.054	0.054	0.054	0.054	0.054
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
9	0.053	0.053	0.053	0.054	0.054	0.054	0.054
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
10	0.052	0.052	0.052	0.053	0.054	0.053	0.053
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
11	0.053	0.054	0.054	0.056	0.057	0.056	0.056
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
12	0.052	0.052	0.053	0.055	0.057	0.055	0.057
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Note. Con. = simulation condition. The terms "fixed" and "visual" refer to the threshold method used to flag rapid guesses. Displayed is the empirical standard error and its Monte Carlo standard error in parentheses as described in Morris and colleagues (2019).

level and combined models performed better, with practically no bias in the condition with low RG rate. When the RG rate was high, the models based on the fixed threshold slightly underestimated μ_{θ_2}, while the response-level model based on the visual threshold slightly overestimated μ_{θ_2} ($|bias| \leq 0.06$). All models slightly underestimated both standard deviations when the overall RG rate was high, most notably the baseline and person-level models and when the estimand was the standard deviation of the low ability group ($bias \geq -0.12$). The empirical standard error of μ_{θ_1} and μ_{θ_2} ranged between 0.03 and 0.04, and the empirical standard error of σ_{θ_1} and σ_{θ_2} between 0.02 and 0.03 (see Tables S7 to S10 in the Supplemental Material). The MSE and the coverage for all five estimands are reported in the Supplemental Material.

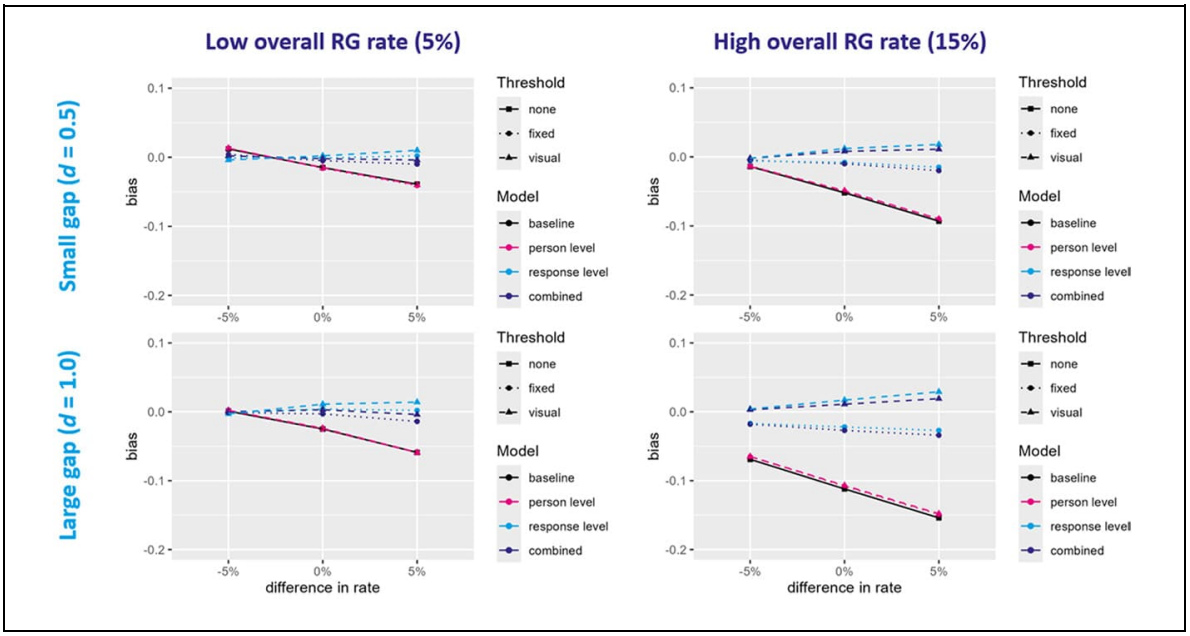


Figure 2. Absolute Bias in Estimated Ability Difference.

Note. RG = rapid guessing, gap = true ability difference. The figure depicts the absolute bias in the estimated ability difference for (1) the baseline model that does not account for rapid guessing and (2) three proposed approaches of accounting for rapid guessing (person-level model, response-level model, combined model) using two threshold methods (fixed, visual). The 12 simulation conditions vary by rapid guessing rate, size of the true ability difference, and difference in rapid guessing rate.

Overall, the models that accounted for RG on the response level (i.e., response-level and combined model) performed best, while only accounting for RG on the person level did not improve the performance compared to the baseline model. The differences between the models were most pronounced when the overall RG rate was high, the true ability difference was large, and/or the high ability group engaged in more RG than the low ability group. Especially in these conditions, the performance of the response-level and combined model also depended on the threshold method.

Empirical Application

To illustrate the different models and investigate differences in conclusions that may be observed in practice, the same models as in the simulation study were used to compare groups of (former) university students in different subject areas of a German LSA.

Methods

Sample. The sample consisted of 478 former university students who participated in the German *National Educational Panel Study* (NEPS; Blossfeld & Roßbach, 2019) that follows different cohorts across their life course. The participants (43% female) gave written informed consent and were currently or have been previously enrolled

in different study subjects. For the group comparison, the sample was divided into two groups covering different study fields with distinct mean reading ability levels. Group 1 consisted of 302 participants (27% female) who were (formerly) enrolled in sports or engineering (age: $M = 27.99$, $SD = 3.04$), and Group 2 consisted of 176 participants (70% female) who were (formerly) enrolled in (veterinary) medicine or arts (age: $M = 28.15$, $SD = 4.26$).

Instrument. The participants took part in an unproctored, web-based test measuring reading comprehension. The test was specifically developed for the NEPS (Gehrer et al., 2013) and comprised 21 items with different response formats distributed over five reading tasks. However, to be comparable with the simulation study, only the 14 multiple-choice items were used in the present analysis. Each of these items contained four response options with one correct solution each (i.e., 25% chance level of a correct response).

Analysis. The response time data were preprocessed with R code developed by Welling et al. (2024). Item response time was defined as the time the participant spent in total on the respective item page (if the participant visited the page more than once, the different visit times were summed). As in the simulation study, a fixed and a visual threshold were set for each item. Because the computational approach for determining the visual threshold (see design section of the simulation study) could not identify clear thresholds for most items, the histograms of these item response times were visually scanned to find the intersection of the bimodal distribution. In case of ambiguous distributions, we followed Rios' (2022) suggestion and preferred a liberal threshold. Consequently, the visual thresholds varied between 5 and 10 seconds (see Table 4).

Before model estimation, measurement invariance between the two groups was examined by testing for differential item functioning (DIF) for each dataset (i.e., the original data and for each threshold approach one dataset with missing values for all responses flagged as rapid guesses). To evaluate DIF, we examined whether item difficulties differed significantly between the two groups. In addition, we compared the DIF model to a model without DIF using information criteria.

The background models for the PV estimation were chosen to resemble the simulation study and included only the grouping variable in all models and the dichotomized RTE in the person-level and combined model. As in the simulation study, seven models were computed: a baseline model that ignores RG, and six models accounting for RG by combining three modeling approaches (person-level, response-level, combined) with two threshold-setting methods (fixed threshold, visual threshold). The computation of the models, the drawing and pooling of PV sets, and the estimation of all five outcomes ($d, \mu_{\theta_1}, \mu_{\theta_2}, \sigma_{\theta_1}, \sigma_{\theta 2}$) followed the same procedure as in the simulation study. Analyses were conducted in *R* (R Core Team, 2024). The raw data are available at NEPS Network (2024), while the analysis code, analysis results, and supplemental material are provided at https://osf.io/gdnse/.

Table 4. Descriptives by Item in the Empirical Example.

			Valid responses	Response time	Fixed threshold		Visual threshold	
Item	Task	Position	*N*	*M* (*SD*)	Threshold	% RG	Threshold	% RG
1	1	1	477	29.29 (17.61)	5.00	0.2	5.00	0.2
2	1	2	477	35.53 (17.54)	5.00	0.8	7.00	1.1
3	1	3	477	26.93 (14.32)	5.00	1.5	6.00	2.1
4	2	6	466	27.00 (20.12)	5.00	1.7	8.00	2.4
5	2	7	466	22.29 (12.84)	5.00	1.9	7.00	2.2
6	3	10	463	27.71 (14.26)	5.00	2.4	9.00	3.7
7	3	11	463	21.22 (13.35)	5.00	3.0	6.00	3.2
8	3	12	460	31.06 (19.81)	5.00	3.3	6.00	3.3
9	3	13	460	34.62 (18.79)	5.00	3.5	10.00	3.9
10	3	15	450	27.14 (12.51)	5.00	3.6	8.00	4.0
11	4	16	424	32.13 (15.53)	5.00	3.8	9.02	5.4
12	4	18	396	32.93 (17.65)	5.00	5.6	9.12	6.6
13	5	19	358	22.97 (11.13)	5.00	4.5	6.79	5.3
14	5	20	348	18.36 (17.74)	5.00	4.9	5.00	4.9

Note. Response time and thresholds are specified in seconds.

Results

There was no DIF between the two groups (see Tables S21 and S22 in the supplement). In total, 7.5% of the responses were missing, resulting in 6,185 valid item responses provided by the participants (Group 1: 3,895 responses, Group 2: 2,290 responses).

In the whole sample, 2.8% of the responses were identified as rapid guesses and 5.7% of the participants were classified as disengaged (i.e., $\geq 10\%$ rapid guesses) when using the fixed threshold, compared to 3.3% rapid guesses and 6.7% disengaged participants when using the visual threshold. The RG rates differed between

the groups: in Group 1, 3.5% (fixed) or 4.0% (visual) of the responses were identified as rapid guesses and 6.3% (fixed) or 7.6% (visual) of the examinees classified as disengaged, while in Group 2, only 1.6% (fixed) or 2.2% (visual) of the responses were rapid guesses and 4.6% (fixed) or 5.1% (visual) of the participants were disengaged. In total, 32.0% (fixed) or 34.1% (visual) of the responses flagged as rapid guesses were solved correctly, thus exceeding the chance level of 25% only slightly.

Overall, the estimated parameters and their confidence intervals did not vary considerably between the different models (see Table 5). The baseline model that ignored RG estimated the ability difference between the groups to be $d = 0.66$, 95% *CI* [0.42, 0.90]. In all six models that accounted for RG, the ability difference was estimated to be slightly lower, ranging between $d = 0.59$, 95% *CI* [0.36, 0.82], in the person-level model using the fixed threshold and $d = 0.64$, 95% *CI* [0.40, 0.87], in the combined model using the visual threshold. The estimated mean ability of Group 1 (μ_{θ_1}) ranged between 0.10 in the person-level model and 0.16 in the response-level model, both models using the visual threshold. In contrast, the estimated mean ability of Group 2 (μ_{θ_2}) ranged from 0.54 for the person-level model to 0.60 for the response-level model, using the fixed and visual threshold, respectively. The standard deviation in Group 1 (σ_{θ_1}) was estimated to be 0.70 in the baseline and person-level models, and to be 0.73 in both combined models. The estimated standard deviation in Group 2 (σ_{θ_2}) ranged between 0.73 for the baseline model and 0.76 for the combined model using the fixed threshold.

The results of the empirical example mostly reflect the finding from the simulation study for the condition with a low overall RG rate, a higher RG rate for the low ability group, and a moderate ability difference. As all confidence intervals overlap, there seems to be no pronounced bias in the five outcomes when (not) accounting for RG, resulting in similar estimates for all examined models.

Discussion

Ability estimates can be seriously impaired if respondents do not invest motivation and effort in an administered competence test but engage in RG behavior (e.g., DeMars et al., 2013; Osborne & Blanchard, 2011; Wise & DeMars, 2010). Previous research, however, primarily focused on the consequences of RG on point estimates of ability, whereas its effect on PV has not yet been systematically explored. The present study filled this gap by determining the influence of RG on group comparisons based on PVs and by introducing and evaluating three different approaches of accounting for RG in PV estimation.

The results of the simulation study showed that the bias in the estimated ability difference varied across models, thresholds, and conditions, highlighting the importance of both model and threshold selection. When the RG rate was low, the corresponding bias in the estimated ability difference across all models was small. However, as the RG rates increased, the bias grew substantially. The models that ignored RG or accounted for RG only on the person level consistently

Table 5. Estimated Parameters in the Empirical Example.

No	Model	Threshold	Cohen's *d*	μ_{θ_1}	μ_{θ_2}	σ_{θ_1}	σ_{θ_2}
1	**Baseline**		0.66 [0.42, 0.90]	0.11 [0.01, 0.21]	0.58 [0.45, 0.71]	0.70 [0.63, 0.77]	0.73 [0.63, 0.84]
2	**Person-level**	fixed	0.59 [0.36, 0.82]	0.12 [0.02, 0.23]	0.54 [0.40, 0.69]	0.70 [0.62, 0.78]	0.74 [0.64, 0.83]
3		visual	0.63 [0.36, 0.91]	0.10 [–0.02, 0.23]	0.56 [0.42, 0.70]	0.70 [0.64, 0.77]	0.75 [0.65, 0.84]
4	**Response level**	fixed	0.60 [0.35, 0.84]	0.14 [0.02, 0.25]	0.57 [0.42, 0.72]	0.72 [0.64, 0.79]	0.75 [0.65, 0.85]
5		visual	0.60 [0.36, 0.84]	0.16 [0.05, 0.26]	0.60 [0.44, 0.75]	0.72 [0.65, 0.79]	0.76 [0.66, 0.85]
6	**Combined model**	fixed	0.63 [0.40, 0.86]	0.12 [0.01, 0.24]	0.59 [0.45, 0.73]	0.73 [0.65, 0.81]	0.77 [0.66, 0.87]
7		visual	0.64 [0.40, 0.87]	0.12 [0.02, 0.22]	0.59 [0.45, 0.73]	0.73 [0.65, 0.81]	0.75 [0.65, 0.85]

Note. μ_{θ_1} = estimated mean reading comprehension in Group 1, μ_{θ_2} = estimated mean reading comprehension in Group 2.
σ_{θ_1} = estimated standard deviation in Group 1, σ_{θ_2} = estimated standard deviation in Group 2.

underestimated the ability difference, particularly under a high RG rate, large true ability difference, and when the high ability group exhibited more RG than the low ability group. In contrast, models that accounted for RG on the response level, either alone or in combination with person-level adjustments, demonstrated substantially lower bias. Only when the RG rate was high, models based on a fixed threshold slightly underestimated the gap, while models using a visual threshold slightly overestimated it. In addition, the response-level models slightly overestimated the mean ability of the low ability group under a high RG rate, while the baseline, person-level, and combined models exhibited minimal bias. Conversely, the baseline and person-level models substantially underestimated the mean ability of the high ability group, particularly under a high RG rate and a large true ability difference. All models underestimated at least slightly the standard deviations when the RG rate was high, with the baseline and person-level model exhibiting the most pronounced bias. Across all outcomes and conditions, the combined model using the visual threshold (i.e., a low misclassification rate) performed best, although differences between the combined and response-level model were mostly marginal.

