UMEÅ UNIVERSITY

Exploring and Modeling Response Process Data from PISA

Inferences related to Motivation and Problem-solving

Erik Lundgren

Department of Applied Educational Science
Umeå 2023

Doctoral Thesis
Department of Applied Educational Science
Umeå University
SE-901 87 Umeå

Academic Dissertations in Educational Measurement No. 15
ISBN: 978-91-8070-057-3 (print)
ISBN: 978-91-8070-058-0 (pdf)
ISSN: 1652-9650
Cover art and design by: Björn Sigurdsson

Printed by: Cityprint i Norr AB
Umeå, Sweden 2023

Contents

List of studies

The thesis is based on the following studies:

I. 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

II. Lundgren, E. (2022). Latent program modeling: Inferring latent problem-solving strategies from a PISA problem-solving task. *Journal of Educational Data Mining*, *14* (1), 46–80

III. Lundgren, E., & Eklöf, H. (2023). Questionnaire-taking motivation: Using response times to assess motivation to optimize in the PISA student questionnaire. *(Under Review)*

IV. Lundgren, E. (2023). Investigating satisficing in PISA 2018 questionnaire items by jointly modeling response times and subitem responses. *Manuscript*

The biggest obstacle to creativity,
is breaking through the barrier of disbelief

— Rodney Mullen

Abstract

This thesis explores and models response process data from large-scale assessments, focusing on test-taking motivation, problem-solving strategies, and questionnaire response validity. It consists of four studies, all using data from PISA (Programme for International Student Assessment).

Study I processed and clustered log-file data to create a behavioral evaluation of students' effort applied to a PISA problem-solving item, and examined the relationship between students' behavioral effort, self-reported effort, and test performance. Results show that effort invested before leaving the task unsolved was positively related to performance, while effort invested before solving the tasks was not. Low effort before leaving the task unsolved was further related to lower self-reported effort. The findings suggest that test-taking motivation could only be validly measured from efforts exerted before giving up.

Study II used response process data to infer students' problem-solving strategies on a PISA problem-solving task, and investigated the efficiency of strategies and their relationship to PISA performance. A text classifier trained on data from a generative computational model was used to retrieve different strategies, reaching a classification accuracy of 0.72, which increased to 0.90 with item design changes. The most efficient strategies used information from the task environment to make plans. Test-takers classified as selecting actions randomly performed worse overall. The study concludes that computational modeling can inform score interpretation and item design.

Study III investigated the relationship between motivation to answer the PISA student questionnaire and test performance. Departing from the theory of satisficing in surveys a Bayesian finite mixture model was developed to assess questionnaire-taking motivation. Results showed that overall motivation was high, but decreased toward the end. The questionnaire-taking motivation was positively related to performance, suggesting that it could be a proxy for test-taking motivation, however, reading skills may affect the estimation.

Study IV examines the validity of questionnaire composite scores assessing reading metacognition, using a Bayesian finite mixture model that jointly considers response times and sequential patterns in subitem responses. The results show that the relatively high levels of satisficing

(up to 29%) negatively biased composite scores. The study highlights the importance of considering response time data and subitem response patterns when evaluating the validity of scores from questionnaire items.

The thesis summary discuss the main findings of the studies from a validity perspective, emphasizing the use of response process data. Although further research is necessary to evaluate model accuracy, ideally against established ground truth criteria, the conclusion is that response process data can offer valuable insights into test-takers' motivation, problem-solving strategies, and questionnaire validity.

KEYWORDS: response processes, large-scale assessments, motivation, problem-solving, computational modeling, Bayesian modeling.

Sammanfattning

Datorbaserade prov gör det möjligt att samla in data om provtagarnas beteende när de utför uppgifter. Genom att analysera dessa data kan vi lära oss mer om hur deltagarna tänker och vilka mål som styr deras handlingar. Detta kan vara viktigt för att förstå provpoängens validitet, det vill säga huruvida provpoängen betyder det vi tror att den betyder. Ett av de områden där analys av responsprocesser kan vara särskilt användbar är för att undersöka provtagarnas motivation att göra sitt bästa i provsituationen. Det vanligaste sättet att undersöka testmotivation är genom självrapport, men analys av responsprocessdata skulle kunna ge kompletterande information. Ännu ett område som responsprocessdata kan ge information om är deltagarnas problemlösningsstrategier när de löser provuppgifter, vilket kan bidra med mer nyanserade tolkningar av provtagares svar. Avhandlingens mål har varit att undersöka hur responsprocessdata kan öka förståelse för prov- och enkätresultat genom att pröva och utveckla nya metoder för att bedöma provtagarnas motivation och dess effekt på poäng, samt att utforska och utveckla metod för att härleda problemlösningsstrategier.

Avhandlingen är en sammanställning av fyra delstudier där det empiriska materialet baseras på PISA-studien. Förhoppningen är dock att forskningen ska vara relevant även för andra prov och enkäter. I Studie I delades provtagare in i grupper med låg, medel och hög ansträngning baserat på klustring av responsprocessdata från en PISA 2012 problemlösningsuppgift. Resultaten visade att högre ansträngning var relaterad till högre provresultat i PISA i de fall då provtagare *inte* löste uppgiften, vilket givet uppgiftens design innebar att provtagaren hade *gett upp* att lösa uppgiften. Om provtagaren lyckades lösa uppgiften så fanns inget positivt samband mellan ansträngning och prestation. Resultaten har implikationer för hur testmotivation kan mätas, då de indikerar att det endast är tillförlitligt att mäta testmotivation från grad av ansträngning före provtagare ger upp, något som inte fungerar för provtagare som har lätt att lösa alla uppgifter, och något som är problematiskt då uppgifters design sällan ger information om huruvida provtagare har gett upp.

Studie II syftade till att undersöka om och hur problemlösningsstrategier kunde identifieras från responsprocessdata i en uppgift från PISA 2012. Metoden var att träna en textklassificeringsmodell med datorsimulerade strategier. Resultaten från studien visade att träffsäkerheten för

att klassificera problemlösningsstrategier var cirka 72%, men att träffsäkerheten kunde ökas till 90% om uppgiften modifierades. Uppgiften kunde lösas med flera strategier, vissa mer effektiva än andra. Provtagare vars beteende med störst sannolikhet passade in på en strategi som valde handlingar helt slumpmässigt uppvisade lägre provprestation kontra andra strategier. Resultaten tyder på att vi kan få en mer detaljerad bild av deltagares problemlösningsprocess, metoden är dock komplicerad och inte direkt generaliserbar till andra uppgiftstyper.

Studie III syftade till att uppskatta provtagares motivation att svara på bakgrundsenkäten i PISA 2018 genom att härleda motivationen från responstider på enskilda enkätuppgifter. Resultaten visade att de flesta deltagare hade hög motivation att svara på enkäten. Dock var vissa frågor, och frågor i slutet av enkäten, mer utsatta för icke-motiverade svar. Enkätmotivation var positivt relaterad till provprestation, ett resultat som tyder på att enkätmotivation kan vara en indikation på provmotivation. Dock kan estimering av enkätmotivation ha påverkats av vilket läsflyt provtagarna har, något som framtida studier bör undersöka.

Studie IV syftade till att skapa en modell som, baserat på responstider och svarsmönster, uppskattade andelen svar som resulterade från ytligt genomtänkta svar (eng. satisficing), och om sådana svar påverkade poäng på frågor som avsåg mäta medvetenhet om lässtrategier. Resultaten visade att mellan 15% och 29% av svaren på de studerade enkätfrågorna kan ha genererats genom en ytlig svarsprocess, och att andelen opålitliga svar var tillräckligt stor för att påverka poängen. Slutsatsen är att validitet i poäng från enkäten kan hotas av skenbart valida svar vilka inte kan upptäckas utan att beakta responstider och svarsmönster.

Sammanfattningsvis visar studierna att responsprocessdata från storskaliga prov ger värdefulla insikter om provtagres motivation, problemlösningsstrategier och enkätsvars tillförlitlighet, och att genom att beakta responsprocessdata kan validiteten i vår tolkning av provresultat öka. Resultaten tyder inte på att självrapporterade motivationsupskattningar bör ersättas av motivationsuppskattningar baserade på responsprocessdata, snarare att de har kompletterande funktioner. Den mest praktiska tillämpbara användningen av responsprocessdata bedöms vara i samband med bakgrundsenkäter. Trots att responsprocessdata ger detaljerade observationer är de inte kompletta beskrivningar av provtagares beteenden, en begränsning som måste beaktas. Betänk dock även

att detaljerade beskrivningar av beteenden sällan otvetydigt kan avslöja den underliggande kognitionen. Mer forskning behövs för att utvärdera hur väl slutsatser från modeller över responsprocessdata överensstämmer med exakta sanningskriterier, och om sådana saknas, med modeller baserade på mer detaljerade observationer.

Preface

There are many people that I would like to acknowledge; without them, this thesis would have been different or not possible at all. First, I want to thank my main supervisor, Hanna, and supporting supervisors, Pecke and Inga, for guiding me along the way while allowing me to pursue some of the more adventurous ideas I've had. Also, thanks to all the doctoral, post-doctoral, and visiting researchers at the department whom I've met and discussed with during my studies: Sofie, Anette, Marcus, Daniel, Jonathan, Jonatan, Theodora, Anna, Susanne, Militsa, Peter, Anders, Gavin, and Michalis. Björn, thank you for designing the cover of the thesis, and thank you, Marianne, for letting me use your photo of Joksgaejsie and Ryjvejegaejsie (which can be glimpsed in the background of the front cover). Thanks to Linus, who, during my master's thesis, introduced me to the magical and powerful world of computer programming. Irina, thank you for interesting discussions related to Bayesian modeling and psychometrics. Big up to everyone who makes learning material available online, be it course material, blog posts, walkthroughs, source codes, etc.; this work would not have been possible without it. I want to thank my doctoral study colleagues, Maria, Andreas, and Pär, with whom I travelled by train to Ljubljana for a PhD meetup. It was quite a journey! (Speaking of which, I want to thank the hosts at Hostel Bohinj who, after I had summited Triglav, rehydrated me and drove a very tired me to the bus so I managed to catch my train back home.). I want to thank all skateboarders in Umeå and its vicinity with whom I have met and rolled during my years at Umeå University! Always grateful to the mountains for providing both sanctuary and recreation. A shout-out to all my friends in Vilhelmina! I want to thank my family, my parents Gudrun and Lars Lundgren, farmor Birgitta, and my siblings Petter and Ella. Finally, Erika, thank you for always believing in me and supporting me during this time!

Umeå, April 2023
Erik Lundgren

1 Introduction

Increasingly, educational tests are administered via computers. For example, International assessments like PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) are already digitalized and parts of the Swedish national tests are set for computer administration in 2024 (Skolverket, 2023). Computer-based assessment allows for the collection of response process data in the form of logs of inputs test-takers give computers while solving tasks, data that could provide insights into test-takers' behaviors and thought processes. Although this data can be complex and messy, it offers valuable context for interpreting test scores. Despite its importance, validity evidence from response processes has rarely been utilized due to practical challenges in data collection and analysis (Ercikan & Pellegrino, 2017; Zumbo & Hubley, 2017). As we have transitioned from a digital desert to a digital ocean (Behrens & DiCerbo, 2014), we can now take a dive into the depths and explore some areas of test-takers' response processes relevant to the validity of assessments.

As Robert Mislevy (2018) points out, when a student acts in an assessment situation, among other things, "affect and motivation play significant roles. They vary with stakes, time pressure, and the purpose of the assessment from the examinee's point of view" (p. 105). High-stakes assessments typically have clear incentives for test-takers (influencing final grades, authorizing driver's licenses, etc.), while low-stakes assessments, such as PISA, lack direct consequences. A common concern in low-stakes assessment contexts is that test-takers are not motivated to spend effort and do their best on the assessment. Research shows a positive relationship between test-taking motivation and test performance (Cole et al., 2008; Eklöf, 2010; Eklöf & Nyroos, 2013; Eklöf et al., 2014; Knekta & Eklöf, 2015; Penk et al., 2014; Penk & Schipolowski, 2015; Silm et al., 2020; Wise & DeMars, 2005), with score validity potentially improved by monitoring attentiveness (Wise et al., 2006), providing external incentives (Gneezy et al., 2019), or increasing perceived relevance (Rios, 2021). The influence of varying motivation levels on test scores implies that scores capture factors beyond the intended construct, raising a validity concern.

Test-taking motivation, often interchangeable with terms like test-

taking effort, engagement, etc., refers to striving for the best possible performance in a test situation. While increased effort can signal higher motivation, it's sometimes important to differentiate between effort and motivation, as the effort-motivation relationship does not necessarily need to be in direct proportion to each other. Consider for example that attempts to achieve the same goal can require varying amounts of effort due to factors like skill and knowledge. Motivation is the desire to achieve a goal, effort the work invested to reach it.

Test-taking motivation is typically assessed through post-test self-reports (Rios, 2021; Silm et al., 2020) while offering insight into test-takers' subjective experiences can be influenced by memory limitations, differing interpretations, or social norms (Schwarz, 2007). Results from Schmeck et al. (2015) show that delayed effort and difficulty ratings (ratings made after completion of an entire test) was higher than immediate ratings (made directly after each problem), however self-reports of interest and motivation was unaffected by delay (Schmeck et al., 2015). Also, test-takers may conflate effort ratings with perceived performance, which has been demonstrated to affect subjective effort ratings (Moore & Picou, 2018), and results from a study comparing self-reported and behavioral measurements of effort show discrepancies between the modes of measurements (Apascaritei et al., 2021). To address self-report limitations, response process data can provide a more direct assessment of motivation by capturing behaviors *in situ*, with the additional benefit that it enables continuous "stealth" assessments (Shute, 2011) requiring no additional work for students.

Despite the prevalence of self-report in social sciences, there is limited research on theory and modeling of response processes data related to self-reports (Launeanu & Hubley, 2017). In large-scale assessments like PISA, motivation to answer background questionnaires is almost as important to consider as motivation for cognitive tasks, particularly since questionnaire information include students' background variables and their perceptions and attitudes related to school, information that influences the creation of plausible value performance estimates (Mislevy, 1991; OECD, 2017; von Davier, 2013). Analyzing questionnaire response process data is a valuable yet understudied area, with potential implications for score validity at both on the level of test items and on aggregated levels.

Besides the case of motivation, response process data from large-scale assessments could increase our understanding of problem-solving strategies, offering insights beyond binary item scores, which has been shown by Greiff et al. (2015). While PISA states that they used response process data to track strategies, and that this in some cases influences scoring, e.g. when it could be determined that students guessed the right answer no credit was given, it is however not much further expanded upon (OECD, 2014). Expectations on analyses of log-file response process data are that they will reveal differences among problem solvers that cannot be fully accounted by the test scores, which points toward interesting uses related to score validity. However, according to Hubley and Zumbo (2017), most research on response processes have mostly descriptive rather than explanation-based which is a limitation.

One strand of problem-solving research pioneered by Newell and Simon in their book *Human Problem Solving* (Newell & Simon, 1972) is to use computer programs, or computational models, to explain, understand, and analyze the cognition underlying problem-solving. Such models are explanatory in the way that they sufficient to generate and simulate behaviors regarding some task while modeling underlying aspects of cognitive processes. Such models require rigorous clarity in expression of assumptions since the theories need to be specified so precisely that they can be implemented as computer programs, see Stewart (2007) for discussions and philosophy related to computational cognitive modeling. Computational models have been suggested to be useful for assessment and item development, evaluation of validity, automated scoring, and evaluation of human raters (Moon et al., 2018). Use of computational modeling in educational context includes for example LaMar (2018) and Rafferty et al. (2020), Rafferty et al. (2015), though more often research on response processes has used data-driven methods (see, e.g., Han et al., 2022; He, Borgonovi, & Paccagnella, 2019; He, Liao, et al., 2019; Qiao & Jiao, 2018; Tang, Wang, et al., 2021; Tang, Zhang, et al., 2021; Ulitzsch et al., 2021; Zhu et al., 2016). Applying computational modeling approaches to large-scale assessments is a novel application where existing research is scarce, and since more theoretically grounded and explanation-based models are asked for, it is an important area to explore.

Since the empirical data that was used in this thesis comes from

PISA studies, a brief introduction to PISA is appropriate. Starting in 2000, PISA has assessed 15-year-olds every three years, to measure their skills in reading, math, and science, as well as other innovative areas like problem-solving, global competence, etc. The PISA 2018 study included about 700,000 students from 79 countries (OECD, 2019). The content of the test is determined by the OECD and its partners, test items are designed by various contractors and the test is administered by national organizations. Students also complete a questionnaire to provide additional information on demographics and education-related factors. According to the PISA 2018 Technical Report, the test takes around 80 minutes on average, followed by a 35-minute questionnaire. PISA uses an incomplete multiple-matrix sampling design where different combinations of test items are administered to students to efficiently cover as much content as possible. Due to this design PISA uses a scoring approach where multiple proficiency scores called plausible values[1] are calculated for each student to account for the uncertainty that arises from the incomplete and missing data.

PISA is a low-stakes test for individual students, with no direct consequences feedback given on their performance. Although PISA doesn't impact individual students, its results can affect future educational systems and policies. Many studies focus on PISA's policy impact (Hopfenbeck et al., 2018), and while assessing its effects is challenging (Pons, 2017; Rutkowski et al., 2020), PISA is perceived to influence policymaking (Breakspear, 2012), used to legitimize existing policies (Waldow, 2009), and provide a description of the current state of educational system which policymakers depart from (Pettersson, 2008). Besides the possible policy implications, data from PISA is used in various educational research studies. So, any research related to examining the validity of data and scores from large-scale assessments can be argued to be of importance.

[1] draws from a posterior distribution of a multivariate latent regression Item Response Theory (IRT) model

1.1 Aim

This thesis aims to explore and model response process data from a large-scale assessment to better understand test-takers' effort, motivation, problem-solving strategies, and potential validity implications. The general research questions are:

- How can response process data be used to infer test-taking motivation?
- How can response process data be used to infer problem-solving strategies?

More specific questions include:

- What is the relationship between motivation and test performance?
- What is the relationship between motivation inferred from response process data and self-reported effort?
- How does insufficiently motivated responses (satisficing) impact questionnaire scores?
- How accurately can problem-solving strategies be inferred?
- Are some problem-solving strategies more efficient than others?
- Do different problem-solving strategies relate to differences in test performance?

The thesis is organized as follows: Section 2 presents theories related to validity, motivation, and problem-solving. Section 3 provides information on the samples, analyses, models, and parameter estimation. Section 4 includes a short summary of each study. Finally, Section 5 concludes with a discussion of the main findings, validity implications, and limitations.

2 Theory

2.1 Validity theory

Assessments in the social and behavioral sciences aim to provide information about *something* often referred to as a *construct*, which in the educational sciences is often related to knowledge, skills, and competencies in some domain. The key questions surrounding assessments involve *validity* — whether the test measures the intended construct and if taking action on the information obtained from the assessment leads to desired outcomes. According to The Standards for Educational and Psychological Testing (Standards, AERA et al., 2014): "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p.11). Mislevy reformulates this definition: "A test is valid for a given interpretation or use to the degree to which empirical evidence and theoretical rationales support reasoning *as if* "it measures what it is purported to measure" (Mislevy, 2018, p. 202) highlighting the use of models indicated by the "as if" modifier. Models are simplifications of reality, which is often what makes them useful[2]. From a philosophical standpoint, a constructivist-realist approach to assessments and validity is in my mind reasonable (Messick, 1989; Mislevy, 2009, 2018). This approach boils down to *realist* in the sense that behaviors, cognitions, and attributes could exist and (at least in principle) be possible to understand, and *constructivist* such that researchers create models that, while not capturing the full reality, help us comprehend and interpret the world.

The Standards mention various sources of validity evidence, including: test content, response processes, internal structure, and relation to other variables. As the present thesis focuses on validity based on response processes, this type of validity evidence will be discussed more closely.

2.1.1 Validity evidence based on response processes

2.1.1.1 What are response processes? There is no universally agreed-upon definition of response processes in the context of educational assessments. While most definitions of response processes include both behavioral and cognitive processes (Bergner & von Davier, 2019; Cizek,

[2]Consider, for example, a world-sized map attempting to represent everything in the world precisely and completely; it would be quite difficult to use for navigation.

2020; Ercikan & Pellegrino, 2017; Hubley & Zumbo, 2017), some definitions put more weight on behavioral processes (AERA et al., 2014; Ercikan & Pellegrino, 2017; Russell & Hubley, 2017) while other on the underlying cognition (Bergner & von Davier, 2019; Hubley & Zumbo, 2017). Launeanu and Hubley (2017) highlights the distinction between *observed processes* and *inferred response processes* which indicate that the interest is on latent cognitive processes, states, etc., and not their overt actions, similar to how (Zumbo et al., 2023) suggest we not conflate response process with response process data. Departing from the previously outlined definitions, I have settled on the following: response processes refer to the underlying *cognition* that coordinates test-taker behavior in the assessment situation. This is also what I believe most other definitions cover, either explicitly or implicitly. This definition implies that response process data is any data that can inform about the cognitive process. However, as Levy (2020) points out: even though response process data is often contrasted against *product data* (e.g. right/wrong answer), the distinction is fuzzy since one assessor's product data might be another's process data depending on their construct definitions.

On the value of response processes for validity, Messick (1989) discussed how cognitive psychology has developed methods to analyze processes underlying item and task performances, which could be valuable in construct validation. Markus and Borsboom (2013) suggest that attempts to model and investigate item response processes will be useful for validity since it will force the investigator to explain the link of how the construct cause item responses, which in turn relates to Mislevy's (2018) reasoning about how an understanding of the lower-level processes can inform and constrain inferences about higher-level construct interpretations. In summary, thinking about response processes helps us to think more clearly about the link between construct and observations in the test-situation.

Against this background on response processes and validity, it is surprising that the Standards (2014) take the view that evidence about response processes is only crucial to validity arguments when test developers make explicit claims about response processes, while in many other cases they are not. Contrary, Newton (2019) writes that despite Standards' unfortunate wording, response processes can offer crucial information even when no claims in the assessment are based on the response

processes. For example, if response process data would indicate that a response was a blind guess, this should alter a score's meaning and validity. Furthermore, Newton argues, the value of response processes for validation depends on cost-effectiveness and how well micro-validations from response processes can be integrated into the overall assessment's macro-validations.

There are many different methods of gathering response process data such as cognitive interviews, think-aloud protocol (Ericsson & Simon, 1984), video-ethnography (Maddox, 2017), eye-tracking (Krstić et al., 2018), and response times (Li et al., 2017). Regarding large-scale assessments and PISA, in 2012 more detailed time-stamped log-file were released. However, all publicly released PISA data has since only included aggregated response process data on total numbers of actions, and total response time on items. Still, this data could be regarded as response process data since it could provide key information on test-takers way of responding not reflected by the raw or scored item responses.