These findings indicate that ignoring RG in PV estimation can distort group comparisons of ability and may compromise the validity of inferences on educational inequalities. However, when the RG rate was small, the bias tended to be rather small. This aligns with previous research on point estimates of abilities, indicating that bias in group comparisons tends to be limited in many settings (Soland, 2018). A potential explanation for the small bias are group-level aggregation effects. As described by Wise et al. (2020), two mechanisms can reduce the impact of RG on aggregated scores. On the one hand, a diluting effect may occur if most respondents engage in SB, which reduces the impact of RG, especially when the RG rates are low. On the other hand, a cancellation effect can occur if individual score distortions vary in direction. As shown by these authors, some rapid guesses lead to negatively biased scores, while others result in positively biased scores (e.g., when the probability of a correct response is larger under RG than under SB), thereby offsetting each other in aggregated scores.

Furthermore, the findings of the simulation study suggest that the extent of the bias varies depending on the prevalence of RG in both groups and the size of the true ability difference. The results support key theoretical assumptions about bias caused by RG when not or inadequately accounted for. Bias increases (a) with higher RG rates, because a larger proportion of distorted responses is incorporated into the estimation process of PVs (Rios & Soland, 2021a), (b) in groups with higher competence since the probability of a correct response under SB and thus the difference between the probability of a correct response under SB vs. RG increases with proficiency (Rios et al., 2017), and (c) in group comparisons additionally when the more able group engages in more RG than the less able group, as a direct consequence of (a) and (b) (Anaya & Zamarro, 2023; Soland, 2018). The finding that the bias increased with the size of the true ability difference is attributable to the effect of (b): the larger the true

difference, the greater was the simulated ability of the high ability group and thus the distorting effect of RG.

The evaluation of the different approaches of accounting for RG in PV estimation showed that adjusting for RG solely on the person level by including a proxy for RG yielded little improvement over the model that ignored RG, whereas the models accounting for RG on the response level effectively minimized bias. These findings suggest that also in PV estimation, RG should be considered at the response level, and are thus consistent with previous research on point estimates (Wise & DeMars, 2006). Leveraging the unique framework that offers PVs by including a proxy for RG as a covariate in the background model, as suggested by Buchholz and colleagues (2022), does not seem to suffice in minimizing the biases induced by RG. The person-level model operates under the assumption that response patterns accurately represent ability, an assumption that is violated under RG. As the individual posterior predictive distribution of PVs is based on both the measurement and the background model, adjustments in the background model may be insufficient if the likelihood in the measurement model is already distorted by responses that do not reflect true ability. Nevertheless, even the response-level models remained imperfect, particularly in estimating the ability of lower-performing groups, which tended to be overestimated. In these cases, combining person- and response-level adjustments provided the most accurate estimates by also addressing the relationship between RG and competence.

However, the performance of the models that were effective in accounting for RG (i.e., the response-level and combined models) varied more between the different threshold methods than between the different approaches used for accounting for RG. Notably, the combined model only yielded the best results when using a threshold method with a low misclassification rate. These findings suggest that the choice of threshold can also impact model performance in PV estimation and align with those of Rios and Deng (2023) on point estimates of ability. They highlight the importance of thorough threshold selection before choosing the correct model for accounting for RG.

The empirical study demonstrated that the proposed approaches of accounting for RG in PV estimation are both applicable and feasible in practice. In the sample used for the empirical application, all models of (not) accounting for RG yielded similar results. This is consistent with the findings from the simulation study for the condition with a low overall RG rate, a higher RG rate for the low ability group, and a moderate group difference—characteristics that were reflected in the actual sample. As the overall RG rate was even lower than under the low RG rate condition of the simulation study, these findings further support the notion that group comparisons do not seem to be heavily biased when RG rates are low.

Implications for Group Comparisons in Large-Scale Assessments

The study demonstrates that accounting for RG in the estimation of PVs in LSAs can reduce the bias induced by RG and is feasible, as the evaluated models are

computationally efficient and easy to implement. However, our findings indicate that the bias introduced by RG is influenced by the ability distribution across groups. Failure to account for RG poses a risk of misinterpreting ability differences, potentially leading to biased conclusions about educational outcomes. This, in turn, could have far-reaching implications for educational policies and decision-making, such as changes in the ranking of countries or demographic subgroups. Therefore, it is crucial to be aware of the potential bias introduced by RG and consider appropriate corrective measures when interpreting assessment outcomes.

Limitations and Future Directions

The findings presented offer several opportunities for follow-up research. First, future research should examine whether including additional covariates related to test-taking engagement into the background model as proposed by Buchholz et al. (2022) can improve the adjustment for RG in PV estimation on the person level. For reasons of clarity, RTE as a proxy for test-taking engagement and the grouping variable were included as the only covariates in the current study. Although our simulation study showed that the RTE can reliably detect RG behavior (see Table 2 in the supplements) and thus explains a large part of the variance of test-taking engagement, it was not sufficient to adjust for the impact of RG on PV. Further studies could investigate whether an integration of a broader set of covariates, which explain an even greater part of the variance of test-taking engagement and ability, can improve the estimation of PVs.

Second, threshold selection plays a critical role in model performance, as different threshold methods can lead to significant variations in results (Rios & Deng, 2023). This study employed two well-established and easily applicable methods: the common-*k* method being practical and straightforward, and the visual-inspection method optimized for better performance. However, these methods also have their limitations in practice. The common-*k* approach disregards item-level variability, while the visual-inspection method can only detect RG when a bimodal distribution is present—an assumption that may not hold for items requiring relatively short response times under SB. Therefore, future studies could evaluate the performance of the approaches presented in this study also using different threshold techniques, for example, the normative-threshold method (by classifying all responses with a response time shorter than fixed percentage of the average response time as RG; Wise & Ma, 2012) or by incorporating accuracy information in threshold selection (e.g., Lee & Jia, 2014).

Third, to keep the scope and interpretation of the results in the simulation study comprehensible, the correlation between RG and ability was not varied in the present study and mean item response times were chosen to resemble the time demands of simple multiple-choice items. Based on previous research using empirical data (e.g., Silm et al., 2020), we selected a relatively strong correlation between RG and ability, recognizing its potential impact on model performance and bias in estimating ability

differences. However, future studies should also take variations in the relationship between ability and RG into account and evaluate the performance of the models under different conditions. Moreover, they could examine whether different time demands of items influence the classification accuracy of the different threshold methods, thereby potentially impacting the performance of the models.

Fourth, the present study focused on easily applicable approaches of accounting for RG on different analytical levels, leveraging the unique framework provided by PVs and investigating whether adjusting for RG in the background model improves performance. Building on these findings, future research should consider exploring alternative modeling approaches, such as multidimensional IRT models (Liu et al., 2019; B. Wang et al., 2025) or specialized mixture models (Nagy & Ulitzsch, 2022; C. Wang & Xu, 2015). These models account for RG on the response level while factoring in the relationship of RG and ability and may thus offer improved accuracy in handling RG effects. Finally, incorporating further process data, such as item nonresponse and text-reread, could provide a more comprehensive understanding of test-taking behavior and enhance the precision of RG adjustments in LSAs (see Welling et al., 2024, for an example).

Conclusion

This study examined the impact of RG on group comparisons based on PVs and evaluated different approaches of accounting for RG in PV estimation. While a simulation study systematically varied RG rates, true ability difference, and variations in RG prevalence between groups to assess whether the bias of RG on the group comparisons can be minimized by using different approaches of accounting for RG in PV estimation, an empirical study subsequently demonstrated that these approaches are applicable in practice. The results indicate that ignoring RG leads to systematic underestimation of ability differences, particularly when RG rates are high, the true ability difference is large, and the more able group engages in more RG. Models that accounted for RG at the response level, either alone or in combination with person-level adjustments, significantly improved estimation accuracy, though slight over- and underestimations of ability differences remained in some conditions. Person-level adjustments alone showed little advantage over the baseline model that ignored RG. Furthermore, the choice of threshold method influenced model performance, stressing the importance of thorough threshold selection.

These findings emphasize the necessity of incorporating RG adjustments in PV estimation to avoid biased group comparisons with potential consequences for educational policies and decision-making. They further indicate that accounting for RG solely on the person level with a proxy for RG does not suffice in attenuating the bias induced by RG. Future research should explore additional background covariates, refined thresholding techniques, and alternative modeling approaches, such as multidimensional IRT or integration of additional process data, to further improve RG adjustments in LSAs.

Acknowledgments

We would like to thank Elias Bohn for assisting with parts of the simulation code. We further thank Esther Ulitzsch for providing valuable feedback on the conceptualization of the study and for critically reviewing parts of the manuscript. No funding was received to assist with the preparation of this manuscript.

Author Contributions

Conceptualization—JW and EZ
Data Curation—JW and EZ
Formal analysis—JW and EZ
Methodology—JW and EZ
Supervision—TG
Visualization—JW
Writing—original draft—JW and EZ
Writing—review & editing—JW, EZ, and TG

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Ethical Approval

The study was carried out under the supervision of the German Federal Commissioner for Data Protection and Freedom of Information and in coordination with the Standing Conference of the Ministers of Education and Cultural Affairs as well as the education ministries of the participating federal states. All data collection procedures and instruments were reviewed and approved by NEPS's designated data protection and security officer, in accordance with applicable national ethical and legal standards.

Informed Consent Statements

All participants provided written informed consent prior to enrolment. Participation was voluntary, and participants could withdraw from the longitudinal study at any time.

ORCID iDs

Jana Welling https://orcid.org/0000-0002-6377-8472
Eva Zink https://orcid.org/0000-0002-3671-6372
Timo Gnambs https://orcid.org/0000-0002-6984-1276

Data Availability Statement

This article uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network. The scored test data are provided at NEPS Network (2024), whereas the process data cannot be made publicly available due to legal restrictions. The computer code, analysis results and supplemental material are provided at https://osf.io/gdnse/.

Supplemental Material

Supplemental material for this article is available online.

References

Anaya, L. M., & Zamarro, G. (2023). The role of student effort on performance in PISA: Revisiting the gender gap in achievement. *Oxford Economic Papers*, *76*(2), 533–560. https://doi.org/10.1093/oep/gpad018

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328. https://doi.org/10.1080/19345740802400072

Blossfeld, H.-P., & Roßbach, H. G. (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed.). Springer VS.

Bondarenko, I., & Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, *35*(17), 3007–3020. https://doi.org/10.1002/sim.6926

Buchholz, J., Cignetti, M., & Piacentini, M. (2022). *Developing measures of engagement in PISA*. OECD. https://doi.org/10.1787/2d9a73ca-en

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23–45. https://doi.org/10.1080/10627190709336946

DeMars, C. E. (2024). Treating noneffortful responses as missing. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644241297925

DeMars, C. E., Bashkov, B., & Socha, A. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69–82.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online*, *5*(2), 50–79. https://doi.org/10.25656/01:8424

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC*. https://doi.org/10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Martens, T., & Lüdtke, O. (2017a). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, *5*(1). https://doi.org/10.1186/s40536-017-0051-9

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017b). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Eds.), *Competence assessment in education: Research,*

models and instruments (pp. 407–425). Springer. https://doi.org/10.1007/978-3-319-50030-0_24

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), 608–626. https://doi.org/10.1037/a0034716

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Jewsbury, P. A., Jia, Y., & Gonzalez, E. J. (2024). Considerations for the use of plausible values in large-scale assessments. *Large-scale Assessments in Education*, *12*(1), 24. https://doi.org/10.1186/s40536-024-00213-y

Jin, K. Y., Siu, W. L., & Huang, X. (2022). Exploring the impact of random guessing in distractor analysis. *Journal of Educational Measurement*, *59*(1), 43–61. https://doi.org/10.1111/jedm.12310

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, *62*(2), 147–177.

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics—Theory and Methods*, *46*(22), 11341–11357. https://doi.org/10.1080/03610926.2016.1267764

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*(8), 1–24.

Liu, Y., Li, Z., Liu, H., & Luo, F. (2019). Modeling test-taking non-effort in MIRT models. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00145

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558. https://doi.org/10.1214/ss/1177010269

Michaelides, M. P., & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: Country and item-type differences. *Psychological Test and Assessment Modeling*, *3*(64), 304–338.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

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. https://doi.org/10.1111/bmsp.12117

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*, *82*(5), 845–879. https://doi.org/10.1177/00131644211045351

NEPS Network. (2024). *National Educational Panel Study, scientific use file of starting cohort first-year students*. Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SC5:19.0.0

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00220

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rios, J. A. (2022). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement*, *82*(1), 122–150. https://doi.org/10.1177/00131644211003640

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-scale Assessments in Education*, *9*(18). https://doi.org/10.1186/s40536-021-00110-8

Rios, J. A., & Deng, J. (2023). A comparison of response time threshold scoring procedures in mitigating bias from rapid guessing behavior. *Educational and Psychological Measurement*, *84*(2), 387–420. https://doi.org/10.1177/00131644231168398

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, *27*(4), 356–373. https://doi.org/10.1080/10627197.2022.2110465

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. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2021a). Investigating the impact of noneffortful responses on individual-level scores: Can the effort-moderated IRT model serve as a solution? *Applied Psychological Measurement*, *45*(6), 391–406. https://doi.org/10.1177/01466216211013896

Rios, J. A., & Soland, J. (2021b). Parameter estimation accuracy of the effort-moderated item response theory model under multiple assumption violations. *Educational and Psychological Measurement*, *81*(3), 569–594. https://doi.org/10.1177/0013164420949896

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, *22*(2), 154–184. https://doi.org/10.1080/15305058.2022.2036161

Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test Analysis Modules* (R Package Version 4.2-21). https://CRAN.R-project.org/package=TAM

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley. https://doi.org/10.1002/9780470316696

Schnipke, D. L. (1995, April 19). *Assessing speededness in computer-based tests using item response times* [Paper Presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA, United States.