2.1.1.2 Examples of using response processes as validity evidence

There exist numerous proposed examples of uses of response process data and models of them: validating score interpretations for both trait and process constructs (Kane & Mislevy, 2017); assessing complex thinking in history (Nichols & Huff, 2017); understanding score meaning in the context of learning disabilities (Tindal et al., 2017); investigations of validity of multiple language versions of tests (Solano-Flores & Chıa, 2017); assessing misunderstanding in math (Rafferty et al., 2020); and understanding tool use in digital assessment platforms (Jiang et al., 2023). In the case of large-scale assessments such as PISA, applications related to response processes have for example been using response times to address issues of rapid guessing, and engagement (Michaelides et al., 2020; Pools & Monseur, 2021; Schnipke & Scrams, 1997; Ulitzsch, Pohl, et al., 2022; Ulitzsch et al., 2020; Wise, 2017). Behavioral data from log-files has also been used to provide a better understanding of: score meaning and successful or unsuccessful strategies in problem-solving items, e.g. VOTAT (Greiff et al., 2015); differences in information processing strategies by high- and low performers (Hu et al., 2017); navigational patterns in reading tasks (Hahnel et al., 2022; He et al., 2022); problem-solving behavior in PIAAC (He, Borgonovi, & Paccagnella, 2019); the influence

of mindset on learning (Holden et al., 2021); fairness of group comparisons (Ercikan et al., 2020); and early prediction of failure on interactive tasks Ulitzsch, Ulitzsch, et al. (2022).

2.1.2 An argument-based approach to validity

This thesis will discuss the validity implications of process data using an argument-based approach (see, Bachman, 2005; Cronbach, 1988; Kane, 2012; Mislevy, 2018; Mislevy et al., 2003). The argument-based approach can be seen as an application of Toulmin's (2003) general analysis of arguments to the case of arguments about constructs that are made in assessments. The following paragraph presents a slimmed-down version of Mislevy's (2018) approach, closely aligned with Toulmin's (2003). This approach was chosen since it requires fewer components to illustrate how response process data can impact assessment arguments.

According to Toulmin, an argument comprises components such as data, claim, warrant, backing, alternative explanation, and evidence for the alternative explanation. A claim (C) is an asserted conclusion, for example "Maggie has a Macintosh," supported by data (D) say "Maggie's laptop has a 'command key' ⌘." The warrant (W) connects data to the claim, e.g., "Only Macintosh laptops have command keys". Backing (B_W) provides evidence for the warrant, say empirical observations from many thousands of laptops. The argument seems valid unless an alternative explanation (AE) is available, for example: "Maggie added a command key sticker over Windows keyboard button" which in turn also requires backing evidence (B_{AE}), say revealing the original keyboard button symbol beneath the sticker. In assessments, the argument structure is analogous: A claim is made about a construct, such as a test-taker's competence in some domain. Data is generated by a person in an assessment situation, like answering multiple-choice items. The warrant justifies that the performance reflects the construct, for example, knowledgeable students answer such items correctly. The backing for warrants can stem from research, experience, expert knowledge, or reasoning. The extent to which claims about constructs are valid depends on the warrants, their backing, and ruling out any alternative explanations that could explain the test performance. See Figure 1 for a diagrammatic depiction of how the argument flows from data to claim.

Toulmin discusses cases of uncertain arguments where an analytic

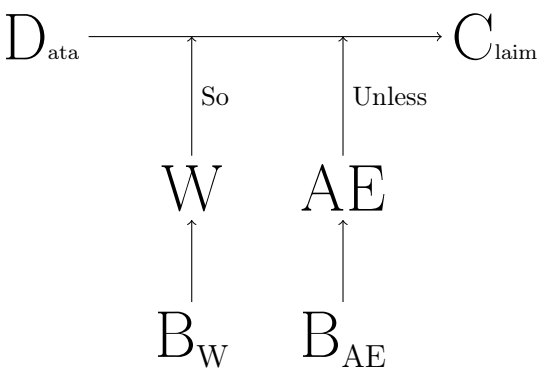

Figure 1: A Toulmin style depiction of an argument

conclusion is unattainable. In these cases qualifiers are used to modify the argument's weight, such that in ambiguous situations, the data and warrant *probably* or *almost certainly* lead to the conclusion. For instance, "Maggie's keyboard has a command button (P), so →, she almost certainly (Q) has a Macintosh (C)". Assessment arguments can similarly incorporate qualifiers through probability theory, as Mislevy (2018) suggests, this transforms claims into probabilistic statements in which the weight of an argument can be adjusted by the evidence.

One particular threat to validity is *construct-irrelevant variation* which involves variance in the claim caused by factors unrelated to the construct of interest. The threat can be linked to alternative explanations for performance, such as variance in score due to lower level of motivation, which would indicate construct-irrelevant variation (given that level of motivation was not considered part of the construct). Regarding response processes and validity, in Mislevy's argument-based approach, one case is response process data supporting the claim about the construct. However, if not part of the claim response processes data falls into Mislevy's *other information*, which, although often not explicitly included in measurement models, plays a critical, hidden, role in assessment reasoning and can lead to alternative explanations for performances. In this thesis, a recurring theme of alternative explanations is related to the motivation of test-takers, a concept that is introduced in the next section.

2.2 Motivation

One of the originators of theories related to motivation is Kurt Lewin (Weiner, 2013), who in the educational context has influenced contemporary motivational theories such as the expectancy-value theories (Wigfield et al., 2009). Lewin's original theorizing is relevant to the present thesis since it focuses on motivation in immediate situations which becomes useful when considering motivation within a test situation.

Lewin (1938) discusses psychological force (motivation) as comprised of three components: *direction*, *strength*, and a *person*[3] whom the motivation pertain to. Lewin's conceptualized the motivational forces within a mathematical space he called hodological space. For the present purposes only the conceptual parts will be needed to explain assumptions and rationale for measuring test-takers' motivation with response process data. One of Lewin's suggested methods of measuring psychological forces is by opposing forces. One suggestion is that in the case of impassable barriers (some obstacles that cannot be overcome to enter a region), a person's psychological force can be measured by the time spent or attempts made before giving up trying to pass the barrier. The greater the force towards a goal (and the barrier that lies in the way), the greater the restraining force provided by the barrier. Withdrawal from attempts to pass the barrier indicates that making attempts have become negatively valenced, or the goal region has lost its positive valence. Withdrawal can also happen due to a person believing that the goal is unachievable with any of the available means. In Study I, test-takers attempted a PISA problem-solving task (see, Figure 2) that was designed such they they could know when the task was solved and the goal state was reached. For some, the problem was a passable barrier, while for others, it was impassable. With impassable barriers, the force toward the barrier was proportional to motivational force, whereas with passable barriers we only know that the force applied was enough to overcome the barrier, which is not likely to reflect the maximum amount of force that the person could have applied if the barrier was stronger. Lewin also proposes measuring psychological force by *restraining forces which permit locomotion* by considering the friction an individual is willing to overcome at each step toward a goal region. Overcoming greater friction indicates a higher motivational force. Study III utilizes the restraining forces ap-

[3]Lewin uses the term it *point of application*

proach. Each questionnaire item is a step allowing progress towards goal of completing the questionnaire, at each step test-takers can choose either low-friction satisficing responses or higher-friction optimizing responses. The more optimizing responses that were used to move through the questionnaire, the greater the motivational force applied.

Test-taking motivation is often defined in terms of effort, referring to a student's engagement and energy expenditure towards achieving the highest possible score (Wise & DeMars, 2005), and researchers have suggested using Response Time Effort (RTE) to assess motivation of test-takers, categorizing test-takers as engaged in *solution behavior* or engaged in *non-solution behavior* depending on whether their response times fall below a threshold that indicates a rapid guessing response (non-solution behavior) (Wise, 2017; Wise & Kong, 2005). A related approach has been to use mixture models to represent response times generated by solution and non-solutions behaviors (Lu et al., 2021; Nagy & Ulitzsch, 2022; Schnipke & Scrams, 1997; Ulitzsch et al., 2020). Test-taking motivation can be understood within Lewin's theory as motivation toward a region of having achieved a maximum test score, which drives solution behaviors.

2.2.1 Motivation in surveys

One aim of this thesis is to gauge test-takers' motivation to complete the background questionnaire. Using the terminology and theory related to "test-taking motivation" may in this case be misleading, as questionnaires do not typically have right or wrong answers, and guessing does not imply the same thing as in knowledge tests. Thus, for this *questionnaire-taking motivation*, a relevant theory to apply is that of satisficing and optimizing in surveys (Krosnick, 1991). *Optimizing* involves respondents providing high-quality data by thorough deliberation, while *satisficing* involves skipping necessary cognitive steps, producing seemingly valid answers that might not represent the responders opinion. Krosnick identifies task difficulty, respondent ability, and motivation as the determinants of satisficing. For surveys, defining task difficulty and ability can be challenging, and as questionnaires should be easy to understand and answer for all test-takers, it is assumed that these factors remain constant among respondents. Thus, motivation becomes the number one suspect of driving satisficing, and since each item is of

equal difficulty they provide equal friction, which allows for comparable measurements of test-takers motivation. In the present work, satisficing and optimizing response processes are inferred from response times on questionnaire items using models. It's important to note that response times also include motor behavior time for communicating responses, an aspect not explicitly addressed by Krosnick. Human reading speed ranges around 250-300 words per minute (Brysbaert, 2019), mainly limited by eye movement speed (Primativo et al., 2016). Visuomotor processes like moving a mouse pointer and clicking take about 1 second. Using these basic process times, and knowing that optimizing responses require reading, decoding, and communicating answers, we can get an idea of the time this processes could take. As satisficing involves only simple visuomotor actions, of pointing and clicking, it should take less time. We can see parallels in satisficing and optimizing responses to solution and non-solution behavior in RTE terms, with rapid guessing as a type of satisficing response and solution behavior as an optimizing response. Questionnaire-taking motivation, as measured in Study III, is operationalized as the probability of a test-taker to engage in an optimizing response style as indicated by models over the optimizing response times relative to models over response times generated by models of rapid satisficing and idle satisficing. As such the approach is similar to Wise's RTE, in that the estimate is based on response times, however it is in many ways different in that models over response times are used and the inferred parameter reflect questionnaire-taking motivation.

2.3 Problem-solving

The problem-solving theory used in this thesis is inspired by Allen Newell (Newell, 1994), incorporating elements from basic AI problem-solving agents (Russell & Norvig, 2020). Newell identified problem-solving as a *search* process in a *problem space* containing *operators* which can be applied to in attempting to transform a current situation into a desired one. An agent starts in state S_{start} and aims to reach a goal state S_{goal}. The *task environment* outlines actions A that the agent can perform and the effects they have. Chess is an example of a *well-defined problem* with a clear goal state and action rules, while other problems like writing novels are not as well defined. The agent searches for solutions, often using prior knowledge to narrow down which possible actions to

try. Agents prioritize to take actions leading to favorable future states given the state current state S they are in, defined by an evaluation function or heuristic $h(S, A)$. For completeness, an agent is often able to perceive the environment by some means and has effectors (muscles) for executing actions and manipulating external objects. In Study II, which focuses on inferring test-takers' problem-solving strategies, strategies are understood as identical to the evaluation functions used while searching for solutions. For further information about human problem solving see Newell and Simon (1972) and Simon (1996).

To aid understanding of these terms and how the problem-solving process is viewed in the present work, I will exemplify with the help of the problem-solving task analyzed in Study II, (see Figure 2). The task presents a map, representing a road network graph with nodes (suburbs) and edges (travel connections) labeled with travel times. The student receives the following prompt:

> Here is a map of a system of roads that links the suburbs within a city. The map shows the travel time in minutes at 7:00 am on each section of road. You can add a road to your route by clicking on it. Clicking on a road highlights the road and adds the time to the Total Time box.
>
> You can remove a road from your route by clicking on it again. You can use the RESET button to remove all roads from your route.
>
> ...
>
> Maria wants to travel from Diamond to Einstein. The quickest route takes 31 minutes.
> Highlight this route. (OECD, 2014, p.41–42)

Suppose the test-taker, our problem-solving agent, aims to solve the task. The starting state has no highlighted edges, and the possible actions are to highlight edges, click the reset button, and submit the solution. The PISA Traffic task features 16 nodes and 23 edges. The edges can be either highlighted or unhighlighted, thus the *task environment* can be configured in $2^{23} = 8388608$ unique states, only one of which is the goal state. As the correct solution is not immediately apparent or available to retrieve from memory, the test-takers must search within a problem space, where they can narrow down possibilities using their

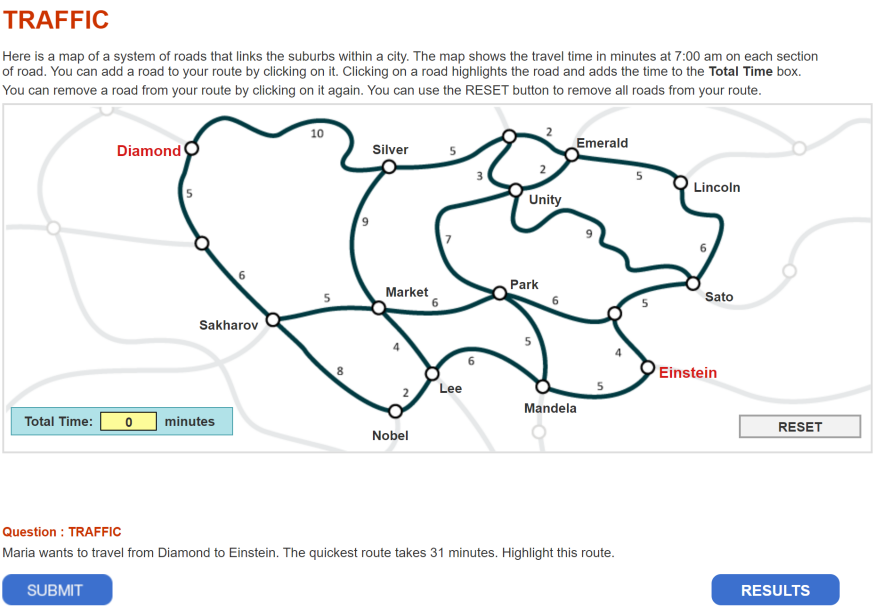

Figure 2: The PISA 2012 Traffic Item 2 (CP007Q02)

subjective understanding and knowledge. Luckily, the Traffic task is connected in such a way that there "only" exist 95 unique paths between Diamond and Einstein, and due to the way our perceptual system works, it is quite easy for test-takers to identify such paths. It could however be difficult to remember and discriminate between them exactly, leading to repetitions of paths already attempted. Even though the task can be solved by blind trial and error (on the level of paths), some strategy is likely to be followed by test-takers. Study II explores what such strategies could be and how they could be identified.

3 Methods and methodological choices

3.1 Data source and samples

The empirical data used in this thesis comes from the PISA study developed by the Organisation for Economic Co-operation and Development (OECD). Some domains in PISA became computer-based in 2012, and since 2015, it has been fully computer-based for most participating countries. From PISA 2012, response process data were gathered in the form of time-stamped logs of within-item actions. However, since 2015 and onward, only aggregated total response times and total numbers of actions on items have been released publicly. The kind of data provided by PISA, including log-files and response times, together with it being a low-stake assessment, make it a suitable empirical data source for exploring methods of using response process data to understand test-takers' motivation.

For Study I and Study II, log-file data from a PISA 2012 problem-solving item called Traffic (code: `CP007Q02`) was used as empirical data. Study III and Study IV used response time data from the 2018 student questionnaire. The samples for each study were as follows: Study I: 3,231 participants from Denmark, Finland, Norway, and Sweden. Study II: no additional restrictions besides passing data cleaning procedures resulted in 24,859 observations from the entire PISA sample. Study III: 5,124 Swedish test-takers on a selection of 61 out of the 79 items from the questionnaire. Study IV: Metacognition items `ST164`, `ST165`, and `ST166`, with Swedish test-takers, sample sizes N = 5,407, 5,400, 5,387 for each item respectively. Data can be found on the PISA website.

3.2 Bayesian data analysis and modeling

Bayesian inference aims to update prior beliefs with observed data to posterior beliefs in a logically consistent way. The "beliefs" refer to the probability estimates of parameter values in models. I will provide a very brief introduction and exemplification of Bayesian inference and then provide some arguments for why I chose this approach.

In the most basic sense, Bayesian inference requires four components: a *parameter* of interest, a *prior* probability distribution over the parameter values, *data* that can inform the value of the parameter, and a *likelihood* or observational model that defines how different parameter

settings relate to the data. To explain this, let's consider a small example related to a simple model in a hypothetical assessment scenario. Suppose we have some data y in the form of right and wrong scores in an assessment. These are integers, either 1 for a correct response or 0 for an incorrect response. We have a parameter of interest θ, which reflects the probability that a test-taker will solve an item in the assessment when presented with it. The parameter can take any value between 0 and 1, where values closer to 0 indicate that the test-taker is less likely to solve items and values closer to 1 indicate that the test-taker is more likely to solve items.

We need to define a prior belief about θ over all possible values that θ can take. Depending on what we know, this could be anything from values being equally likely to beliefs that there are values of θ that are more likely. The prior belief is expressed as a probability distribution $P(\theta)$. We then have the likelihood $P(y|\theta)$, which expresses the probability of observing data given different values of the parameter. In our case, the data we observe can only take values between 0 and 1, which are incorrect and correct responses. Therefore, we can use the *Bernoulli distribution* as the likelihood function, which takes as input parameter p that can vary continuously between 0 and 1. If we simulate a response from a Bernoulli distribution, it will return a 0 or a 1. If the value of p is close to 1, it will almost always produce 1s, and if p is close to 0, the simulation will most often produce 0s.

We can use Bayesian inference to infer which values of our parameter θ are most credible given some sequence of data. Figure 3 shows the posterior inference of θ given two different priors (dashed and solid line) and 10 observations of data. You can imagine that the 1s are correct responses to items and the 0s are incorrect responses. The model's "belief" about the parameter values after having "seen" the data is referred to as the *posterior distribution*. The dashed prior starts with initial beliefs of lower values of θ while the solid prior starts with a uniform prior (all values equally likely). We can see in Figure 3 that the different priors lead to varying posterior conclusions. The dashed vertical line represents the true parameter value used to generate the data.

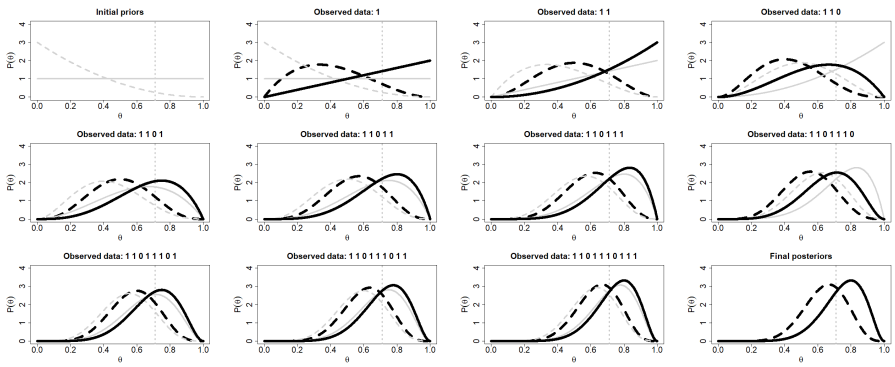

Figure 3: Iterative Bayesian updating. The gray lines show the priors in each iteration. The black line shows the posteriors at each iteration. The dotted vertical line shows the true underlying value ($\theta = 0.71$) used to simulate the data sequence which is indicated on top of each subplot.

3.2.1 Parameter estimation

Estimating the posterior can be done analytically for simple models, but for most complex models, numerical methods such as Markov Chain Monte Carlo (MCMC) algorithms are required (Brooks et al., 2011). Traceplots, showing the sampling sequence, can be used to assess convergence, good mixing, and stationarity of the Markov chains. For models with many parameters, it is more practical to use the convergence diagnostic statistic $\hat{R}$, which is recommended to be below 1.01 (Vehtari et al., 2021).

One way to assess model fit in the Bayesian paradigm is to plot posterior predictive checks by simulating outcomes from the model using posterior parameter values (Gabry et al., 2017). Posterior predictive checks compare model-generated data to observed data, and if they are similar, it suggests a good model fit, while a large discrepancy suggests a bad fit. Such plots were used in Studies III and IV.

3.2.2 Why Bayes?

Bayesian analysis, though not mainstream, offers, as I have experienced it, several benefits. The Bayesian paradigm provides modeling flexibility, enabling tailored model design without automatically forcing assumptions to be made; delivers intuitively interpretable results in terms of

probability statements, reflecting certainty/uncertainty about parameter values. Bayesian inference allows the use of prior information, which if contemplated could enhance understanding of models and parameters, and lead to better inferences. Furthermore, it emphasizes inference of parameters over hypothesis testing, while at the same time providing intuitive ways of testing hypothesis by comparing posterior parameter estimates. It encourages thinking about the data generating process — how we think the data were created and what model could simulate it. And finally, I think the Bayesian approach seem to be more in tune with how we think in science and everyday life. For example, consider that two researchers, A and B, faced with the same data, may have different prior understandings of the phenomena under study, and thus different opinions on appropriate prior parameter values. Consequently, they might reach slightly different conclusions, balancing their prior beliefs with information from empirical observations. However, as more data is gathered, and assuming they share the same model structure, their beliefs will eventually converge.

Also, in the case of PISA or other assessments using plausible values (draws from a posterior distribution), the Bayesian approach offers a straightforward method of incorporating information and uncertainty from all plausible values in analyses.

3.3 Code repositories

Repositories including data and code sufficient to reproduce the results are available at the following locations. Study I: https://osf.io/hfsjy/; Study II: https://doi.org/10.5281/zenodo.6670627; Study III: https://doi.org/10.5281/zenodo.7831743. As Study IV is still in manuscript stage no repository has been set up for now but will be available in the future.

3.4 Method and models

The methods are all generally related to finite mixture models, which are used to represent how sub-populations with different characteristics are responsible for creating observed data. Even though k-means (used in Study I) cannot really be considered a model, it is related to mixture modeling, as a kind of simplification of Gaussian mixtures. And multino-

mial Naïve Bayes used in the second study is a kind of multinomial finite mixture model. Study III and IV use Bayesian finite mixture modeling.

3.4.1 Study I

In Study I the approach to identify test-taking effort was to create various summary variables out of log-file data (e.g. time on task, number of actions, number of attempts, time to first actions) and then to cluster them with k-means clustering. K-means clustering segments the data set into K-components using K mean points in an iterative process: starting values are chosen for the component means μ_k. In the first step, each data point d_n is assigned to the nearest μ_k, using Euclidean distance. In the second step, each μ_k is updated by assigning it the average value of all data points that lie nearest to it. Step 1 and step 2 are repeated until no data points are chaining their cluster assignment, which is the convergence criterion. K-means can be sensitive to starting values, so usually, they are assigned randomly and many such starting values are tested. K-means is also sensitive to non-spherical data, and could be considered a somewhat unsophisticated method since it cannot take any uncertainty of classifications into consideration.

3.4.2 Study II

For Study II the approach was to model problem-solving behavior with computer programs that implemented different problem-solving strategies. The actions the computer programs took while problem-solving were coded into words (symbol sequences) which were then processed into *n-grams* and used as input to train a text classifier (Multinomial Naïve Bayes). To infer the strategies from real test-takers their action sequences were classified with the text classifier that was trained on the synthetic data generated by the computational models.

Roughly, Naïve Bayes works as follows: We have some data (sequences of actions) d which can be of k different classes $C \in \{c_1, c_2, \ldots, c_k\}$ (in our case problem-solving strategies). To retrieve a class given some data d Naïve Bayes applies Bayes' Theorem:

$$P(c|d) = \frac{P(d|c)P(c)}{p(d)}$$

Then classification is done by choosing the class with the highest posterior probability (Maximum a posteriori, MAP):

$$c_{MAP} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(c|d)$$

In the study, equal prior probability of all strategies was assumed, and the posterior probability of the class given the data $P(c|d)$ was calculated based on of how often n-gram action-sequences occur given data generated by the different strategies. For a specific strategy c, some action sequences are more commonly executed compared to other strategies, information provided by distribution $P(d|c)$ which is estimated by maximum likelihood: the number of times that n-gram action sequences occurred in data generated by strategy c divided by the total number of n-gram sequences. If a test taker's data include actions that have a high probability of having been generated by strategy c as compared to the other strategies, it will be classified as generated by this strategy. For more information on Naïve Bayes, see Schütze et al. (2008).