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 Fall*, *34*(3), 213–232.

Sengewald, M.-A., & Mayer, A. (2024). Causal effect analysis in nonrandomized data with latent variables and categorical indicators: The implementation and benefits of EffectLiteR. *Psychological Methods*, *29*(2), 287–307. https://doi.org/10.1037/met0000489

Setzer, J. C., Wise, S. L., Van Den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 83–112. https://doi.org/10.1111/BMSP.12188

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*, 9–36.

Wang, B., Huggins-Manley, C., Kuang, H., & Xiong, J. (2025). Enhancing effort-moderated item response theory models by evaluating a two-step estimation method and multidimensional variations on the model. *Educational and Psychological Measurement*, *85*(2), 401–423. https://doi.org/10.1177/00131644241280727

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Welling, J., Gnambs, T., & Carstensen, C. H. (2024). Identifying disengaged responding in multiple-choice items: Extending a latent class item response model with novel process data indicators. *Educational and Psychological Measurement*, *84*(2), 314–339. https://doi.org/10.1177/00131644231169211

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

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

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

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment*, *26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86–105.

Wise, S. L., Kingsbury, G. G., Kong, X. J., & Thomason, J. (2004, April 15). *An investigation of motivation filtering in a statewide achievement testing program* [Paper Presentation].

National Council on measurement in education 67th annual meeting, San Diego, CA, United States.

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. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement. In M. J. Margolis & R. A. *Integrating Timing Considerations to Improve Testing Practices* (pp. 150–164). Routledge. https://doi.org/10.4324/9781351064781-11

Wise, S. L., & Kuhfeld, M. R. (2021). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130–149. https://doi.org/10.1111/jedm.12275

Wise, S. L., & Ma, L. (2012, April 13). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper Presentation]. 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. https://doi.org/10.1080/08957340902754650

Wise, S. L., Soland, J., & Bo, Y. (2020). The (non)impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57–77. https://doi.org/10.1080/15305058.2019.1605999

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Appendix C

Welling, J., Fischer, L., Thums, K., & Gnambs, T. (2026). *Prevalence, predictors, and consequences of rapid guessing among children in elementary school* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/tsvyf_v1

Supplemental Material: https://osf.io/q2ebk

Prevalence, Predictors, and Consequences of Rapid Guessing among Children in Elementary School

Jana Welling[1], Luis Fischer[2], Kathrin Thums[3], Timo Gnambs[1]

[1]Leibniz Institute for Educational Trajectories, Bamberg, Germany

[2] University of Konstanz, Konstanz, Germany

[3] Technische Hochschule Nürnberg Georg Simon Ohm, Nürnberg, Germany

Author Note

Jana Welling: https://orcid.org/0000-0002-6377-8472

Luis Fischer: https://orcid.org/0009-0002-4668-5482

Kathrin Thums: https://orcid.org/0000-0002-4285-0236

Timo Gnambs: https://orcid.org/0000-0002-6984-1276

Corresponding author information:

Correspondence concerning this article should be addressed to Jana Welling, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany.

E-mail: jana.welling@lifbi.de

Abstract

Background. The lack of consequences for test-takers in educational large-scale assessments can lead to rapid guessing, a test-taking behavior characterized by unusually short response times. Despite its potential threat to the validity of assessment outcomes, little is known about rapid guessing among elementary school students. Because children differ in their cognitive development from older individuals, the generalizability of findings from older samples is questionable. Therefore, the current study investigated the prevalence, predictors and consequences of rapid guessing among children in elementary school.

Methods. The two-wave study is based on data from the German National Educational Panel Study (N = 1,887) including a reading and a mathematics assessment in Grades 2 (n = 1,512) and 4 (n = 1,470). Rapid guessing was identified using visual inspection methods in the Grade 2 mathematics assessment and two normative thresholds (NT10 and NT20) in all other assessments; rapid guesses were scored as missing. Mixed effects logistic regressions were conducted to determine predictors of rapid guessing. Its consequences were operationalized as the impact on ability group comparisons. The estimated ability difference was compared between the filtered and unfiltered data using bootstrap confidence intervals.

Results. The prevalence of rapid guessing was generally low but differed substantially between assessments (Grade 2: 3.3% in reading, 0.8% in mathematics; Grade 4: 0.5% in reading, 0.4% in mathematics). Test performance estimated solely with engaged responses was negatively related to rapid guessing across assessments, while gender and socio-economic status were no significant predictors. Accounting for rapid guessing did not affect the estimated ability differences between girls and boys or children with low and high parental socio-economic status in any of the assessments.

Conclusions. The findings support previous research indicating that rapid guessing may be less common among elementary school children than among older populations. However, test and assessment characteristics seem to have a strong influence on the occurrence of rapid guessing in this age group. Even when some rapid guessing was present, it did not affect ability comparisons between

subgroups. Only proficiency but not socio-demographic characteristics could be confirmed as a predictor of rapid guessing among children.

Introduction

Educational large-scale assessments (LSAs), such as the *Programme for International Student Assessment* (PISA) or *Programme for the International Assessment of Adult Competencies* (PIAAC), assess competencies across different age groups and subpopulations, thereby informing policies and educational decisions. However, because these low-stakes assessments do not have personal consequences for the test-takers, they may not invest their best effort while responding to the test items. In extreme cases, test-takers may even guess the response without properly reading the item – a behavior known as *rapid guessing*, since the responses are typically characterized by unrealistically short response times and an accuracy rate close to chance level (Wise, 2017). Because such responses do not reflect the test-taker's true ability, rapid guessing poses a threat to the validity of inferences based on competence assessments (Rios et al., 2017, 2022). For instance, if subpopulations (e.g., boys and girls) differ in their rapid guessing behavior, estimates of achievement gaps between these groups may be distorted (e.g., Welling et al., 2026), threatening the validity of such comparisons.

To identify potential threats to the validity of assessments, it is crucial to examine subgroup differences in the prevalence of rapid guessing as well as its impact on assessment outcomes such as ability group comparisons. While numerous empirical studies have investigated these aspects in adolescent and adult samples (e.g., Goldhammer et al., 2016; Kroehne et al., 2020; Michaelides & Ivanova, 2022; Rios & Soland, 2022), rapid guessing in children is less well understood. Since children differ from older individuals in their cognitive development (Piaget, 2016; see also Borgers et al., 2000), their rapid guessing behavior may differ as well. For example, children's overconfidence (Van Loon et al., 2017; Xia et al., 2024) and limited capacity for strategic and prospective thinking (Lombardi et al., 2017) could lead to increased test-taking engagement. Initial empirical findings suggest that rapid guessing may indeed be less prevalent in elementary school students (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019), raising questions about the generalizability of findings based on older samples.

The current study aims to address this gap by investigating the prevalence, predictors, and impact of rapid guessing behavior in elementary school students. Following an overview of rapid guessing behavior with its determinants and consequences, we present an empirical study based on data from the German *National Educational Panel Study* (NEPS; Blossfeld & Roßbach, 2019) with reading and mathematics assessments in Grades 2 and 4. Finally, we discuss the results and their implications for LSAs involving child samples.

Rapid Guessing: Definition, Prevalence, Predictors, and Consequences

In low-stakes assessments, test-takers typically face no direct consequences for their performance. Therefore, not all examinees are motivated to respond to every item with their best effort (Schnipke & Scrams, 1997; Wise & DeMars, 2005; Wise & Smith, 2011). Test-taking behavior is usually divided into *solution behavior*, where examinees respond thoughtfully and to the best of their abilities, and *disengaged responding*, where responses are given superficially or even without fully processing the question (Schnipke & Scrams, 1997). Rapid guessing is a common form of disengaged responding where the test-taker responds so quickly to an item that it seems implausible that they have properly read and processed the item (Wise, 2017).

Two theoretical frameworks can explain the occurrence of rapid guessing: the *expectancy-value theory* as summarized by Eccles and Wigfield (2002) and the *demands-capacity model* by Wise and Smith (2011). According to expectancy-value theory, motivation depends on both the expectancy of success and the perceived value of the task (Eccles & Wigfield, 2002). Thus, test-takers will invest effort if they believe performing well is important and achievable. In low-stakes tests, however, the task value can be limited as the performance on the test has no consequences for the test-taker (Eklöf, 2010; Wise & Demars, 2010), leading to a lack of test-taking motivation. While some empirical studies support this theory (LaFave et al., 2022; Penk & Schipolowski, 2015), others do not (Baumert & Demmrich, 2001; Pools & Monseur, 2021), indicating that it cannot fully explain test-taking motivation.

The demands-capacity model (Wise & Smith, 2011), on the other hand, focuses not only on the examinee, but also on test characteristics to explain low test-taking effort. The model assumes that test-taking effort depends on a combination of both the examinee's willingness and ability to respond

effortfully (i.e., *effort capacity*) as well as the required amount of effort for solving an item correctly (i.e., *resource demands*). In low-stakes tests, the examinees can be less willing to invest effort, leading to reduced effort capacity, which in turn can diminish their actual test-taking effort. Empirical studies support the model, showing that test-taking motivation indeed depends on both item and person characteristics (Goldhammer et al., 2016; Michaelides & Ivanova, 2022; Rios & Soland, 2022).

Previous research has demonstrated that rapid guessing is indeed a prevalent issue in low-stakes assessments (Goldhammer et al., 2016; Kroehne et al., 2020; Michaelides & Ivanova, 2022; Soland & Kuhfeld, 2019). Rapid guessing rates can be reported at the examinee, item, or response (i.e., item-by-examinee) level. In a meta-analysis by Rios and colleagues (2022), approximately 7% of the responses were, on average, classified as rapid guessing, while around 28% of the examinees engaged in rapid guessing on at least one item in the test. However, reported rates vary considerably, as they depend heavily on test (e.g., domain, setting) and sample (e.g., country, age) characteristics as well as the methods used to identify and handle rapid guessing (Kroehne et al., 2020; Michaelides & Ivanova, 2022; Rios & Soland, 2022; Soland & Kuhfeld, 2019). For instance, Goldhammer and colleagues (2016) reported response-level rapid guessing rates ranging from 0% to 8% in PIAAC, depending on the country and test domain.

Several item (e.g., position, type) and person characteristics (e.g., gender, socio-economic status [SES]) have been shown to be related to rapid guessing (Goldhammer et al., 2016; Michaelides & Ivanova, 2022; Rios & Soland, 2022; Soland, 2018; Wise et al., 2021). Most notably, rapid guessing seems to be strongly associated with proficiency, with less proficient individuals generally engaging in more rapid guessing than their more proficient peers (Silm et al., 2020). While identifying predictors on the item or test level is important for designing tests that encourage examinees to invest effort into their responses, person-level predictors are critical for identifying potential threats to the validity of assessment outcomes such as ability group comparisons.

Since rapid guessing can be quite prevalent in some assessments, it is important to determine its actual impact on assessment outcomes. Because randomly guessed responses should be, on average, less likely to be correct than engaged ones, proficiency estimates compromised by rapid

guessing could be systematically underestimated. In fact, a meta-analysis by Rios and colleagues (2022) concluded that proficiency estimates are, on average, underestimated by 0.13 standard deviations when not accounting for rapid guessing. A simulation study by Rios and colleagues (2017) demonstrated that the distortion depends on both the rapid guessing rate and test difficulty, with significant downward distortion (Cohen's $d > 0.20$) when rapid guessing exceeded 6% on the easy test, or 12% on the moderate or difficult tests.

However, average score distortion is no meaningful measure for the actual impact of rapid guessing. Because educational LSAs provide information on the population level rather than on the individual level, the important aspect is *differential* score distortion – that is, are the proficiency estimates of some subpopulations (e.g., adolescents, boys) more distorted than others? Only differential score distortion can lead to biased assessment outcomes such as ability comparisons, development across time, or predictors of proficiency. For example, if one subpopulation (e.g., boys) is less engaged than others (e.g., girls), the proficiency estimates for the less engaged group may be more severely underestimated, resulting in biased achievement gap estimates or incorrect school or country rankings. Some studies suggest that the impact on assessment outcomes may be rather limited with only minimal effects on achievement gaps, school rankings, school effectiveness, and summer learning loss, even when the rapid guessing rates vary greatly between the subpopulations (Kuhfeld & Soland, 2020; Soland, 2018; Wise et al., 2020). Since findings are still rather scarce and inconclusive, it is crucial to provide more evidence on the effects of rapid guessing on educational and inequality outcomes.

Rapid Guessing in Elementary School Students

So far, most findings on rapid guessing are based on adolescent or adult samples (Goldhammer et al., 2016; Kroehne et al., 2020; Michaelides & Ivanova, 2022; Welling et al., 2024; Wise et al., 2021), while less is known about it in younger populations. This raises the question of whether existing findings on rapid guessing can be generalized to children.

Certain theoretical arguments suggest that rapid guessing may be less prevalent among children than among older populations. For instance, elementary school students tend to be more overconfident in their abilities (Van Loon et al., 2017), an effect which declines with age (Xia et al.,

2024). This heightened confidence may increase their expectancy of success and effort capacity in competence assessments. In addition, strategic and prospective thinking is more developed in adolescents and adults than in younger children (Lombardi et al., 2017). Older test-takers may therefore be more aware of the low stakes of the test and make more deliberate decisions about how much effort to invest, potentially reducing both the perceived task value and their willingness to engage fully. However, there are also reasons to expect higher rates of rapid guessing among children. Compared to older individuals, children possess shorter attention spans (Simon et al., 2023), which further decline during testing (Betts et al., 2006). If assessments are not specifically designed to accommodate their limited attentional capacity, children may perceive tasks as more demanding, which could result in rapid guessing.