In contrast to solely relying on pattern recognition and data mining techniques (e.g., Study I), computational modeling provides precise definitions of strategies, reducing misconceptions that may arise from varying interpretations of verbally described models. Additionally, it reveals uncertainties from different strategies that yield identical behaviors, which might be overlooked by purely data-driven pattern recognition methods.

3.4.3 Studies III and IV

For Study III and Study IV, finite mixture models were used. A finite mixture model represents how a finite set of data-generating processes are mixed together to create observed outcomes. This makes them appropriate in our case when we assume that both optimizing and satisficing response processes create the data. Mathematically we can describe a finite mixture model as follows. We assume that the data is made up of K different component distributions f_k, that are mixed in proportion λ, where $\lambda_k \geq 0$ and $\sum_{k=1}^{K} \lambda_k = 1$. Each observed data y_n is associated with a latent variable $z_n \sim \text{Categorical}(\lambda)$, the component it pertains to. The observed data y_n is generated by the component distribution that the z_n indicates: $y_n \sim f_{k=z_n}$. The component distribution could be any kind

of distribution. For example, for response times, lognormal distributions are often used, and in this case, we have that, $f_k = Lognormal(\mu_k, \sigma_k)$ and $y_n \sim Lognormal(\mu_{k=z_n}, \sigma_{k=z_n})$. In our case, we are also interested in using these models for classifications, e.g. which component should be believed to have generated data y_n, and with what certainty? To answer this question Bayes theorem is used to recover the mixture component. If we take the example with the lognormal component distributions:

$$p(z_n = k|y_n, \mu, \sigma, \lambda) \; \frac{p(y_n|z_n = k, \mu_k, \sigma_k) \cdot p(z_n = k|\lambda_k)}{\sum_{k'}^{K} p(y_n|z_n = k', \mu_{k'}, \sigma_{k'}) \cdot p(z_n = k'|\lambda_{k'})}$$

This was the general approach, however, various extensions were made in both Study III and Study IV. In Study III, 61 finite mixture models were used, one for each item in the questionnaire, to estimate the questionnaire motivation parameter for each test-taker. And in Study IV the finite mixture model was multivariate, response times and subitem responses were modeled as generated jointly from the same underlying, satisficing, or optimizing, mixture component.

Additionally, various forms of regression modeling were applied. Study I uses clusters as predictors of test-taking performance and, ordinal regression modeling to estimate how clusters predicted self-reported effort. In Study II test-taker strategy classifications were used to predict test-performance. In Study III posterior estimates of questionnaire-taking motivation were used as predictors of test-taking performance, using both a linear regression model and a segmented regression model. Study IV used counterfactual simulations from the models to investigate what the effects would be on composite scores assuming that all test-takers would have used an optimizing response style. For detailed presentations and explanations of the exact models and their implementation see the methods sections in each of the papers and the code in the repositories. Models in Studies III and IV were coded in the probabilistic programming language Stan (Carpenter et al., 2017).

3.5 Ethical considerations

The data in this thesis relies solely on secondary analyses of previously collected data that has been made publicly available. As the data is anonymized at both individual and school levels and lacks sensitive personal information, and because PISA participation is voluntary and poses

no health or psychological risks to students, the studies within this thesis fall outside the formal requirements of an ethical review by the Swedish Ethical Review Authority. However, ethics in research involves more than determining whether data are sensitive and ensuring that the anonymity of participants is protected. For example, it also includes maintaining scientific transparency and honesty, for which I have aimed to provided repositories for each study containing code sufficient to reproduce the results. The code further makes explicit the exact decisions related to data cleaning, statistical analyses, models, figures, and results, which can sometimes be impractical to discuss in detail in research papers.

4 Summary of studies

4.1 Study I – Effort and motivation in a problem-solving task

Log-file data collected from a PISA 2012 problem-solving task, reflecting test-takers' actions when interacting with the task, was analyzed to develop a behaviorally informed classification of test-taking effort. The relationship between this behavioral classification and two other indicators, test-taking performance and self-reported test-taking effort, was examined.

Test-takers' log data were summarized into seven variables relevant to understanding their effort on the specific task. K-means clustering grouped test-takers with similar behavioral patterns into four clusters: low, medium, and high effort, and a "planner" cluster characterized by an initial long period of apparent inactivity or planning which made this cluster's level of effort difficult to interpret. These four clusters were split by item scores, as the effort-motivation relationship could vary depending on whether effort was exerted before giving up or before successfully completing the task*.

The results indicated that for test-takers who gave up (did not solve the task), higher effort levels were related to higher performance scores; however, this relationship was absent for those who solved the task. Regarding the relationship with self-reported effort, the only notable effect was from the low-effort cluster that did not solve the task, which had comparatively lower self-reported motivation.

Regarding the discrepancy in the relationship between effort and test performance, this depends on whether the effort signals test-taking motivation or not. When a test-taker finds a solution, we can only be certain that they had enough motivation to make the necessary efforts needed to solve the task, without knowing how much additional effort their motivation would have allowed them to use if they had needed to. Conversely, when test-takers give up on solving a task, the sum of their efforts must have reached a limit set by their motivation; in this case, test-takers' efforts are proportional to their subjective level of motivation. Reasons for the discrepancy between behavioral and self-reported measures are

*The task was designed in such a way that test-takers could ensure they had found the correct solution

also discussed.

4.2 Study II – Inferring problem-solving strategies

Results from Study I revealed variation in problem-solving behavior among test-takers, with some requiring substantial effort to solve the task while others requiring less. This indicates that test-takers rely on different, more or less efficient, strategies when solving items. Study II aimed to model test-takers' problem-solving strategies on the PISA 2012 problem-solving task (Traffic) using computational programs. Action sequences produced by simulated strategies were used to train a text-classifier used to inferred the most likely strategy applied by real test-takers. Variations of the task were generated to investigate the effects on strategy retrieval accuracy, and the performance of different problem-solving strategies. The relationship between test-takers' predicted strategies and their problem-solving performances was also explored.

Strategies were defined as the evaluation function (heuristic) used to prioritize action selection. Additionally, parameters related to lookahead, mistakes, forgetting, motivation, and backtracking method were drawn stochastically to create more realistic and noisy data.

The results showed that the overall classification accuracy of simulated strategies for the original PISA 2012 problem-solving task was 0.72, with some strategies being easier to classify than others. Varying the task design improved classification accuracy to 0.90. While various strategies had high chances of solving the original PISA version of the task, strategy choice becomes more important for success in larger-sized versions of the problem. The best-performing strategy in the original PISA task was the "travel time" strategy, while in the generalized Traffic task, the "straight line" strategy was slightly more efficient. Test-takers classified as using the "travel time" strategy were almost guaranteed to solve the task, whereas those using a "random actions" strategy had a low probability of success. The relationship between predicted strategies and PISA problem-solving performance indicated that "random actions" was associated with lower performance, while other strategies showed similar PISA performances.

The study demonstrates that combining computational modeling with text-classification can be used to infer strategies from response process

data, and make suggestions on task design that could increase the accuracy of diagnosing strategy use.

4.3 Study III – Questionnaire-taking motivation

Assessing motivation from response process data and response times in large-scale assessments can be challenging due to the varying item designs and the influence of test-takers' skills and knowledge affecting the effort required to solve items. However, the student questionnaire, administered closely to the test part of the assessment, is not to the same extent affected by these confounding factors. As all test-takers take the same, similarly designed, items, that should not be affected by skill, questionnaire-taking motivation could perhaps be more easily inferred, making the response processes on the questionnaire an interesting proxy for test-taking motivation. Study III aimed to explore the prevalence of questionnaire-taking motivation in PISA 2018 and its relationship with test performance.

The motivation to answer the questionnaire was conceptualized using a satisficing framework introduced by Krosnick (1991). Test-takers' motivation was operationalized as their probability of using an optimizing response style on questionnaire items. Optimizing was assumed to be probabilistically observable through item response times. A hierarchical Bayesian model accumulating information from finite mixture models was developed to infer questionnaire-taking motivation, and the relationship with test-taking performance.

Using a sample of Swedish test-takers, the model estimated that the overall questionnaire-taking motivation was very high, with a mean of 0.97 and a 95% credible interval of [0.86, 1.00]. The results showed increased satisficing for items at the end of the questionnaire. The correlation between questionnaire-taking motivation and test performance was 0.75 (Spearman), and 0.31 (Pearson), suggesting a stronger monotonic than linear relationship, possibly indicating a nonlinear positive relationship between motivation and performance. This led to the application of a piecewise regression model, which revealed a breakpoint for very high levels of motivation where the relationship between questionnaire-taking motivation and performance became stronger.

The discussion considered the possibility that the strong positive relationship after the breakpoint in the piecewise regression might be due

to measuring reading ability rather than motivation. The model demonstrated a good fit to most items' response time distributions. However, a few of the items' response time distributions did not fit well as their distributions exhibited an additional mode that could not be accounted for by the model. These findings were explored further in Study IV.

4.4 Study IV – Satisficing in questionnaire items

The final study aimed to develop a model that could capture different response styles of omitting, satisficing, and optimizing on questionnaire items, by considered both response times and patterns within sub-item responses. The model was used to investigate if the composite scores calculated from the items were influenced by the satisficing responses.

The investigation focused on three items (`ST164`, `ST165`, and `ST166`) from the PISA 2018 questionnaire. The items were assumed to measure awareness of metacognition strategies in reading, three composite scores were calculated by PISA, one for each item. Using data from Swedish test-takers, the results showed that the mean point estimate of satisficing was 15%, 29%, and 28%, respectively for each item. The subitem patterns in the responses indicated that straight-lining was a common response pattern associated with faster response times. The individual item responses had different distributions depending on the response style. Additionally, counterfactual simulations revealed that the amount of satisficing was biasing the composite scores, especially so for two of the items.

The study concludes by acknowledging that the results indicate that these items receive, compared to results in other studies, a large proportion of satisficing responses. Findings further suggest that the proportion of satisficing impacted the composite score results, thus seemingly valid responses could cause validity issues that may go unnoticed. However, by considering information from response times and response patterns, this threat can be mitigated.

5 Discussion

This thesis aimed to explore and model response process data from an international large-scale assessments to increase the understanding of test-taking motivation, problem-solving strategies, and integrity of questionnaire responses. The thesis is composed of four studies and a summary. Study I used log-file data from a PISA 2012 problem-solving task to explore task-related effort and motivation. Study II aimed to provide a method of inferring problem-solving strategies from a PISA 2012 problem-solving task. Study III focused on modeling response times to make inferences regarding students' motivation to answer the PISA 2018 student questionnaire, and Study IV examined the influence of low-motivation on composite scores from a set of PISA 2018 questionnaire items. The studies indicate that response process data can be a promising way forward to obtain better estimates of test-taking motivation and to improve the validity of score interpretations from items in both the cognitive and questionnaire parts of the assessment. In this concluding chapter, main findings are presented and discussed from a validity perspective.

5.1 Main findings related to test-taking motivation

5.1.1 Relationship to test performance

The results from Studies I and III show a positive relationship between test-taking motivation as inferred from response process data and test performance. These results are in line with previous research and imply that some variation in the score is due to varying levels of motivation, which, as long as test-takers' level of motivation is not explicitly considered by the construct, present some amount of construct-irrelevant variance.

However, it is important to note that the relationship between behavioral observations of test-takers effort and their latent test-taking motivation is not always straightforward. This point is further discussed whiting the next section.

5.1.2 Relationship to self-reported effort

Regarding the relationship between response process-inferred motivation and self-reported test-taking effort, the studies in this thesis do not fully investigate this. However, as it is an interesting and relevant question, additional analyses (new to the kappa) have been conducted to explore the relationship between the response process and self-report test motivation indicators, and how they contribute to the prediction of test-taking performance (see Appendices 1 and 2). The results showed a very low positive correlation between self-reported effort and response process indicated motivation, even lower than what previous research comparing RTE and self-reported effort has shown, a finding that suggests that response time and self-report measurements of effort and motivation do not measure the same construct (see, Silm et al., 2019; Wise & Kong, 2005).

Interestingly, both response process-indicated motivation and self-reported effort contribute to test performance predictions, aligning with (Silm et al., 2019). However, the strength and relative importance vary based on whether motivation was assessed from the Traffic task or the questionnaire. Results in Appendix 2 show that the questionnaire-derived motivation from Study III was a close to four times stronger predictor of performance scores compared to self-reported effort. In Study I, the level of effort signaling motivation (effort exerted before giving up), the predictive value was similar to self-reported effort. However, level of effort before *solving* the task, became a weak negative predictor of test-performance. An explanation for this result could be that effort required to solve cognitive tasks interacts with skill, such that more skilled test-takers need to exert less effort regardless of their motivation levels. These results provide further insight into Study I and convey the key message that it was level of effort before giving up that signaled motivation, while effort before solving the task did not. This finding thus complicates the use of response process data from test items as indicators of test-taking motivation. That the relationship between self-reported effort and test performance was unaffected by whether or not test-takers solved the Traffic task, suggests that self-reported motivation is also applicable to motivated and efficient test-takers, for whom the effort-motivation relationship might not be clearly inferred from the kind of response process data that was used in this thesis.

5.1.3 Impact of unmotivated responses (satisficing) on questionnaire composite scores

Results from Study III suggest that overall, test-takers in the Swedish sample were motivated to respond to the PISA student questionnaire. The small proportions of satisficing should not have a dramatic effect on either the performance score results or impact score distributions *for most* individual items. However, the approach developed in Study III could not model all items, a limitation addressed by Study IV.

Study IV focused on more detailed modeling of response processes for a set of questionnaire items appearing to receive higher rates of satisficing responses and investigated if satisficing had an impact on composite scores. Results indicated a relatively large proportion of satisficing responses and counterfactual analyses assuming only optimizing responses suggested that composite scores were biased. This indicates that results from these items could lead to unjustified interpretations and might affect secondary analyses. Therefore, a recommendation is to conduct detailed analyses of satisficing on items in the background questionnaire, if possible. This finding shows that secondary, post-hoc analyses of response processes can affect validity, irrespective of whether explicit claims related to these response process are made in relation to the construct by the assessment framework.

5.1.4 Conclusion and implications of results

In conclusion, these findings have significant implications for measuring test-taking motivation. Firstly, using response process-derived effort assessments from cognitive tasks may be misleading when interpreting test-taking motivation, as efficient test-takers' efforts may not reveal their level of motivation. Secondly, assessing effort via questionnaires is a promising complement to self-reports and appears more practicable compared to using items from the cognitive part of the test. Finally, modeling response processes on questionnaires can help detect potential validity issues related to the composite scores derived from them, which is important finding for large-scale assessments and survey research in general

5.2 Main findings related to inference of problem-solving strategies

5.2.1 Accuracy of inference

Results from Study II indicate that inferring problem-solving strategies from response process (log-file data) achieved 72% accuracy on the original PISA item design, improving to 90% with design changes. This classification performance aligns with related studies by Holden et al. (2021) and Tang, Zhang, et al. (2021) showing 83% classification accuracy. Note however, that these studies did binary classification, while the Study II had 6 different categories of problem-solving strategies to classify, some of which could be classified with accuracy of 87%, 97%, and 82%. More advanced classifiers, like various deep learning methods, might boost accuracy. However nonidentifiability due to different strategies leading to the same behavior, which can be quite common in small-sized versions of the Traffic task, could limit improvements. Although this level of accuracy is not ideal for assessment purposes, it could perhaps be useful in learning situations to hypothesize which strategies are implemented by test-takers offer suggestions to try different ones.

5.2.2 Efficiency of strategies

Strategies were of different efficiency, requiring different amounts of effort and motivation before solving the problem. Strategies using the travel time or straight line heuristics were most efficient, while the strategy of trying paths randomly without the guidance of heuristics required increased effort. Random action selection proved ineffective. In light of these findings, that various problem-solving strategies can be applied successfully, results suggest that the item mainly requires persistence and motivation, which, although crucial for success in problem-solving, might not have been what was intended to capture by the construct.

The result that the majority of test-takers searched randomly among paths suggests unfamiliarity with the strategies that were most efficient on the Traffic task, indicating potential benefits from teaching them. However, any such benefit is conditional on strategies effective in the Traffic task also being useful for finding the shortest paths in maps encountered in everyday life. Previous research on PISA problem-solving has emphasized the vary-one-thing-at-a-time (VOTAT, Tschirgi,

1980) strategy that has showed to be positively linked to performance in problem-solving tasks (Greiff et al., 2015; Wu & Molnár, 2021). However, success depends on tasks having direct effects and no changes happening to the task environment without interaction. In the latter case, a "NOTAT"-strategy (observing system behavior without interaction) is needed to understand the system (Lotz et al., 2022). Understanding multivariable systems with interaction effects demands other strategies (Teig et al., 2020). This highlights the importance of metastrategic knowledge: knowing which strategy to apply based on the task's structure. As neither strategy applies directly to the Traffic task, more relevant strategies related to shortest-path problems were explored. But it is possible to think of all strategies except the random actions as a kind of VOTAT strategy, if understanding the path trials as the "thing" that is varied. However, perhaps the random paths strategy would be the most similar since it is not guided by any heuristic information (something that could be regarded as metastrategic knowledge). A Critique of PISA problem-solving tasks is that more complex tasks need to be attempted during longer time periods in order to capture problem-solving relevant to real-life situations (Dörner & Funke, 2017; Schoppek & Fischer, 2015). While I think this critique is valid, it could probably be difficult to fit many longer, more complex, problem-solving tasks within one assessment.

5.2.3 Relationship to performance

Test-takers classified as using a random-actions strategy exhibited lower performance. The lack of differences between other strategies' relationship with problem-solving competency may stem from their specificity to the Traffic task. Though the strategies differ in efficiency on the Traffic task, knowledge of these strategies does not guarantee success in other problem-solving items. The positive relationship between VOTAT (Greiff et al., 2015) and test-performance is perhaps due to it being a more general strategy, or since other items in the PISA 2012 problem-solving assessment shared underlying structural features for which the application of VOTAT is successful. The performance difference between the random actions strategy and other strategies in Study II, is however on most occasions, likely to be attributed to differences in motivation to make a serious attempt or not, rather than due to the strategy used.

5.2.4 Conclusion and implications of results

In summary, Study II demonstrates that computational modeling of response process data helps explain how different strategies leads to behaviors and can be used to infer problem-solving strategies, inform item design, evaluate item difficulty, and efficiency of problem-solving strategies, thus adding interpretative value to scores. This study revealed that out of nearly 25,000 test-takers, 70% had unique action sequences, highlighting the vast behavioral variation humans can exhibit even in a simple well-defined environment, far less complex than most everyday situations. As noted by Leighton and Gierl (2007) cognitive models of task performances are the only ones that can support claims about thinking processes. While currently impractical for implementation across all different items in a large test, in an ideal world they could be used for all items.

5.3 Validity implications

Overall, the studies that are part of the present thesis presents two general validity implications. First, test-taking motivation inferred from response process data can complement self-reported estimations, and improve the evaluation of the prevalence of test-taking motivation and how it influences scores. Second, response processes help identify alternative explanations which could influence the interpretation of test scores. The following two examples illustrate how findings from response process data could be incorporated into assessment arguments.

5.3.1 Attempts to incorporate evidence from response processes within an assessment validity argument

To offer concrete examples of using evidence from response processes in assessment arguments, Toulmin-style arguments have been constructed, incorporating information from Study I and Study IV.

Beginning with Study I, which used response processes to develop a behaviorally informed indicator of test-takers' motivation and effort on the Traffic item, this can be formalized as shown in Figure 4.

Figure 4 presents an assessment argument structured as a Toulmin diagram. The observed data (D), an incorrect score, leads to the claim (C) that the student lacked the required skills, supported by warrant

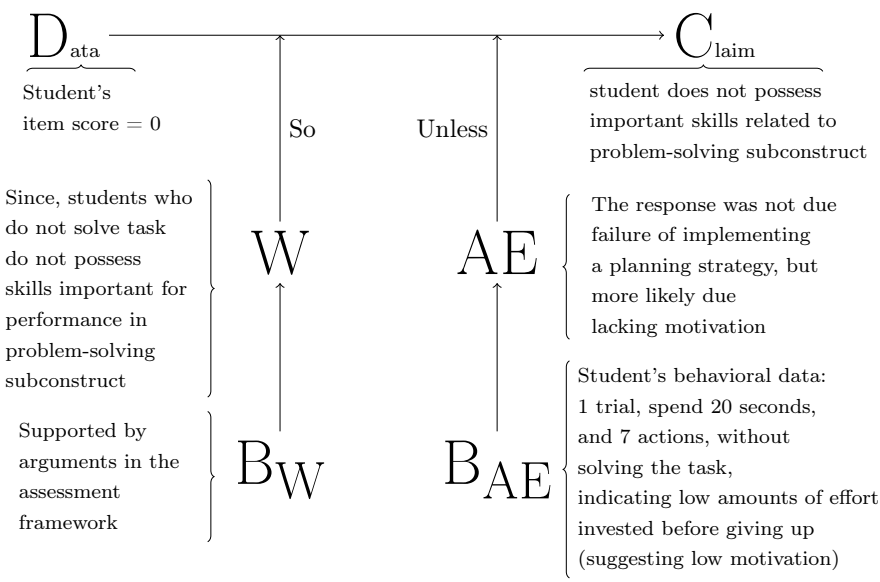

Figure 4: Example of a validity argument regarding a response to the Traffic item.

(W) and backing (B_W). This claim holds unless an alternative explanation (AE) is backed by evidence (B_{AE}). In this hypothetical example, backing evidence for an alternative explanation suggests the test-taker was not motivated. There is no qualification made, that limits the scope of precision, to this claim. It is stated without a qualifier, as an either-or argument, as the method of classifying test-takers according k-means do not permit probabilistic assignment of alternative explanations.

Figure 5 shows the test-taker's questionnaire responses data (D) leading to the claim (C) of the level of awareness in reading metacognition strategies, unless an alternative explanation (AE) attributes it to satisficing. Backing evidence (B_{AE}) suggests a 0.99 probability of satisficing, which qualifies the claim, revealing a 0.01 certainty that the score reflects the test-taker's knowledge of reading strategies. Here the method of inference allows us to make a probabilistic qualification of the claim. Although not impossible, an optimizing response yielding the same score is highly unlikely, and the score should be considered missing data or attributed to satisficing.

While these arguments may seem trivial, considering alternative ex-

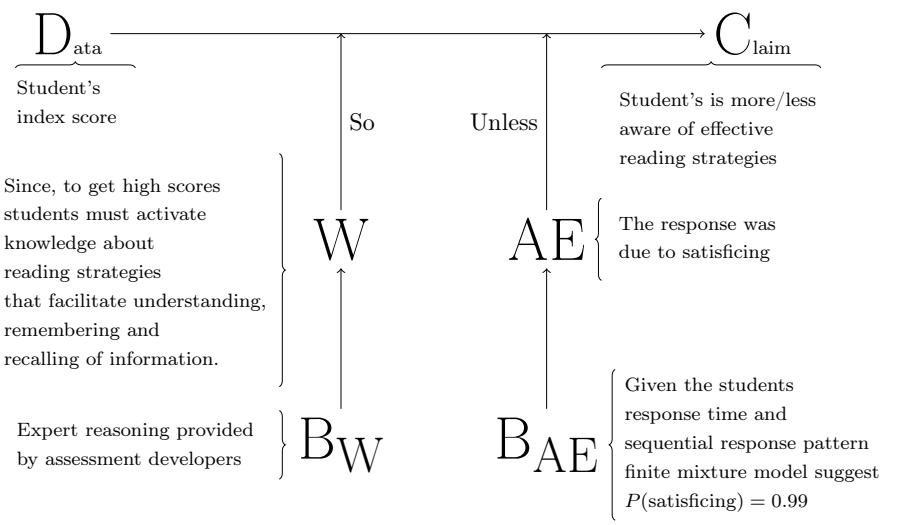

Figure 5: Validity argument regarding a questionnaire metacognition composite score.

planations for observed data generation can strengthen the validity of claims made by assessments. Absence of strong evidence for alternative explanations would increase the credibility of the claim about the construct, whereas support for an alternative explanation would decrease the credibility.