Despite these differences, only few studies have reported rapid guessing rates among children across different age groups. The available evidence points towards increasing rates of rapid guessing from primary to secondary school (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023; Wise & Kingsbury, 2022). For example, Jensen and colleagues (2018) reported that about 1-2% of third graders engaged in rapid guessing, while the respective share increased to 7-12% among eighth graders. Soland and Kuhfeld (2019) identified an increase in rapid guessing rates between Grades 2 and 8 from about 0–1% to 8–23%. Taken together, these results suggest that rapid guessing may in fact be relatively uncommon among elementary school students but may become more prevalent during adolescence, when test-taking engagement tends to decline. This raises important questions about the generalizability of findings based on older individuals, particularly those from adolescent samples (e.g., PISA).

Not only the prevalence of rapid guessing but also its influencing factors may change with age. For instance, rapid guessing behavior may be less strongly associated with SES in children than in adolescents, because the influence of parental SES on children tends to increase with age (Li et al., 2025). Nevertheless, empirical evidence on predictors of rapid guessing in children remains scarce. Svetina Valdivia and colleagues (2023) descriptively compared rapid guessing rates across different subgroups. They found that boys, less proficient children, and children who were *not* eligible for free

lunch at school (a proxy for SES) exhibited more rapid guessing than their peers. While the findings on gender and proficiency align with patterns observed in older samples, the result regarding SES is somewhat unexpected: children from lower SES backgrounds were more engaged than those from higher-status groups, contradicting trends typically found in older populations.

Since the impact of rapid guessing on assessment outcomes depends – among other factors – on the rapid guessing rate itself (e.g., Rios et al., 2017), the extent of its distorting effects may be considerably lower in elementary school students than in older samples. Moreover, potential biases depend strongly on the relationship between rapid guessing and proficiency (Wise, 2020). If this relationship were weaker or stronger in younger children, the effect of rapid guessing on assessment outcomes could also diverge from what has been observed in older test-takers. So far, two U.S. studies in classroom-based assessments have reported that the effect of rapid guessing on teacher value-added estimates and educational effectiveness metrics in older children was negligible (Jensen et al., 2018; Kuhfeld & Soland, 2020). However, evidence remains scarce, highlighting the need for studies to replicate these findings in younger samples and different assessment settings.

To summarize, children‘s tendencies to engage in rapid guessing may be linked to their developmental stage, thereby also affecting its predictors and impact on the validity of assessment outcomes. Yet empirical research on this age group remains limited, with existing studies largely confined to U.S. samples and classroom-based testing contexts. To address this gap, the current study examines predictors and consequences of rapid guessing in a German sample of elementary school students in an individual assessment setting.

Research Questions and Hypotheses

This study seeks to improve the understanding of rapid guessing behavior in elementary school students by investigating its prevalence, predictors, and impact on ability group comparisons. Specifically, we address three research questions (RQ): (1) How does the prevalence of rapid guessing change across elementary school? (2) Which person-level characteristics predict rapid guessing?, and (3) To what degree does rapid guessing bias ability group comparisons (e.g., between girls and boys)? This longitudinal two-wave study is based on a sample of German elementary school students who

took a reading and a mathematics test in Grades 2 and 4. The specific hypotheses (H) addressed in the study are as follows:

In line with prior studies (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023; Wise & Kingsbury, 2022), we assume that older children (Grade 4) will exhibit more rapid guessing behavior than younger children (Grade 2). Because of rapid cognitive developments during this period (Betts et al., 2006; Piaget, 2016; Xia et al., 2024), with overconfidence (Xia et al., 2024) decreasing and prospective thinking increasing (Lombardi et al., 2017), older children may be more inclined to engage in rapid guessing than younger ones.

H1: The prevalence of rapid guessing increases between Grades 2 and 4.

Furthermore, rapid guessing rates may vary with person characteristics such as proficiency, gender, and SES (Rios & Soland, 2022). Highly proficient individuals are typically accustomed to succeed in tasks, which is associated with increased self-efficacy, higher outcome expectations, and more task enjoyment (e.g., Schunk, 1995). They may thus possess high expectancy of success, task value, and effort capacity, as well as perceive the task demands to be low, leading to heightened test-taking engagement. Indeed, proficiency has consistently been found to negatively correlate with rapid guessing behavior (Silm et al., 2020; Ulitzsch et al., 2021; Wise et al., 2020). Since younger children generally tend to be highly confident in their abilities (Van Loon et al., 2017; Xia et al., 2024), this relationship may be less pronounced in younger individuals. Because high proficiency is also linked to greater task enjoyment, we still expect that proficiency is negatively associated with rapid guessing.

Gender differences in rapid guessing are usually explained by differences in socialization and personality: agreeableness and conscientiousness are both more prevalent in females and related to test-taking engagement (Demars et al., 2013). Accordingly, most previous studies have found boys and men to engage in more rapid guessing than girls and women (Demars et al., 2013; Deribo et al., 2022; Kroehne et al., 2020; Setzer et al., 2013; Soland, 2018; Wise et al., 2021). As gender differences in personality traits do not seem to change significantly across childhood and adolescence (Slobodskaya & Kornienko, 2021), we expect to replicate these findings in elementary school students.

Parental SES shapes children's developmental environments. For instance, parents with higher SES provide more support for their children's academic development and select extracurricular activities that further foster learning (e.g., Hemmerechts et al., 2017; Pfost & Heyne, 2023), resulting in higher proficiency across domains (Li et al., 2025; White, 1982; Zhang et al., 2020). Even beyond competence, SES may influence rapid guessing. Children with highly educated parents exhibit higher self-efficacy (Han et al., 2014) and, consequently, may have a greater expectation of success and effort capacity in competence assessments (Soland & Kuhfeld, 2019). Therefore, we expect that parental SES is negatively related to rapid guessing behavior, even beyond the influence of proficiency.

H2: (a) Proficiency, (b) gender, and (c) SES predict test-taking engagement, with less proficient children, boys, and children with a low SES exhibiting more rapid guessing than more proficient children, girls, and children with a high SES.

The third research question concerns the impact of rapid guessing on ability differences between subgroups (e.g., girls and boys). To answer this question, ability group differences are compared between methods that ignore or account for rapid guessing. Because the estimated effect may be attributed to both the actual effect of rapid guessing and the effect of the scoring method itself (e.g., person-level filtering can lead to an upward bias in ability estimates; Rios et al., 2017), we refer to it as the impact of *accounting for* rapid guessing. Since the impact depends – among others – on the rate of identified rapid guesses (Rios et al., 2017), we expect that it will be smaller in younger children (Grade 2) than in older children (Grade 4).

H3: The impact of accounting for rapid guessing on ability group comparisons increases between Grades 2 and 4.

Methods

Sample and Study Design

The current study employed a two-wave longitudinal design using data from Starting Cohort 1 (Newborns) of the German *National Educational Panel Study* (NEPS; Blossfeld & Roßbach, 2019). The NEPS assesses competence development and educational trajectories across the life span following a

multi-cohort, multi-stage design. In Starting Cohort 1, a test battery including measures of reading speed, reading comprehension, and mathematics was administered on tablets or computers in Grades 2 and 4. The assessments took place in an individually proctored setting. In Grade 4, trained interviewers visited each child's home and supervised the testing session in person. Due to COVID-19–related restrictions, the Grade 2 assessment was conducted under comparable conditions but with remote proctoring via telephone.

Because the reading speed test differs significantly from other assessments in design and presentation, only the reading comprehension and mathematics assessments were used for the analyses. The longitudinal sample comprised 1,877 students who provided at least three valid responses with corresponding response-time information in one of the four competence assessments. However, only 1,204 students participated in both reading assessments and 1,190 in both mathematics assessments. Overall, 50.0% of the students were girls and 48.3% of the students had at least one parent with a university degree. The students were, on average, 8.29 years old (*SD* = 0.13) in Grade 2 and 10.23 years old (*SD* = 0.14) in Grade 4. Descriptive information for the subsamples of the four competence assessments is presented in Table 1.

Instruments and Measures

Reading Competence Test

Early reading comprehension was assessed in Grades 2 and 4 with the text comprehension subscale of the ELFE-II test (Lenhard et al., 2017), a widely used instrument designed for children from first to seventh grades. In both grades, the test was presented second in the test battery, following the reading speed assessment. It consisted of 26 multiple-choice items based on 17 short texts, each comprising two to seven sentences. For each item, students selected one of four response options by tapping or clicking on the screen. Students could neither skip nor revisit items. The assessment was limited to seven minutes. Following the test authors' scoring rules (Lenhard et al., 2017), all items that were not reached within this period were scored as incorrect.

Mathematics Competence Tests

Mathematics competence was assessed in Grades 2 and 4 with two different tests specifically developed for the NEPS. The framework and test development are described in Neumann et al. (2013). Both tests were presented third in the test battery, following the reading speed and reading comprehension assessments. They addressed five content areas: (a) space and shape, (b) change and relationships, (c) data and chance, and (d) sets, numbers, and operations, as well as (e) units and measuring. The tests comprised simple and complex multiple-choice items, which students could omit.

In Grade 2, the mathematics test consisted of 20 items, all administered to each student. Unlike many other competence tests, the item questions were presented in spoken form via pre-recorded audio, while the response options were displayed on the screen. An answer could only be provided after the audio had finished. In Grade 4, mathematics competence was assessed using a multi-stage test (Pohl, 2013) that presented different items depending on the responses during the test. The first stage consisted of eight items and was administered to all students. Based on their performance in the first stage, students were assigned to an easy, medium, or difficult second stage, each consisting of eight items. Thus, each student completed a total of 16 items. Both the questions and response options were presented in text form. The tests were limited to 28 minutes for completion; items not reached within this period were scored as missing.

Socio-Demographic Characteristics

Binary gender was assessed using parent-reports (diverse gender categories were not included). Parental SES was operationalized using two indicators: educational attainment and occupational status. Educational attainment was measured with the *Comparative Analysis of Social Mobility in Industrial Nations* index (CASMIN; Brauns et al., 2003), which classifies the highest educational qualification into nine categories ranging from 0 (no formal degree) to 8 (university degree). Occupational status was assessed with the *International Socio-Economic Index of Occupational Status* (ISEI-08; Ganzeboom et al., 1992), which measures the prestige of each occupation on a scale ranging from 10 (e.g., garbage collector) to 98 (e.g., senior judge).

For each indicator, the highest value of either parent (HCASMIN, HISEI) was used as the measure of parental SES. Because HCASMIN is categorical and HISEI was not normally distributed, both variables were dichotomized at the median for the analyses. Parental educational attainment was classified as high when at least one parent held a university degree (HCASMIN = 8) and low otherwise. Parental occupational status was classified as high when HISEI was greater than 72, corresponding to occupations with upper-middle or higher status, and low otherwise.

Gender and both measures of parental SES were repeatedly assessed in Starting Cohort 1 of the NEPS. For each child and their parents, the first available non-missing value for gender, CASMIN, and ISEI up to Grade 4 was used in the analyses.

Item Response Time

Response times were computed from raw process data, which recorded when a test-taker accessed each page. For each item, response time was defined as the total time the test-taker spent on the corresponding item page.

Rapid Guessing

In the literature, a plethora of methods to identify and handle rapid guessing has been proposed. To enable comparisons with previous findings on rapid guessing in adolescence and adult samples, the current study uses the threshold-based method, which is the most widespread and established approach for the measurement of rapid guessing (Rios & Deng, 2021; Silm et al., 2020). The threshold-based approach assumes that rapid guesses are characterized by unusually short response times. For each item, a response-time threshold is set to distinguish between rapid guessing and solution behavior. If the item response time t_{ij} of person *i* on item *j* is shorter than the item-specific threshold ρ_j, the engagement ϕ_{ij} is classified as rapid guessing (ϕ_{ij} = 0), otherwise as solution behavior (ϕ_{ij} = 1):

$$\phi_{ij} \in \begin{cases} 0, & if\ t_{ij} < \rho_j \\ 1, & if\ t_{ij} \geq \rho_{j.} \end{cases} \quad (1)$$

This method requires the selection of appropriate thresholds. In the current study, we adopted different threshold methods depending on the assessment's characteristics. Due to the test design,

some of the items in both reading and the Grade 4 mathematics assessments exhibited only few responses (e.g., 67 in the last item of Grade 2 reading), rendering item-specific threshold selection based on the visual inspection of response times or additional information (e.g., accuracy, item-total correlation) impossible. Therefore, rapid guessing was identified using the normative threshold (NT) method, which defines the threshold as a fixed proportion of the average item response time (e.g., 10%; Wise & Ma, 2012). This approach determines thresholds based on item-specific response-time information without relying on the observed response time distribution, making it suitable even for contexts with few responses per item or infrequent rapid guesses. Because of its broad applicability, it is among the most commonly used threshold approaches in the rapid guessing literature (e.g., Jensen et al., 2018; Kuhfeld & Soland, 2020; Michaelides & Ivanova, 2022; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023). Because no threshold procedure perfectly classifies rapid guesses and the true rapid guessing rate is unknown, it is essential to compare different threshold approaches in empirical applications. The current study examined two thresholds for both reading and the Grade 4 mathematics assessments: *NT10* (10% of average response time) and *NT20* (20% of average response time).

In the Grade 2 mathematics assessment, on the other hand, all items exhibited sufficient responses for threshold selection based on visual inspection. Furthermore, because the children could not select a response before the audio presenting the item was finished, this assessment contained no short response times, rendering threshold selection not based on visual inspection (e.g., NT) difficult. Therefore, rapid guessing was identified using the visual inspection of bimodality (VIB; Wise et al., 2009) and the visual inspection of proportion of correct scores (VICS; Guo et al., 2016; Lee & Jia, 2014). The VIB method assumes that response times in low-stakes assessments follow a bimodal distribution with the first mode corresponding to rapid guessing and the second mode to solution behavior. The threshold is set at the intersection of both modes. If no bimodal distribution was discernible, the threshold was set to zero (meaning no rapid guessing). The VICS method, on the other hand, assumes that disengaged responding is not only characterized by short response times but that the average accuracy differs between rapid guessing and solution behavior. For rapid guessing, it should range

around chance, whereas for solution behavior, it should be higher. Thus, response times are divided into different response time bins (e.g., one second) and the average accuracy is computed for each bin and plotted across the response time distribution. The threshold is set at the beginning of the first response-time bin (e.g., 6 seconds for the (6;8] seconds bin) for which the average accuracy exceeds chance level. However, because the average accuracy under rapid guessing does usually not correspond exactly to chance (Wise, 2017), and because the average accuracy under solution behavior may also lie below chance level, we defined the threshold in the current study as the beginning of the first response-time bin for which the average accuracy fell above chance level plus 10 percentage points *or* below chance level minus 10 percentage points. For example, in a multiple-choice question with five response options (i.e., chance level = 20%), the threshold would be set at the start of the first response-time bin for which the average accuracy exceeds 30% or falls below 10%. If the accuracy of the first response-time bin lay already above or below these criteria, the threshold was set to zero (meaning no rapid guessing). To prevent random fluctuations in accuracy due to small bin sizes, response times were combined in two-second bins. The accuracy was computed if a bin contained at least five responses.

For each assessment, the thresholds were validated at the test level using three measures: (1) comparing the accuracy rate between rapid guessing and solution behavior, (2) the overall accuracy rate as a function of response time, and (3) comparing the item-total correlation between rapid guessing and solution behavior. If classification were perfect, the accuracy for rapid guesses would approach chance level, providing an estimate of the threshold's classification performance (Goldhammer et al., 2016; Wise, 2019; Wise & Gao, 2017). Similarly, overall accuracy should be near chance for very fast responses and increase with longer response times. Plotting overall accuracy against response time therefore offers an additional indication of threshold performance (Guo et al., 2016; Lee & Jia, 2014). However, because this criterium is identical to the VICS method, it was only used for the reading and Grade 4 mathematics assessments. The last criterium assumes that rapid guesses provide no psychometric information and should therefore correlate only marginally with the overall test performance (Goldhammer et al., 2016; Wise, 2019; Wise & Gao, 2017; Wise, 2019, specifically suggests that rapid guessing can be assumed when $r < 0.20$). However, when the thresholds

do not distinguish perfectly between rapid guessing and solution behavior but misclassify at least some engaged responses as rapid guessing, the score of short response times may indeed be related to test performance, because disengaged responses should be more often incorrect than engaged responses and because disengaged test takers should perform, on average, worse than engaged test takers. Furthermore, when the accuracy is close to 1 or 0, the correlation may be small even under solution behavior due to the low variance in scores. Therefore, this criterium should be handled with caution.

Proficiency Estimates

After identifying rapid guesses, effort-moderated scoring (Wise & Demars, 2006) was applied, which yields better parameter recovery than other scoring methods (Rios et al., 2017; Rios & Deng, 2024). This approach scores all rapid guesses as missing values before estimating respondents' proficiencies. For each test and dataset (rapid guesses contained versus scored as missing), proficiency scores were estimated as weighted likelihood estimates (WLE; Warm, 1989) using the one-parametric logistic test model (Rasch, 1960), which is consistent with the ELFE and NEPS test designs (Lenhard et al., 2017; Neumann et al., 2013).

Statistical Analyses

Data preparation and statistical analyses were conducted in R (Version 4.5.2; R Core Team, 2025), using the packages *effectsize* (Version 1.0.1; Ben-Shachar et al., 2020), *boot* (Version 1.3-32; Canty & Ripley, 1999), and *lme4* (Version 1.1-37; Bates et al., 2015). The scored test data is available at NEPS Network (2025), while the process data cannot be made publicly available due to legal restrictions. The analysis code, results, and supplemental material are provided at https://osf.io/q2ebk/.

The demographic data and the reading tests contained no missing values. Because small numbers of missing responses do not introduce severe parameter bias in item response modeling (Rose et al., 2010), missing data in the mathematics tests were ignored. As sensitivity analysis, statistical analyses were conducted for both SES measures. The results for HCASMIN are reported in the main part of the manuscript, while the results for HISEI are summarized at the end of the results section and reported in detail in the supplemental material. Longitudinal analyses were conducted only for the

reading domain, as the same test was administered in both grades. In contrast, the mathematics tests differed substantially between Grades 2 and 4, meaning that differences in rapid guessing rates may reflect test differences rather than true changes in test-taking behavior.

RQ1 and RQ2 were addressed using mixed-effects logistic regression analysis with a binary response-level engagement indicator (0 = rapid guess, 1 = solution behavior) as dependent variable and student as person-level grouping factor. To determine changes in rapid guessing behavior between Grades 2 and 4 (RQ1), the engagement indicator of both reading tests was regressed on a wave indicator (0 = Grade 2, 1 = Grade 4). To identify predictors of rapid guessing behavior (RQ2), the engagement indicator was regressed on proficiency, gender, and SES for each assessment separately. Proficiency was operationalized using the WLEs adjusted for rapid guessing. Effects were interpreted using average marginal effects (Arel-Bundock et al., 2024) and Pseudo-R^2 (McFadden, 1974).

RQ3 concerns the effect of accounting for rapid guessing on ability comparisons across gender and SES groups. This effect was determined using a three-step procedure. First, for each test and dataset (based on either the original or corrected WLEs), we estimated the ability difference by computing Cohen's *d* with pooled standard deviation (Goulet-Pelletier & Cousineau, 2018) between the two groups (girls vs. boys, or low vs. high SES). Group variances were assumed to be equal when the ratio did not exceed 1.5 (Blanca et al., 2018). Second, for each test *T*, we determined the difference between the estimates $\Delta_T = d_{T,corrected} - d_{T,uncorrected}$, with a positive Δ_T indicating that the estimated ability difference is more positive after accounting for rapid guessing. Third, the 95% confidence intervals for Δ_T were computed using the bias-corrected and accelerated (BCa) bootstrap method (Efron, 1987), which adjusts for potential bias and skewed distributions in the bootstrapped estimates. We used 5000 bootstraps to ensure stable and precise interval estimates (Davison & Hinkley, 1997).

Results

Response-Time Thresholds

The average item response times in the reading domain varied between 9.36 and 59.03 seconds (*Mdn* = 23.16) in Grade 2 and between 9.55 and 33.34 seconds (*Mdn* = 19.49) in Grade 4. In

the mathematics domain, average item response times ranged between 22.37 and 115.05 seconds (*Mdn* = 46.48) in Grade 2 and between 29.2 and 87.96 seconds (*Mdn* = 57.70) in Grade 4. In the reading and the Grade 4 mathematics assessment, thresholds were defined as 10% (NT10) or 20% (NT20) of the average item response times. For example, in the reading assessment of Grade 2, the NT10 thresholds ranged between 0.94 and 5.90 seconds (*Mdn* = 2.32), while the NT20 thresholds ranged between 1.87 and 11.81 seconds (*Mdn* = 4.63), see Tables S1, S2, and S4 in the supplemental material. In the Grade 2 mathematics assessment, VIB thresholds were set at the intersection of the bimodal response-time distribution and VICS thresholds at the time when the average accuracy considerably exceeded chance level. For all items, no clear bimodal distribution was discernible, but few extremely short response times (identified as outliers in the histogram) were classified as disengaged on some items. Average accuracy of the shortest responses ranged only in two items around chance level.

The thresholds were validated at the test level using three measures: (1) comparing the accuracy rate between rapid guessing and solution behavior (Table 2), (2) the overall accuracy rate as a function of response time (Figure 1), and (3) comparing the item-total correlation between rapid guessing and solution behavior (Table 2). Except for the VIB thresholds in the Grade 2 mathematics assessment, the accuracy rate of solution behavior was substantially higher than that of rapid guessing. In the Grade 2 reading assessment, the accuracy rate among rapid guesses identified by the NT20 thresholds was closer to chance level than the accuracy rate based on the NT10 thresholds. In the Grade 4 reading assessment, the NT10 thresholds did not detect any rapid guesses, whereas the NT20 thresholds identified rapid guesses with an accuracy rate slightly higher than chance level (+ 9.9 percentage points). In the Grade 2 mathematics assessment, the accuracy rate for rapid guessing resembled chance for the VICS method but was extremely high for the VIBS method. In the Grade 4 mathematics assessment, both threshold methods exhibited an accuracy rate of disengaged responses close to chance level. These findings support the NT20 thresholds in the reading assessments, the VICS thresholds in the Grade 2 mathematics assessment and both methods in the Grade 4 mathematics assessment.

Figure 1 shows that in both reading assessments and in the Grade 4 mathematics assessment, accuracy rates were close to chance level for very short responses and increased with longer response times. The plots also indicate that accuracy exceeded chance level at approximately 6–7 seconds in the Grade 2 reading assessment, 4–5 seconds in the Grade 4 reading assessment, and 7–8 seconds in the Grade 4 mathematics assessment. When comparing these findings to the average threshold for each method (Table 2), they support the NT20 thresholds for the reading assessments and the NT10 thresholds for the mathematics assessment.

The correlation between the score and overall test performance was higher for disengaged responses than for engaged responses in all assessments except for Grade 2 reading. In the Grade 2 reading assessment, the correlation was comparable between disengaged and engaged responses using the NT20 thresholds and substantially smaller for rapid guesses than for solution behavior using the NT10 thresholds. This finding demonstrates that all thresholds except for the NT10 thresholds in the Grade 2 reading assessment have probably misclassified at least some engaged responses as disengaged. However, it does not show which proportion of the responses classified as rapid guesses were in fact engaged.

Considering all validation measures and following Rios' (2022) recommendation to favor liberal thresholds, we concluded that the NT20 method performed best in both reading assessments, the VICS method in the Grade 2 mathematics assessment, and the NT10 method in the Grade 4 mathematics assessment. For clarity, only the results of these primary threshold methods are presented in the main text, while analyses using the second threshold method are included in the supplemental material as sensitivity analysis. Because the VIB method did not seem to accurately classify test-taking engagement, it was not included in further analyses.

Ability Estimates

In the Grade 2 reading assessment, the estimated WLE reliability was 0.88 for both models ignoring and accounting for rapid guessing. In the Grade 4 reading assessment, it was 0.87 for both models. In the mathematics assessments, the WLE reliability was estimated to be lower: In Grade 2, it

was 0.73 for the model ignoring and 0.72 for the model accounting for it, whereas in Grade 4, the reliability was 0.58 for the model ignoring rapid guessing and 0.57 for the model accounting for it.

In the Grade 2 reading assessment, the ability estimates of the model ignoring rapid guessing correlated nearly perfectly with the ability estimates of the model accounting for rapid guessing, r = .99. In all other assessments, the correlation between the ability estimates was perfect, r = 1.00. Figure 2 illustrates that accounting for rapid guessing could lead both to slight over- and underestimations of individual ability estimates. In summary, accounting for rapid guessing had barely any effect on the estimated ability.

Prevalence of Rapid Guessing

In the mathematics assessments, 0.8% of the responses were identified as rapid guesses in Grade 2, compared to 0.4% in Grade 4. Because the assessments differed in test design and identification method, differences in test-taking engagement may not only be attributed to age differences, but also to differences in test characteristics or classification accuracy. In the reading assessments, 3.3% of the responses were identified as rapid guesses in Grade 2, while 0.5% rapid guesses were identified in Grade 4. To test H1 for the reading domain, a mixed-effects logistic regression with test-taking engagement as response-level dependent variable and student ID as person-level grouping factor was conducted (Tables 4 and 5). In total, 89% of the variance in test-taking engagement could be attributed to person-level differences. The predictor (Grade) explained 3% of the variance in test-taking engagement. Contrary to H1, being in Grade 4 did not go along with lower test-taking engagement, b = 2.05, z = 16.04, p_{one_sided} = 1.000. The estimated marginal probabilities of engaging in solution behavior differed by 0.02 percentage points (z = 3.24, p_{one_sided} = .999) between the two grades, meaning that being in Grade 4 was associated with 0.02 percentage points *higher* probability of engaging in solution behavior. In summary, the prevalence of rapid guessing varied considerably between all four assessments but did not increase between Grades 2 and 4 in the reading assessments.

Predictors of Rapid Guessing

Descriptively, girls and children with at least one parent with a university degree were more engaged than boys or children with no parent with a university degree (Table 3). Depending on the

domain and grade, 0.2% to 2.6% of the girls' responses were identified as rapid guesses, compared to 0.5% to 4.0% of the boys' responses. The rapid guessing rates of children with no parent with a university degree ranged between 0.6% and 4.4%, whereas the rates of children with at least one parent with a university degree ranged between 0.2% and 2.3%. To test H2, a mixed-effects logistic regression with test-taking engagement as response-level dependent variable and student ID as person-level grouping factor was conducted for each assessment separately (Tables 4 and 5).

The estimated intraclass correlations showed that 88%, 94%, 30% and 95% of the variance in test-taking engagement could be attributed to person-level differences, for the Grade 2 reading, Grade 4 reading, Grade 2 mathematics, and Grade 4 mathematics assessments, respectively. Altogether, the predictors explained 7% of the variance in test-taking engagement in the Grade 2 mathematics assessment, compared to 2% in the Grade 4 mathematics and 1% in both reading assessments. Supporting H2a, proficiency was a significant predictor of test-taking engagement across assessments (Grade 2 reading: $b = 0.26$, $z = 3.19$, $p < .001$; Grade 4 reading: $b = 0.40$, $z = 2.12$, $p = .017$; Grade 2 mathematics: $b = 0.48$, $z = 6.71$, $p < .001$; Grade 4 mathematics: $b = 1.13$, $z = 2.91$, $p = .002$). In the Grade 2 reading and mathematics assessments, a one-unit increase in proficiency went along with a 0.01 or 0.05 percentage points increase in the probability of engaging in solution behavior, respectively (reading: $z = 2.04$, $p = .021$; mathematics: $z = 4.23$, $p < .001$). In the other two assessments, the marginal effects were not significant. In contrast to H2b and H2c, gender and SES were not significant predictors of test-taking engagement across assessments. In summary, only proficiency predicted rapid guessing behavior, although the effect was rather small across all assessments.