5.3.2 On the importance of item design

To understand behavior is to understand the environment: "the apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves" (Simon, 1996, p. 53). To gather information on how test-takers think, tasks need to be designed such that test-takers' observable actions closely map to their cognitive processes. Items requiring minimal observable response processes data (e.g., click-data from multiple choice reading tasks), naturally offer limited insights into cognitive processes. However, tasks such as the Traffic task, analyzed in Study I and II, yields much process data due to the numerous clicks that most test-takers produce while searching, allowing response process data from this item to be informative about cognitive response processes. However, optimizing tasks for response

process data may compromise validity, resulting in cumbersome, unrealistic tasks that poorly reflect real-world skills. Thus, a balance must be struck between gathering response process data and maintaining task validity.

While utilizing process data "by design" (Maddox, 2023) is preferable. It will however, probably, be as challenging as post-hoc, retrospective analyses. As Mislevy writes: "The hard way is to ask “How do you score it?” after you have built the assessment and scripted the tasks or scenarios. Unfortunately, the contrasting approach is not “the easy way” but a different hard way" (Mislevy et al., 2002, p. 385). While only the second hard way can be justified when creating assessments, the first hard way can nonetheless provide interesting ideas and hypotheses of test-takers' response processes relevant to the validity of their responses.

5.4 Limitations and future research

A key limitation of the present research is the lack of ground truth criteria to validate inferences about test-takers' motivation, questionnaire-taking motivation, and problem-solving strategies. Future research could address this through experiments providing conditions for different levels of motivation, satisficing response, and application of specific strategies, then compare these known conditions, to inferences from models. Another way to address the problem is to triangulate inferences from the data and models applied in this thesis, with inferences from based on different methods such as eye tracking, mouse tracking, video observations, and/or think-aloud protocols. This could and reveal and clarify blind spots of methods applied in this thesis and provide concurrent validity evidence (see, Zoanetti & Griffin, 2017).

Concerning data analysis approaches, Bayesian finite mixture models that combine response times with other types of data are promising means of investigating test-taking motivation and should be developed further. The computational modeling approach used in II is impractical to apply to all items, and the exploratory clustering approaches used in Study I offer only coarse, and atheoretical, descriptions of test-takers' behaviors, hiding much individual variation. Improving Bayesian modeling could involve incorporating better information into priors. This would be especially important for small datasets or for identifying infrequent response processes.

Each individual study has its own limitations. Study I relies on a single task, and the assessment of motivation is limited to test-takers who did not solve the task, restricting generalizations. Further research should consider behavioral evidence from multiple test items; however, this would require detailed knowledge of each item's design, and as Study I results indicate, inferences may not be reliable for test-takers who solve the task, making it impractical.

The main limitation in Study III is using questionnaire-taking motivation as a proxy for test-taking motivation, assuming that motivation is consistent across both contexts. This assumption can be questioned, and future research could challenge it. Another limitation is that questionnaire-taking motivation might be affected by reading fluency, this could also be investigated in future studies.

Study IV offers the strongest, most direct assessment of test-takers' motivation and its effects. However, it has limitations, as it relies on subitem responses on the same scale, which is not applicable to all items. Although the finite mixture model approach could be applied solely to response times, it would benefit from highly accurate prior information gathered from component actions, like prior distributions on response times of mouse-clicks, mouse-movements, reading-speed, and writing speed, along with detailed task environment descriptions. This information could then be used in more comprehensive models, building upon the models presented in this thesis.

In the case of Study II, there might be unconsidered problem-solving strategies besides the modeled strategies. Before any such methods are used in practice, research should identify strategies applied in the field by real test-takers, as well as the extent of shifts in problem-solving strategies during problem-solving. Future studies should also model response times between actions which could provide valuable information for resolving uncertainties in problem-solving strategy inference, potentially using cognitive architectures (Kotseruba & Tsotsos, 2020). Future studies could also investigate improvements by applying more powerful approaches to sequence pattern recognition such as deep learning language processing models (Minaee et al., 2021).

This study addresses validity in assessment arguments from data in test situations to claims about test-taker constructs. However, a full validity argument includes an assessment-uses argument (Mislevy, 2018),

which concerns claims about test-takers in future situations. This aspect is crucial, and future research could involve joint efforts of educational measurement and policy researchers to evaluate and understand the consequences of using large-scale assessment results.

The present research could be critiqued for not being holistic (see, Zumbo et al., 2023), i.e., not taking into consideration social, cultural, and other contextual factors that have shaped the environment and influenced test-takers' motivation or problem-solving strategies. This critique is relevant, however, the focus of the present thesis was mainly to provide information on test-takers from the immediate situation. Future research could extend current models and take such factors into consideration.

While process data have up to this point mostly been seen as a by-product used for occasional secondary analysis, efforts are made to make more systematic use of response processes (Maddox, 2023). The present study is no exception, planning the analyses and designing tasks with response process in mind before delivering the assessment is preferable. However, there is still value in post-hoc secondary analyses of response process data, as they can provide valuable information relevant to score validity, even though the test developers may never have thought about using response process data.

Regarding the potential for unintended and negative consequences of using response process data to understand test-takers, I agree with Maddox (2023) that this area need more research. Concerning the present research, we should consider whether native or second language influences the inference of satisficing, misclassification of highly efficient test-takers as satisficing, and how any disability could influence response times and any inferences made from them (see, Zumbo et al., 2023). Furthermore, to ensure trust, transparency, and informed consent, test-takers, and the public should, to the greatest extent possible, be informed about the response process data collected, along with its purpose and usage within the assessment.

5.5 Conclusions

Considering response process data puts us in a better position to understand test-takers' motivation, as evidenced by positive relationships to test performance, effects on composite scores, and qualitative interpretations of clustering and models. Similarly, regarding problem-solving

strategies, much can be learned by considering the sequential actions that test-takers do while working with test items. Analyzing response process data from PISA is not like looking into a crystal ball, there are limitations, and response process data will surprise any analyst with the rich and chaotic nature of human behavior. Nevertheless, well-reasoned models of response processes can enhance our understanding of test-takers performances, providing crucial information to advance our knowledge of both test-takers and assessments. Used the right way, it can be used to improve assessments, making them more valid and fair.

6 References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Apascaritei, P., Demel, S., & Radl, J. (2021). The difference between saying and doing: Comparing subjective and objective measures of effort among fifth graders. *American Behavioral Scientist*, *65*(11), 1457–1479.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1–34.

Behrens, J. T., & DiCerbo, K. E. (2014). Harnessing the currents of the digital ocean. In *Learning analytics* (pp. 39–60). Springer.

Bergner, Y., & von Davier, A. A. (2019). Process data in naep: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. https://doi.org/https://doi.org/10.1787/5k9fdfqffr28-en

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Brysbaert, M. (2019). How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, *109*, 104047.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & B. H. I. (Eds.), *Test validity* (pp. 3–17). Routledge.

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01153

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European journal of psychology of education*, *28*, 497–510.

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in timss advanced. *Applied Measurement in Education*, *27*(1), 31–45.

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment*, *25*(3), 179–197.

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Taylor & Francis.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. the MIT Press.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2017). Visualization in bayesian workflow. *arXiv preprint arXiv:1709.01449*.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, *1*(3), 291–308.

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

Hahnel, C., Ramalingam, D., Kroehne, U., & Goldhammer, F. (2022). Patterns of reading behaviour in digital hypertext environments. *Journal of Computer Assisted Learning*.

Han, Y., Liu, H., & Ji, F. (2022). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research*, *57*(6), 960–977.

He, Q., Borgonovi, F., & Paccagnella, M. (2019). Using process data to understand adults' problem-solving behaviour in the programme for the international assessment of adult competencies (piaac): Identifying generalised patterns across multiple tasks with sequence mining.

He, Q., Borgonovi, F., & Suárez-Álvarez, J. (2022). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning*.

He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using process data in piaac problem-solving items. *Theoretical and practical advances in computer-based educational measurement*, 189–212.

Holden, L. R., LaMar, M., & Bauer, M. (2021). Evidence for a cultural mindset: Combining process data, theory, and simulation. *Frontiers in Psychology*, 3998.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353.

Hu, Y., Wu, B., & Gu, X. (2017). An eye tracking study of high- and low-performing students in solving interactive and analytical problems. *Journal of Educational Technology Society*, *20*(4), 300–311. Retrieved April 16, 2023, from http://www.jstor.org/stable/26229225

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In *Understanding and investigating response processes in validation research* (pp. 1–12). Springer.

Jiang, Y., Cayton-Hodges, G. A., Oláh, L. N., & Minchuk, I. (2023). Using sequence mining to study students' calculator use, problem solving, and mathematics achievement in the national assessment of educational progress (NAEP). *Computers & Education*, *193*, 104680.

Kane, M. (2012). Validating score interpretations and uses. *Language testing*, *29*(1), 3–17.

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 11–24). Routledge.

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.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/https://doi.org/10.1002/acp.2350050305

Krstić, K., Šoškić, A., Ković, V., & Holmqvist, K. (2018). All good readers are the same, but every low-skilled reader is different: An eye-tracking study using PISA data. *European Journal of Psychology of Education*, *33*(3), 521–541.

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika*, *83*(1), 67–88.

Launeanu, M., & Hubley, A. M. (2017). A model building approach to examining response processes as a source of validity evidence for self-report items and measures. In *Understanding and investigating response processes in validation research* (pp. 115–136). Springer.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3–16.

Lewin, K. (1938). *The conceptual representation and the measurement of psychological forces*. Duke University Press.

Levy, R. (2020). Implications of considering response process data for greater and lesser psychometrics. *Educational Assessment*, *25*(3), 218–235.

Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time data as validity evidence: Has it lived up to its promise and, if not, what would it take to do so. In *Understanding and investigating response processes in validation research* (pp. 159–177). Springer.

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2022). G's little helpers–votat and notat mediate the relation between intelligence and complex problem solving. *Intelligence*, *95*, 101685.

Lu, J., Wang, C., & Shi, N. (2021). A mixture response time process model for aberrant behaviors and item nonresponses. *Multivariate Behavioral Research*, 1–19.

Maddox, B. (2017). Talk and gesture as process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3-4), 113–127.

Maddox, B. (2023). The uses of process data in large-scale educational assessments. (286). https://doi.org/https://doi.org/10.1787/5d9009ff-en

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning.* Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing*, *20*(3), 187–205. https://doi.org/10.1080/15305058.2019.1706529

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, *54*(3). https://doi.org/10.1145/3439726

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

Mislevy, R. J. (2009). *Validity from the perspective of model-based reasoning* (tech. rep. CRESST REPORT 752). National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* Routledge.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, *15*(4), 363–389.

Moon, J. A., Finn, B., LaMar, M., & Irvin R., K. (2018). *Simulations of thought: The role of computational cognitive models in assessment* (Periodical RDC-26). Educational Testing Service.

Moore, T. M., & Picou, E. M. (2018). A potential bias in subjective ratings of mental effort. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2405–2421.

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.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104). Prentice-hall Englewood Cliffs, NJ.