Impact of Accounting for Rapid Guessing

The estimated ability differences and their standard errors for all four assessments and for the two different methods (ignoring rapid guessing vs. accounting for it) are displayed in Figure 3. In the Grade 2 reading assessment, no significant difference between girls and boys was found when ignoring rapid guessing, $d = –0.06$ [CI 95%: –0.16, 0.05], or when accounting for it, $d = –0.05$ [CI 95%: –0.15, 0.06]. Accounting for rapid guessing slightly underestimated the ability difference by $\Delta = 0.01$ [CI 95%: 0.00, 0.02], but the effect was minimal and not significant. In the Grade 4 reading assessment, girls

were estimated to outperform boys when ignoring rapid guessing, d = –0.13 [CI 95%: –0.24, –0.03], and when accounting for it, d = –0.14 [CI 95%: –0.24, –0.03]. There was no effect of accounting for rapid guessing, Δ = 0.00 [CI 95%: 0.00, 0.00]. In the Grade 2 mathematics assessment, boys outperformed girls by d = 0.22 [CI 95%: 0.12, 0.32] for both methods, yielding no effect of rapid guessing. In the Grade 4 mathematics assessment, the ability difference was d = 0.17 [CI 95%: 0.07, 0.28] when ignoring rapid guessing and d = 0.18 [CI 95%: 0.08, 0.28] when accounting for it. Accounting for rapid guessing slightly overestimated the ability difference by Δ = 0.01 [CI 95%: 0.00, 0.02], but the effect was marginal.

Across all assessments, children with at least one parent with a university degree performed better than children with no parent with a university degree. In the reading assessments, the estimated ability difference was d = –0.27 [CI 95%: –0.37, –0.17] in Grade 2 and d = –0.46 [CI 95%: –0.56, –0.36] in Grade 4 for both methods. Therefore, there was no effect of accounting for rapid guessing in Grade 2, Δ = 0.00 [CI 95%: –0.01, 0.02], or in Grade 4, Δ = 0.00 [CI 95%: –0.00, 0.00]. In the Grade 2 mathematics assessment, the ability difference was d = –0.35 [CI 95%: –0.45, –0.25] using both methods, yielding no effect of rapid guessing. In the Grade 4 mathematics assessment, the estimated ability difference was d = –0.54 [CI 95%: –0.64, –0.44] when ignoring rapid guessing or when accounting for it. The effect of accounting for rapid guessing was not significant, Δ = 0.00 [CI 95%: –0.00, 0.01].

In summary, accounting for rapid guessing did not affect the estimated ability difference in any of the four assessments for neither gender nor parental SES. Therefore, the impact did not increase between Grades 2 and 4. H3 was thus not supported and did not need to be further tested.

Sensitivity Analysis

Sensitivity analyses using HISEI as an alternative indicator of parental SES yielded results that were highly comparable to those obtained with HCASMIN. HISEI was not a significant predictor of test-taking engagement in any of the assessments (Tables S5 and S6). Likewise, accounting for rapid guessing did not meaningfully alter the estimated ability differences between children with high and low HISEI (Table S7).

In the reading assessments, the NT10 method could identify some rapid guesses in Grade 2 compared to no rapid guesses in Grade 4. In the mathematics assessments, the NT20 thresholds

identified roughly twice as many rapid guesses as the NT10 thresholds in Grade 4, whereas the VIB thresholds could not accurately classify test-taking engagement in Grade 2. To test the robustness of the results, RQ2 and RQ3 were re-analyzed using the second threshold method in the Grade 2 reading and Grade 4 mathematics assessments. The findings were nearly identical to those obtained with the primary threshold method (see Figure S1 and Tables S8 to S10 in the supplemental material). Across both assessments, proficiency was the only significant predictor of test-taking engagement. Moreover, accounting for rapid guessing did not affect the estimated ability differences across genders or SES-groups.

Discussion

Although rapid guessing has been widely studied over the past two decades, most evidence is based on adolescent or adult samples (e.g., Goldhammer et al., 2016; Kroehne et al., 2020; Michaelides & Ivanova, 2022; Wise et al., 2021), while research in younger populations remains scarce. The current study addressed this gap by investigating the prevalence, predictors, and impact of rapid guessing in a sample of German second and fourth graders who took reading and mathematics assessments. Because the mathematics tests differed across measurement occasions, longitudinal analyses were limited to the reading domain.

Prevalence of Rapid Guessing

Overall, estimated rapid guessing rates in Grades 2 and 4 were low compared to those reported in adolescent or adult samples (Rios et al., 2022), which aligns both with our theoretical assumptions and previous findings on test-taking engagement of children (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023; Wise & Kingsbury, 2022). Whereas a substantial amount of rapid guessing was identified in the Grade 2 reading assessment, all threshold methods estimated that less than 1% of the responses were rapid guesses in the other assessments. Difficulties in the threshold selection based on visual inspection and the high score-total correlations among rapid guesses indicate that the actual prevalence may have been even lower than the results suggest. Because the prevalence varied between assessments, test-taking engagement among children may also depend on assessment characteristics such as test design and setting.

In contrast to Hypothesis 1, rapid guessing did not increase between Grades 2 and 4 in the reading domain. We see two plausible explanations for this finding. First, although test-taking engagement may generally decrease with age as cognitive development advances, the difference between second and fourth graders may be too small to detect. Previous studies that report decreasing rapid guessing rates across childhood and adolescence found only minimal descriptive differences in younger children, with substantial increases typically emerging around puberty (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023).

Second, descriptive results for the Grade 4 reading assessment showed that (a) accuracy rates were extremely high and (b) average item response times were very short, suggesting that this test may have been very easy for most of the fourth graders. Easier items tend to elicit less rapid guessing (Goldhammer et al., 2017), which may explain the lack of an age effect. Moreover, short response times led to very low response-time thresholds in Grade 4, making rapid guesses difficult to detect and potentially increasing misclassifications. This finding highlights a potential limitation of the NT method: when average response times are extremely short or long, thresholds may become unrealistic and classification accuracy may suffer.

In summary, the findings suggest that rapid guessing is generally less common among elementary school children, with their test-taking engagement depending strongly on test characteristics as well as the fit between sample, test, and identification method.

Predictors of Rapid Guessing

In line with Hypothesis 2a, mixed-effects logistic regressions revealed that proficiency was a significant predictor of test-taking engagement across all assessments. Highly proficient individuals exhibited less rapid guessing than their peers, even though the effects were small. This result is consistent with findings from older samples (Silm et al., 2020; Ulitzsch et al., 2021; Wise et al., 2020) and supports the assumption that higher proficiency is associated with greater task enjoyment and motivation, even among young children.

Whereas proficiency was replicated as predictor of test-taking engagement, neither gender nor SES significantly predicted test-taking engagement, thereby contradicting Hypotheses 2b and 2c.

Consequently, gender and SES may indeed play a limited role in rapid guessing at this age. Although descriptive (non-significant) results showed higher rapid guessing rates among boys and children from lower-SES backgrounds, these differences might only become meaningful in later years (e.g., Deribo et al., 2022; Kroehne et al., 2020; Wise et al., 2021). For example, the influence of parental SES increases as children age (Li et al., 2025), and may be too weak in early elementary school to meaningfully affect test-taking behavior. Additionally, situational factors such as test characteristics may be more important than small individual differences.

The null effects may also be attributed to methodological causes. Many previous studies reporting demographic differences in test-taking engagement rely purely on descriptive statistics (Demars et al., 2013; Kroehne et al., 2020; Setzer et al., 2013; Soland, 2018; Wise et al., 2021), and those using inferential methods typically report insignificant or negligible effects (Goldhammer et al., 2016; Lindner et al., 2019; Rios & Soland, 2022). Thus, potential effects of gender and SES may be too small to be statistically detectable. Moreover, rapid guessing was rare in the present study, which can lead to an underestimation of effects due to rare-events bias (King & Zeng, 2001). The extremely small marginal effects support this interpretation: although raw differences in rapid guessing rates between demographic groups sometimes reached 2%, the estimated marginal probabilities of engaging in rapid guessing differed by only about 0.02 percentage points. These results show that identifying predictors of test-taking engagement in samples with very few rapid guesses can be statistically challenging.

Impact of Accounting for Rapid Guessing

Accounting for rapid guessing did not affect ability group comparisons between girls and boys or children with low and high SES in either grade. Therefore, in contrast to Hypothesis 3, its impact did not differ between Grades 2 and 4. While in the mathematics assessments and the Grade 4 reading assessment the null findings could be attributed to the low prevalence in rapid guessing, in the Grade 2 reading assessment this interpretation seems unplausible. However, the null findings align with prior research showing that although aggregated proficiency estimates may be to some degree distorted by rapid guessing (Rios et al., 2017, 2022), the actual impact on assessment outcomes often remains small, even when rapid guessing is non-negligible (Kuhfeld & Soland, 2020; Wise et al., 2020).

Wise (2020) provides two reasons for these rather unintuitive findings. First, although some rapid guessing may be prevalent, most of the students are engaged, minimizing the deleterious effect of rapid guessing. Second, rapid guessing is usually strongly negatively related to proficiency (e.g., Silm et al., 2020). As the probability of a correct response is – even for solution behavior – rather low for less proficient individuals, the impact of rapid guessing on their proficiency estimates can be minimal or – under certain conditions – even positive (Kuhfeld & Soland, 2020).

While Wise's arguments explain why the impact of rapid guessing on aggregated scores may be smaller than expected, they do not explain why they are even smaller in ability comparisons (e.g., by ranking or direct group comparisons). We argue that for detecting effects on *comparisons* between subpopulations or measurement occasions, the differences in rapid guessing rates between the groups need to be large (see also a recent simulation study by Welling et al., 2026), which may rarely occur in practice. Moreover, less-engaged groups are often also less proficient, which in turn reduces the effect of rapid guessing on proficiency estimates (Welling et al., 2026; Wise, 2020). In summary, the results indicate that rapid guessing may not meaningfully distort assessment outcomes in younger populations.

Implications for Research and Practice

The results support previous findings on test-taking engagement among elementary school children, indicating that rapid guessing is relatively uncommon in younger populations (Jensen et al., 2018; Kuhfeld & Soland, 2020; Soland & Kuhfeld, 2019; Svetina Valdivia et al., 2023; Wise & Kingsbury, 2022). Even when the children were partly disengaged, it did not affect ability group comparisons between subgroups. This valuable insight suggests that the deleterious effects of rapid guessing may be limited among elementary school children. However, because the occurrence of rapid guessing may depend more on assessment characteristics than on age itself, and because other assessment outcomes may still be affected, replication studies across different domains, settings, and outcomes remain necessary.

The present study also illustrates that there is no "one size fits all" approach for determining response-time thresholds. While some thresholding procedures are heuristic (e.g., common-*k*

method), others require large sample sizes throughout the assessment (well into the thousands, e.g., identification based on score-total correlation), substantial rapid guessing rates, or specific response time distributions (e.g., VIB). In the current study, different threshold-selection methods were used to account for different test designs. Interestingly, the methods performed differently depending on the assessment. For example, the NT approach determines thresholds based on the average item response times, which may lead to unrealistic thresholds and increased misclassification when response times are unusually short. This finding emphasizes the need to consider and compare several thresholding approaches in empirical applications, using different validation methods to determine the best response time thresholds for the data at hand.

Limitations and Future Directions

This study contributes to the limited literature on rapid guessing in children by investigating the prevalence, predictors, and impact of rapid guessing in a longitudinal multi-domain study of German second and fourth graders. In contrast to previous studies, the assessment took place in an individually proctored setting. By not only determining the prevalence and predictors of rapid guessing, but also its actual impact on ability group comparisons as an example of assessment outcomes, the study contributes to an improved understanding of the validity of proficiency estimates in educational LSAs.

Besides its considerable strengths such as the use of multiple competence domains administered in a longitudinal design, the study also possesses some limitations. First, the design of the study does not allow identifying test domain or setting effects. Because there was no rotation design for the different test domains and because the tests used different time limits and response formats, differences in rapid guessing behavior between the two test domains should be interpreted only carefully. Moreover, the individually proctored setting of the assessments may have affected the test-taking behavior of the students. In addition, the setting differed slightly between the two measurement occasions due to COVID-19–related restrictions in the Grade 2 assessment, which may have influenced test-taking engagement. Kroehne et al. (2020) demonstrated that rapid guessing rates can indeed depend on the assessment setting, with students being more engaged in proctored, group-

based settings than in un-proctored, group-based or online settings. However, to our knowledge, no study so far has investigated the effect of individually proctored settings on test-taking engagement or the effect of assessment setting in younger populations. Future studies should address this gap by conducting experiments with randomized settings, including samples of different age groups.

Furthermore, the study comes along with some methodological limitations. First, proficiency was operationalized as the rapid-guessing-adjusted test performance. Even though this approach should eliminate most disengaged responses, it may be at least partly confounded with test-taking engagement. The proficiency estimates will be based on at least some misclassified responses, and children who rapidly guess on some items may be prone to be only partially engaged on other items. However, in absence of high-stakes test results, the adjusted WLEs are the best estimates of true proficiency, because parents' or teachers' assessments of a child's proficiency were repeatedly shown to be unreliable (Dollinger, 2013; Schmitt et al., 2022). Second, the rarity of rapid guessing may have limited the precision of the estimated effects (King & Zeng, 2001). In assessments with few rapid guesses (i.e., Grade 4 mathematics and reading), regression coefficients exhibited larger standard errors, reflecting increased estimation uncertainty and reduced statistical power to detect potential associations.

Even beyond these limitations, there are several possible directions for future research. The current study focused on rapid guessing, because it is a widespread and well-investigated form of low test-taking engagement (Kroehne et al., 2020; Rios et al., 2022; Rios & Soland, 2022; Schnipke & Scrams, 1997; Wise, 2017). However, recent research has shown that test-taking engagement may not be dichotomous (either rapid guessing *or* solution behavior), but rather a more continuous construct (Nagy et al., 2022; Wise & Kingsbury, 2022; Wise & Kuhfeld, 2021). Therefore, test-takers could not only be either engaged or disengaged, but also partially engaged. Future studies could include partial disengagement in their research questions, thus attaining a fuller picture of the prevalence and impact of low test-taking engagement.

Furthermore, future studies could replicate the results using different identification and scoring methods. While the present study applied a threshold-based approach due to its wide prevalence in

the rapid guessing literature, future research could investigate the applicability of model-based approaches (Leventhal & Pastor, 2023; Nagy & Ulitzsch, 2021; Ulitzsch et al., 2021) in child samples. For example, the low prevalence rates in younger populations may lead to convergence problems in mixture models.

Conclusion

While most research on test-taking engagement has focused on adolescent and adult samples, less is known about rapid guessing in elementary school children. Because children differ from older individuals in their cognitive development, findings from older samples may not be generalizable to child samples. Therefore, the present study investigated the prevalence, predictors, and impact of rapid guessing in a German sample of second and fourth graders. Results showed that the prevalence of rapid guessing was generally low, although it varied considerably between assessments. Proficiency in a test domain predicted test-taking engagement, with less proficient children exhibiting more rapid guessing than their peers. Lastly, accounting for rapid guessing did not affect ability comparisons between subgroups. The results indicate that in the end, rapid guessing may not pose a substantial threat to the validity of LSAs with child samples. However, to be able to make reliable statements, the results need to be replicated and extended to other assessment outcomes. In summary, this study contributes to an improved understanding of both test-taking engagement in younger populations and of the consequences for the validity of proficiency estimates in educational LSAs.

References

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to Interpret Statistical Models Using **marginaleffects** for *R* and *Python*. *Journal of Statistical Software*, *111*(9). https://doi.org/10.18637/jss.v111.i09

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

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. https://doi.org/10.1007/BF03173192

Ben-Shachar, effectsize: E. of effect size indices and standardized parameters, Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Betts, J., Mckay, J., Maruff, P., & Anderson, V. (2006). The development of sustained attention in children: The effect of age and task load. *Child Neuropsychology*, *12*(3), 205–221. https://doi.org/10.1080/09297040500488522

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, *50*(3), 937–962. https://doi.org/10.3758/s13428-017-0918-2

Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-23162-0

Borgers, N., De Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *66*(1), 60–75. https://doi.org/10.1177/075910630006600106

Brauns, H., Scherer, S., & Steinmann, S. (2003). The CASMIN Educational Classification in International Comparative Research. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-*

National Comparison (pp. 221–244). Springer US. https://doi.org/10.1007/978-1-4419-9186-7_11

Canty, A., & Ripley, B. (1999). *Boot: Bootstrap Functions* (Version 1.3-32) [Computer software]. https://doi.org/10.32614/CRAN.package.boot

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511802843

Demars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, (8), 69–82.

Deribo, T., Goldhammer, F., & Kroehne, U. (2022). Changes in the speed–ability relation through different treatments of rapid guessing. *Educational and Psychological Measurement*, 001316442211094. https://doi.org/10.1177/00131644221109490

Dollinger, S. (2013). *Diagnosegenauigkeit von ErzieherInnen und LehrerInnen: Einschätzung schulrelevanter Kompetenzen in der Übergangsphase*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-01660-9

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Ganzeboom, H. B. G., Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. In *OECD Education Working Papers* (133rd ed., Vol. 133).

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education*, *5*(18).

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

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. https://doi.org/10.1080/08957347.2016.1171766

Han, J., Chu, X., Song, H., & Li, Y. (2014). Social capital, socioeconomic status and self-efficacy. *Applied Economics and Finance*, *2*(1), 1–10. https://doi.org/10.11114/aef.v2i1.607

Hemmerechts, K., Agirdag, O., & Kavadias, D. (2017). The relationship between parental literacy involvement, socio-economic status and reading literacy. *Educational Review*, *69*(1), 85–101. https://doi.org/10.1080/00131911.2016.1164667

Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *40*(2), 267–284. https://doi.org/10.3102/0162373718759600

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, *62*(2), 147–177.

Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness*, *13*(1), 147–175. https://doi.org/10.1080/19345747.2019.1636437

LaFave, A. J., Taylor, J. A., Barter, A. M., & Jacobs, A. S. (2022). Student engagement on the National Assessment of Educational Progress (NAEP): A systematic review and meta-analysis of extant

research. *Educational Assessment*, *27*(3), 205–228. https://doi.org/10.1080/10627197.2022.2043151

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). https://doi.org/10.1186/s40536-014-0008-1

Lenhard, W., Lenhard, A., & Schneider, W. (2017). *ELFE II – Ein Leseverständnistest für Erst- bis Siebtklässler. [ELFE II – A Reading Comprehension Test for First to Seventh Graders]*. Hogrefe.

Leventhal, B. C., & Pastor, D. (2023). An illustration of an IRTree model for disengagement. *Educational and Psychological Measurement*, 00131644231185533. https://doi.org/10.1177/00131644231185533

Li, Y., Gao, M., Yu, Y., Zhang, S., & Yang, X. (2025). Influence of socioeconomic status on children's reading abilities: The mediating role of home learning environment and the moderating role of grade level. *BMC Psychology*, *13*(1), 848. https://doi.org/10.1186/s40359-025-03203-z

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, *10*, 1533. https://doi.org/10.3389/fpsyg.2019.01533

Lombardi, E., Di Dio, C., Castelli, I., Massaro, D., & Marchetti, A. (2017). Prospective thinking and decision making in primary school age children. *Heliyon*, *3*(6), e00323. https://doi.org/10.1016/j.heliyon.2017.e00323

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 104–142). Academic Press. https://escholarship.org/uc/item/61s3q2xr

Michaelides, M., & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: Country and item-type differences. *Psychological Test and Assessment Modeling*, *64*(3), 304–338.

Nagy, G., & Ulitzsch, E. (2021). A Multilevel Mixture IRT Framework for Modeling Response Times as Predictors or Indicators of Response Engagement in IRT Models. *Educational and*

Psychological Measurement, 001316442110453. https://doi.org/10.1177/00131644211045351

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12719

NEPS Network. (2025). *Scientific Use File der NEPS-Startkohorte 1: Neugeborene [Scientific Use File of NEPS Starting Cohort 1: Newborns]* (Version 12.0.0) [Dataset]. Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SC1:12.0.0

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online*, *5*(2), 80–109. https://doi.org/10.25656/01:8426

Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences*, *42*, 27–35. https://doi.org/10.1016/j.lindif.2015.08.002

Pfost, M., & Heyne, N. (2023). Joint book reading, library visits and letter teaching in families: Relations to parent education and children's reading behavior. *Reading and Writing*, *36*(10), 2627–2647. https://doi.org/10.1007/s11145-022-10389-w

Piaget, J. (2016). *Meine Theorie der geistigen Entwicklung* (R. Fatke, Ed.; H. Kober, Trans.; 4. Auflage). Beltz.

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement*, *50*(4), 447–468. https://doi.org/10.1111/jedm.12028

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-Scale Assessments in Education*, *9*(1), 10. https://doi.org/10.1186/s40536-021-00104-6

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rios, J. A. (2022). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement*, 001316442110036. https://doi.org/10.1177/00131644211003640

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education*, *9*(1), 18. https://doi.org/10.1186/s40536-021-00110-8

Rios, J. A., & Deng, J. (2024). A comparison of response time threshold scoring procedures in mitigating bias from rapid guessing behavior. *Educational and Psychological Measurement*, *84*(2), 387–420. https://doi.org/10.1177/00131644231168398

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, *27*(4), 356–373. https://doi.org/10.1080/10627197.2022.2110465

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. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, 1–31. https://doi.org/10.1080/15305058.2022.2036161

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1). https://doi.org/10.1002/j.2333-8504.2010.tb02218.x

Schmitt, M., Nusser, L., & Lorenz, C. (2022). Wie nehmen Eltern und pädagogische Fachkräfte die Kompetenzen ihrer Kinder wahr? *Zeitschrift für Erziehungswissenschaft*, *25*(5), 1293–1318. https://doi.org/10.1007/s11618-022-01088-x

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Schunk, D. H. (1995). Self-efficacy, motivation, and performance. *Journal of Applied Sport Psychology*, *7*(2), 112–137. https://doi.org/10.1080/10413209508406961

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335

Simon, A. J., Gallen, C. L., Ziegler, D. A., Mishra, J., Marco, E. J., Anguera, J. A., & Gazzaley, A. (2023). Quantifying attention span across the lifespan. *Frontiers in Cognition*, *2*, 1207428. https://doi.org/10.3389/fcogn.2023.1207428

Slobodskaya, H. R., & Kornienko, O. S. (2021). Age and gender differences in personality traits from early childhood through adolescence. *Journal of Personality*, *89*(5), 933–950. https://doi.org/10.1111/jopy.12624

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment*, *24*(4), 327–342. https://doi.org/10.1080/10627197.2019.1645592

Svetina Valdivia, D., Rutkowski, L., Rutkowski, D., Canbolat, Y., & Underhill, S. (2023). Test engagement and rapid guessing: Evidence from a large-scale state assessment. *Frontiers in Education*, *8*, 1127644. https://doi.org/10.3389/feduc.2023.1127644

Ulitzsch, E., Penk, C., Davier, M., & Pohl, S. (2021). Model meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment*, *26*(2), 104–124. https://doi.org/10.1080/10627197.2020.1858786

Van Loon, M., De Bruin, A., Leppink, J., & Roebers, C. (2017). Why are children overconfident? Developmental differences in the implementation of accessibility cues when judging concept learning. *Journal of Experimental Child Psychology*, *158*, 77–94. https://doi.org/10.1016/j.jecp.2017.01.008

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Welling, J., Gnambs, T., & Carstensen, C. H. (2024). Identifying disengaged responding in multiple-choice items: Extending a latent class item response model with novel process data indicators. *Educational and Psychological Measurement*, *84*(2), 314–339. https://doi.org/10.1177/00131644231169211

Welling, J., Zink, E., & Gnambs, T. (2026). Comparing different approaches of (not) accounting for rapid guessing in plausible values estimation. *Educational and Psychological Measurement*, 00131644251395590. https://doi.org/10.1177/00131644251395590

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*(3), 461–481. https://doi.org/10.1037/0033-2909.91.3.461

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

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, *32*(4), 325–336. https://doi.org/10.1080/08957347.2019.1660350

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5–6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

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

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

Wise, S. L., & Demars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment*, *26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Wise, S. L., & Kingsbury, G. G. (2022). Performance decline as an indicator of generalized test-taking disengagement. *Applied Measurement in Education*, *35*(4), 272–286. https://doi.org/10.1080/08957347.2022.2155651

Wise, S. L., & Kuhfeld, M. (2021). A method for identifying partial test-taking engagement. *Applied Measurement in Education*, *34*(2), 150–161. https://doi.org/10.1080/08957347.2021.1890745

Wise, S. L., & Ma, L. (2012, April 13). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper Presentation]. 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. https://doi.org/10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings.* (pp. 139–153). American Psychological Association. https://doi.org/10.1037/12330-009

Wise, S. L., Soland, J., & Bo, Y. (2020). The (non)impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57–77. https://doi.org/10.1080/15305058.2019.1605999

Xia, M., Poorthuis, A. M. G., & Thomaes, S. (2024). Children's overestimation of performance across age, task, and historical time: A meta-analysis. *Child Development*, *95*(3), 1001–1022. https://doi.org/10.1111/cdev.14042

Zhang, F., Jiang, Y., Ming, H., Ren, Y., Wang, L., & Huang, S. (2020). Family socio-economic status and children's academic achievement: The different roles of parental academic involvement and subjective social mobility. *British Journal of Educational Psychology*, *90*(3), 561–579. https://doi.org/10.1111/bjep.12374

Tables

Table 1

Sample Descriptives by Assessment

Grade	Domain	N_{person}	$N_{response}$	Response Time	Age	Gender	HCASMIN†	HISEI‡
				M (*SD*)	*M* (*SD*)	% female	% high	% high
2	reading	1,537	20,758	28.95 (20.55)	8.29 (0.13)	50.2	50.9	53.1
4	reading	1,470	29,296	19.75 (11.32)	10.23 (0.14)	50.7	49.7	50.9
2	math	1,512	29,366	54.68 (40.20)	8.29 (0.13)	50.5	49.9	52.5
4	math	1,500	23,396	58.92 (37.49)	10.23 (0.14)	50.5	49.5	50.5

Note. Response time is presented in seconds, and age in years. HCASMIN and HISEI are two different measures of parental socio-economic status.

† HCASMIN is the highest parental *Comparative Analysis of Social Mobility in Industrial Nations* index (CASMIN, Brauns et al., 2003), which classifies the highest educational qualification into nine categories ranging from 0 (no qualification) to 8 (university degree). It was dichotomized at the median (0 = no university degree, 1 = university degree; *Mdn* = 7).

‡ HISEI is the highest parental *International Socio-Economic Index of Occupational Status* (ISEI-08, Ganzeboom et al., 1992), which assigns a numerical score to each occupation ranging from 10 (e.g., garbage collector) to 98 (e.g., senior judge). It was dichotomized at the median (0 = low occupational prestige, 1 = high occupational prestige; *Mdn* = 73).

Table 2

Threshold Descriptives and Validation

Do.	Gr.	Method	Thresh.	Solution Behavior			Rapid Guessing		
			M (*SD*)	exp. % correct†	actual % correct	$r_{score.\ total}$	exp. % correct†	actual % correct	$r_{score.\ total}$
Re	**2**	NT10	5.64 (1.61)	52.3	21.0	0.25	20.0	23.3	0.33
		NT20	11.29 (3.21)	52.5	21.0	0.25	23.6	22.3	0.34
	4	NT10	2.02 (0.65)	89.2	25.0	0.22			
		NT20	4.03 (1.30)	89.4	25.0	0.21	34.9	25.0	0.50
Ma	**2**	VICS	2.70 (8.47)	66.1	21.1	0.26	14.6	13.9	0.48
		VIB	13.15 (15.33)	65.7	21.0	0.27	78.9	16.1	0.29
	4	NT10	2.43 (1.07)	79.5	25.0	0.29	20.7	25.0	0.05
		NT20	4.87 (2.13)	80.9	25.0	0.28	24.9	25.0	0.27

Note. Do. = domain, Gr. = grade, Thresh. = thresholds, exp. = expected, Re = reading, Ma = mathematics, NT10 = 10% normative threshold, NT20 = 20% normative threshold, VICS = visual inspection of proportion of correct scores, VIB = visual inspection of bimodality. Thresholds are reported in seconds.

† expected accuracy under random responding

Table 3

Prevalence of Rapid Guessing per Group

Grade	Domain	Rapid Guesses per Group (in %)					
		Girls	Boys	Low HCASMIN†	High HCASMIN†	Low HISEI‡	High HISEI‡
2	reading	2.6	4.0	4.4	2.3	4.7	2.2
4	reading	0.3	0.7	0.7	0.3	0.7	0.3
2	mathematics	0.8	0.9	1.0	0.7	0.8	0.9
4	mathematics	0.2	0.5	0.6	0.2	0.5	0.3

† HCASMIN is the highest parental *Comparative Analysis of Social Mobility in Industrial Nations* index (CASMIN, Brauns et al., 2003), which classifies the highest educational qualification into nine categories ranging from 0 (no qualification) to 8 (university degree). It was dichotomized at the median (0 = no university degree, 1 = university degree; *Mdn* = 7).

‡ HISEI is the highest parental *International Socio-Economic Index of Occupational Status* (ISEI-08, Ganzeboom et al., 1992), which assigns a numerical score to each occupation ranging from 10 (e.g., garbage collector) to 98 (e.g., senior judge). It was dichotomized at the median (0 = low occupational prestige, 1 = high occupational prestige; *Mdn* = 73).

Table 4

Results of Mixed-Effects Logistic Regression: Random Effects and Explained Variance

Model	$N_{persons}$	$N_{responses}$	ICC	$SD_{Random\ Intercept}$	Pseudo-R^2 (fixed effects)	Pseudo-R^2 (total)
Longitudinal Reading	1,839	50,054	0.89	5.27	0.03	0.90
Grade 2 Reading	1,470	20,578	0.94	7.19	0.01	0.94
Grade 4 Reading	1,537	29,296	0.88	4.81	0.01	0.88
Grade 2 Mathematics	1,512	29,366	0.30	1.18	0.07	0.35
Grade 4 Mathematics	1,500	23,396	0.95	8.00	0.02	0.95

Note. ICC = intraclass coefficient. Presented are the results of the mixed-effects logistic regressions with test-taking engagement (0 = rapid guess, 1 = solution behavior) as response-level dependent variable and student ID as person-level grouping factor. Pseudo-R^2 was estimated as described in McFadden (1974). Fixed effects are displayed in Table 5.

Table 5

Results of Mixed-Effects Logistic Regression: Fixed Effects

Model	**Regression Parameter**				**AME†**	
Predictor	Estimate	SE	*z*	*p*‡	Effect	SE
Longitudinal Reading						
Intercept	8.37	0.31	27.21	< .001		
Grade	2.05	0.13	16.04	1.000	0.02	0.01
Grade 2 Reading						
Intercept	7.81	0.40	19.60	< .001		
Gender	0.41	0.30	1.38	.083	0.01	0.01
SES	0.23	0.30	0.76	.223	0.01	0.01
Proficiency	0.26	0.08	3.19	.001	0.01	0.00
Grade 4 Reading						
Intercept	11.10	0.73	15.27	< .001		
Gender	0.72	0.70	1.04	.150	0.00	0.00
SES	0.05	0.68	0.08	.469	0.00	0.00
Proficiency	0.40	0.19	2.12	.017	0.00	0.00
Grade 2 Mathematics						
Intercept	5.33	0.19	28.59	< .001		
Gender	0.30	0.15	1.95	.026	0.18	0.10
SES	0.08	0.15	0.54	.293	0.05	0.09
Proficiency	0.48	0.07	6.71	< .001	0.23	0.05
Grade 4 Mathematics						
Intercept	11.50	0.78	14.70	< .001		
Gender	0.85	0.75	1.13	.130	0.00	0.00
SES	0.14	0.76	0.18	.429	0.00	0.00
Proficiency	1.13	0.39	2.91	.002	0.00	0.00

Note. AME = average marginal effect in percentage points, SES = socio-economic status. Presented are the results of the mixed-effects logistic regressions with test-taking engagement

(0 = rapid guess, 1 = solution behavior) as response-level dependent variable and student ID as person-level grouping factor. SES was operationalized as the dichotomized highest parental *Comparative Analysis of Social Mobility in Industrial Nations* index (HCASMIN, Brauns et al., 2003; 0 = no university degree, 1 = university degree). Gender was operationalized as boys (= 0) or girls (= 1). Random effects and explained variance are displayed in Table 4.

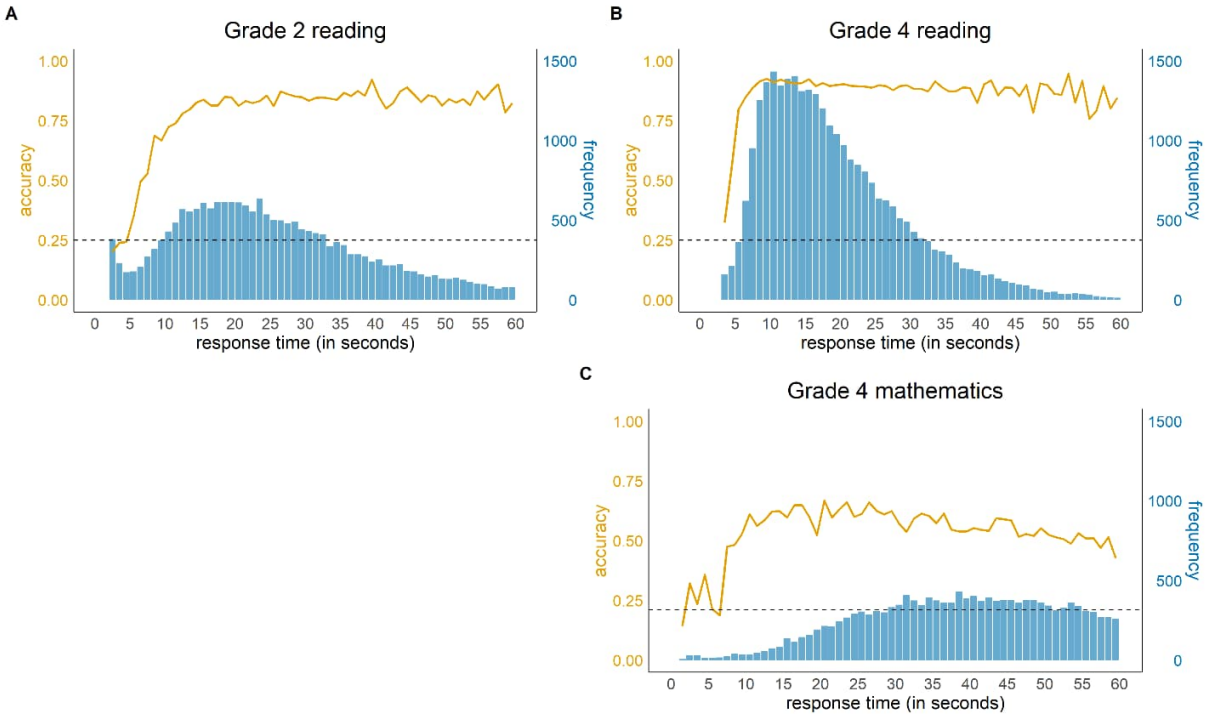
† Marginal effects are presented in percentages. They represent the change in the probability of the dependent variable when the predictor is increased by one unit.

‡ *p*-values of predictors refer to one-sided hypothesis tests, while *p*-values of intercepts refer to two-sided hypothesis tests; the significance level is $\alpha = .05$.

Figures

Figure 1

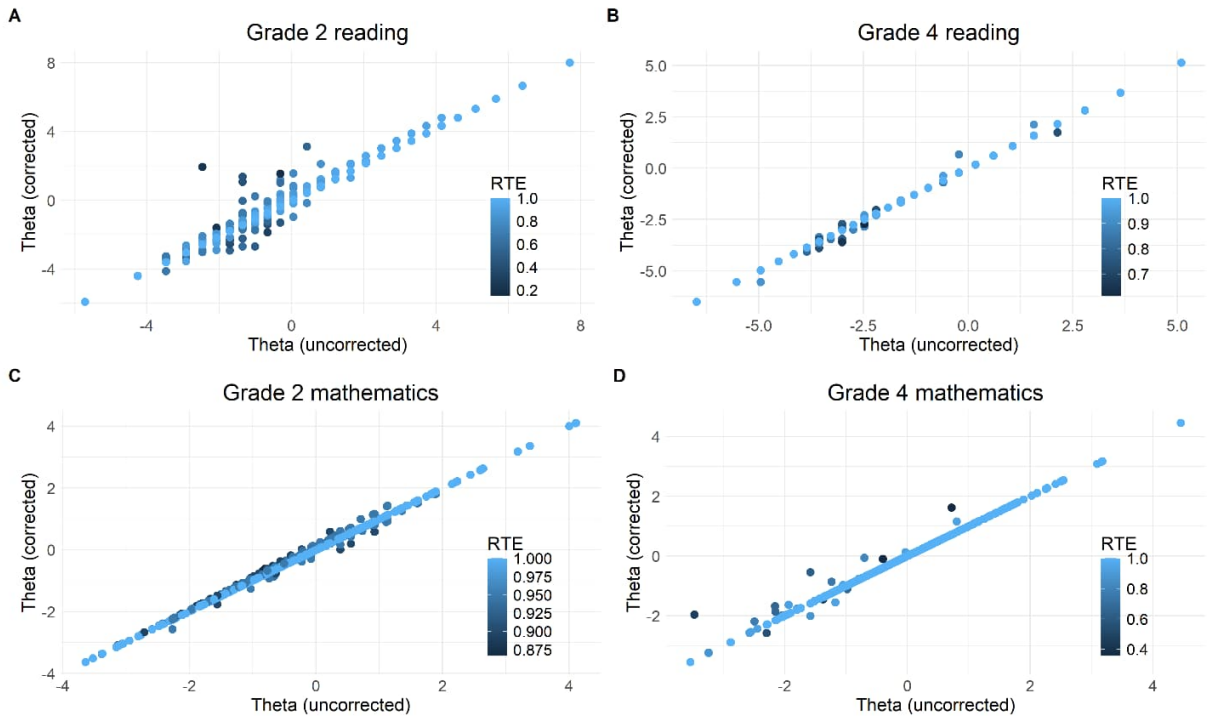
Accuracy Rate by Response Time



Note. Displayed are the accuracy rate (% correct, orange) and frequency of responses (blue) per response time bin (1 second) across all items of one assessment.

Figure 2

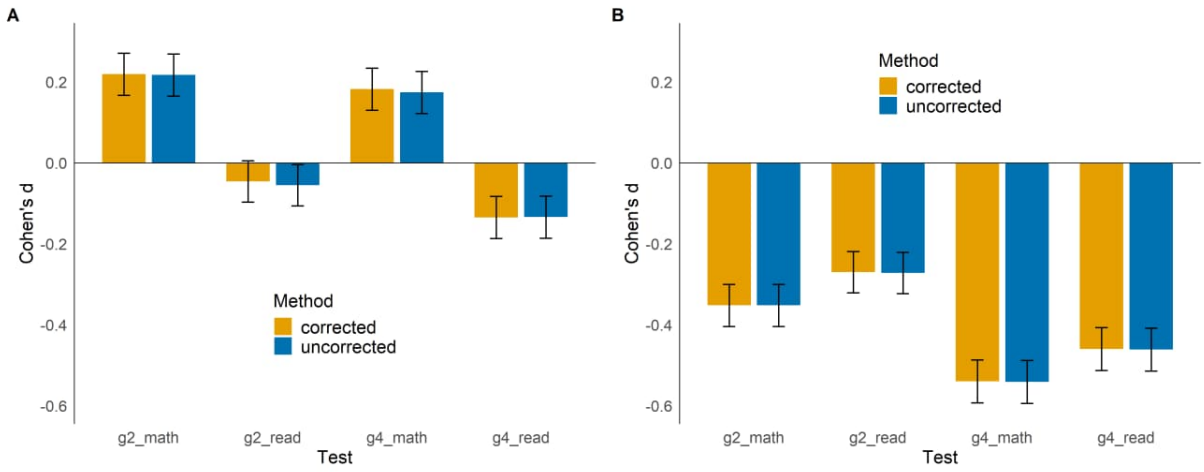
Comparison of Ability Estimates Between Models Ignoring or Accounting for Rapid Guessing



Note. RTE = response time effort. Displayed are the ability estimates of the models ignoring (x-axis) or accounting for (y-axis) rapid guessing. Response time effort is defined as the test-takers' average engagement across the test (1 = engaged on all items, 0 = rapid guessing on all items).

Figure 3

Estimated Ability Differences for Gender and Parental SES by Method



Note. g2_math: Grade 2 mathematics; g2_read: Grade 2 reading; g4_math: Grade 4 mathematics; g4_read: Grade 4 reading. Displayed are the estimated ability differences (Cohen's *d*) with standard errors for the model ignoring rapid guessing (blue) and the model accounting for rapid guessing by scoring all rapid guesses as missing (orange). **A**: Ability difference between boys (= 0) and girls (= 1); **B**: Ability difference between children without (= 0) and with (= 1) at least one parent with a university degree.

Author Contributions

Conceptualization – JW, KT and LF

Data Curation – JW

Formal analysis – JW and TG

Investigation – JW and LF

Methodology – JW, LF and TG

Supervision – TG

Visualization – JW

Writing – original draft – JW

Writing – review & editing – JW, LF and TG

Declarations

Conflicting Interests

No conflicting interests to declare.

Funding Statement

No funding was received.

Ethical Approval and Informed Consent Statements

All legal guardians of the participating children provided written informed consent before study enrolment. All participants could withdraw from the study at any time. The data collection procedures and instruments were approved by the data protection unit at the Leibniz Institute for Educational Trajectories in line with national ethical and legal regulations.

Availability of Data and Materials

This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network. The scored test data is provided at NEPS Network (2025), whereas the process data cannot be made publicly available due to legal restrictions. The computer code and analysis results are provided at https://osf.io/q2ebk/.

Preregistration

This study builds on the unpublished master's thesis submitted by Luis Fischer to the University of Bamberg in July 2025, which developed from an earlier project (preregistration: https://osf.io/z5wsd/). The present manuscript extends the original designs by incorporating additional data, new research questions, and refined analyses.