Newton, P. E. (2019). What is response process validation evidence and how important is it? an essay reviewing ercikan and pellegrino (2017) and zumbo and hubley (2017). *Assessment in Education: Principles, Policy & Practice*, *26*(2), 245–253. https://doi.org/10.1080/0969594X.2018.1447909

Nichols, P., & Huff, K. (2017). Assessments of complex thinking. *Validation of score meaning for the next generation of assessments: The use of response processes. New York, NY: Routledge*, 63–74.

OECD. (2014). PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (volume v).

OECD. (2017). *PISA 2015 technical report*. https://www.oecd.org/pisa/data/2015-technical-report/

OECD. (2019). *Pisa 2018 assessment and analytical framework*. https://doi.org/https://doi.org/10.1787/b25efab8-en

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education*, *2*, 1–17.

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.

Pettersson, D. (2008). *Internationell kunskapsbedömning som inslag i nationell styrning av skolan* (Doctoral dissertation). Universitetsbiblioteket.

Pons, X. (2017). Fifteen years of research on PISA effects on education governance: A critical review. *European Journal of Education*, *52*(2), 131–144.

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), 1–31.

Primativo, S., Spinelli, D., Zoccolotti, P., De Luca, M., & Martelli, M. (2016). Perceptual and cognitive factors imposing "speed limits" on reading rate: A study with the rapid serial visual presentation. *PLOS ONE*, *11*(4), 1–25. https://doi.org/10.1371/journal.pone.0153786

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in psychology*, 2231.

Rafferty, A. N., Jansen, R. A., & Griffiths, T. L. (2020). Assessing mathematics misunderstandings via bayesian inverse planning. *Cognitive science*, *44*(10), e12900.

Rafferty, A. N., LaMar, M. M., & Griffiths, T. L. (2015). Inferring learners' knowledge from their actions. *Cognitive Science*, *39*(3), 584–618. https://doi.org/https://doi.org/10.1111/cogs.12157

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85–106.

Russell, L. B., & Hubley, A. M. (2017). Some thoughts on gathering response processes validity evidence in the context of online measurement and the digital revolution. In *Understanding and investigating response processes in validation research* (pp. 229–249). Springer.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach (4th edition)*. Pearson. http://aima.cs.berkeley.edu/

Rutkowski, D., Thompson, G., & Rutkowski, L. (2020). Understanding the policy influence of international large-scale assessments in education. *Reliability and validity of international large-scale assessment*, *261*.

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*, 93–114.

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.

Schoppek, W., & Fischer, A. (2015). Complex problem solving—single ability or complex phenomenon? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01669

Schwarz, N. (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture. *The science of real-time data capture: Self-reports in health research*, *11*, 26.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press Cambridge.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, *55*(2), 503–524.

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

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.

Simon, H. A. (1996). *The sciences of the artificial*.
Skolverket. (2023). *Redovisning av uppdrag att digitalisera de nationella proven*. https://www.skolverket.se/getFile?file=11127
Solano-Flores, G., & Chıa, M. (2017). Validation of score meaning in multiple language versions of tests. In *Validation of score meaning for the next generation of assessments* (pp. 127–137). Routledge.
Stewart, T. C. (2007). *A methodology for computational cognitive modelling* (Doctoral dissertation). Carleton University.
Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, *74*(1), 1–33.
Tang, X., Zhang, S., Wang, Z., Liu, J., & Ying, Z. (2021). Procdata: An r package for process data analysis. *Psychometrika*, *86*, 1058–1083.
Teig, N., Scherer, R., & Kjærnsli, M. (2020). Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data. *Journal of Research in Science Teaching*, *57*(9), 1400–1429.
Tindal, G., Alonzo, J., Sáez, L., & Nese, J. F. (2017). Assessment of students with learning disabilities: Using students' performance and progress to inform instruction. In *Validation of score meaning for the next generation of assessments* (pp. 113–126). Routledge.
Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.
Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child development*, 1–10.
Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *psychometrika*, *86*, 190–214.
Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619.
Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2022). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, 1–21.
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*, 83–112.
Waldow, F. (2009). What PISA did and did not do: Germany after the 'PISA-shock'. *European Educational Research Journal*, *8*(3), 476–483.
Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for As-

sessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Weiner, B. (2013). Little-known truths, quirky anecdotes, seething scandals, and even some science in the history of (primarily achievement) motivation. *Personality and Social Psychology Review*, *17*(3), 293–304.

Wigfield, A., Tonks, S., & Klauda, S. L. (2009). Expectancy-value theory. In *Handbook of motivation at school* (pp. 69–90). Routledge.

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

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

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

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.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, *8*, 175–201.

Wu, H., & Molnár, G. (2021). Logfile analyses of successful and unsuccessful strategy use in complex problem-solving: A cross-national comparison study. *European Journal of Psychology of Education*, *36*, 1009–1032.

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

Zoanetti, N., & Griffin, P. (2017). Log-file data as indicators for problem-solving processes.

Zumbo, B. D., & Hubley, A. M. (2017). *Understanding and investigating response processes in validation research* (Vol. 26). Springer.

Zumbo, B. D., Maddox, B., & Care, N. M. (2023). Process and product in computer-based assessments: Clearing the ground for a holistic validity framework. *European Journal of Psychological Assessment*.

7 Appendix 1

Intro

This report provides additional analyses extending those in Study I. The report examines the correlation between response process indicated motivation and self-reported test-taking effort, and investigates how the response process indicated motivation and self-reported effort contribute to PISA performance outcomes.

Method

The analysis utilizes the same sample of test-takers who completed the Traffic 2012 problem-solving task, as was used in Study I.

Correlations were used to assess the relationship between response process effort and self-reported effort, using Spearman (ρ_s) and Pearson (ρ_p) methods.

Linear regression modeling was employed to evaluate the contributions of response process effort and self-reported effort in predicting PISA performance. To facilitate comparison and interpretation, each n test-taker's self-reported effort from the effort thermometer y_n was coded into low-, medium-, and high self-report effort scores (SE) using the following procedure:

$$\text{SE}_n = \begin{cases} 1, & \text{if } y_n \in \{1, 2, 3\}, \\ 2, & \text{if } y_n \in \{4, 5, 6, 7\}, \\ 3, & \text{if } y_n \in \{8, 9, 10\}, \end{cases}$$

Thus, 1 represents low-, 2 medium-, and 3 high self-reported effort, respectively. This makes them more comparable to the behavioral effort clustering results (BE) from Study I, which were coded similarly: 1 for low-, 2 for medium-, and 3 for high behavioral effort. The Planner cluster was excluded due to uncertainty about its effort level.

To determine the extent to which behavioral and self-report effort predicted test performance, two linear regression models were applied:

$$\text{Model 1: } \mu = a + \beta_{BE} BE_n + \beta_{SE} SE_n$$
$$\text{Model 2: } \mu = a[s] + \beta_{BE}[s] BE_n + \beta_{SE}[s] SE_n$$

Where $[s]$ indexes item scores and thus imply a model with interaction effects of item score.

The full model with priors is presented below. Simply swap the relationship to the mean as stated above to obtain Model 1 and Model 2.

$$
\begin{aligned}
pv_n &\sim Normal(\mu, \sigma) \\
\mu &= \text{Model 1 or Model 2} \\
a &\sim Normal(0, 1000) \\
BE &\sim Normal(0, 1000) \\
SE &\sim Normal(0, 1000) \\
\sigma &\sim Normal(100, 1000), \text{ s.t. } \sigma > 0
\end{aligned}
$$

The model parameters were estimated using data from all five plausible values. See repository at https://doi.org/10.5281/zenodo.7838627 for data and code sufficient to reproduce the analyses.

Results

The correlation between behavioral and self-reported effort was $\rho_p = 0.06$, and $\rho_s = 0.04$

Table 1: Parameter estimates

	Model 1	Model 2	
		score = 0	score = 1
a	367.95 [359.60, 375.80]	345.50 [332.85, 358.60]	438.21 [428.86, 447.96]
β_{BE}	16.24 [14.15, 18.43]	23.22 [20.07, 26.34]	−5.88 [−8.60, −3.30]
β_{SE}	39.17 [36.55, 41.96]	22.91 [18.07, 27.53]	37.22 [34.49, 40.04]
σ	84.79 [83.70, 85.85]	78.94 [78.00, 79.91]	

Note: Posterior mean estimates, 95% Credible Intervals in brackets.

Discussion

This report compares response process indicated motivation and effort with self-reported test-taking effort, building on data and results from

Study I. There are weakly positive relationships between behavioral effort in the Traffic task and self-reported effort, for both types of correlation. This does not suggest a strong relationship between the variables, meaning test-takers' behavioral effort ratings may not align with their self-reported effort ratings.

Results from Regression Model 1 (see Table 1) show a small positive relationship between both behavioral effort and self-reported effort, indicating they both contribute to predicting performance, with self-reported effort having roughly twice the effect of effort measured from the Traffic task. However, further nuances in these relationships emerge from Model 2. For test-takers with incorrect scores, self-reported test-taking effort and behavioral effort show a very similar relationship to test performance. Conversely, for test-takers with correct scores, increased effort in the Traffic task is slightly negatively related to overall test performance. These findings suggest, as in Study I, that effort before giving up on solving a task is positively related to test-taking performance, while this claim cannot be made for test-takers who solve the task. In the latter case, effort may signify efficiency, which is weakly negatively related to test performance.

The discrepancy between effort on the Traffic task and self-reported effort raises questions about how the response process indicated effort is actually related to the behavioral and mental effort invested. The discrepancy implies that solving tasks "efficiently" might require greater cognitive effort into the few actions they made, while those who used many actions and took a long time may have been less attentive.

That self-report scores did not change the direction of the relationship, suggests that self-report information is a more straightforward and potentially more robust indicator of test-taking effort for test-takers that can solve tasks efficiently. In conclusion, these results indicate that one cannot assume a positive relationship between effort exerted on an item and test-taking motivation for all test-takers, as the relationship could be the opposite for many of them, while self-reported test-taking effort stays roughly similar independent of item score.

8 Appendix 2

Intro

This report provides additional analyses extending those in Study III. The report aims to investigate the relationship between response process indicated motivation and self-reported effort, and how both types of effort measurements contribute to PISA performance scores.

Method

The same sample as in Study III was used, containing test-takers that completed the PISA 2018 questionnaire. After combining with the self-reported effort data, the number of complete cases dropped to 4578 due to some missing self-reported data.

Correlations were used to assess the relationship between response process-derived effort and self-reported effort, using Spearman (ρ_s) and Pearson (ρ_p) correlations.

Linear regression modeling was employed to evaluate the relative importance of response process-derived effort versus self-reported effort in predicting PISA performance. To make these effort ratings more comparable, the following min-max scaling procedure was applied to self-reported effort:

$$\mathrm{SE}_n = \frac{y_n - min(y)}{max(y) - min(y)}$$

The resulting self-reported effort score, ranging from 0 to 1, is on a similar scale as the motivation parameter obtained in Study III, which also varies (though continuously) between 0 and 1. To assess the predictive power of each type of motivation variable, the following model was fitted:

$$\begin{aligned}
pv_n &\sim Normal(\mu, \sigma) \\
\mu &= a + \beta_{QM}\theta_n + \beta_{SE}SE_n \\
a &\sim Normal(0, 1000) \\
BE &\sim Normal(0, 1000) \\
SE &\sim Normal(0, 1000) \\
\sigma &\sim Normal(100, 1000), \text{ s.t. } \sigma > 0
\end{aligned}$$

In this model, pv represents the plausible value, θ refers to the posterior mean point estimate of test-takers' motivation derived from response times in Study III, and SE denotes the min-max-scaled self-reported effort. The self-reported effort was not part of the student questionnaire which the response process questionnaire taking effort was based on. The model parameters were estimated using data from all ten reading plausible values. See repository at https://doi.org/10.5281/zenodo.7838627 for data and code sufficient to reproduce the analysis.

Results

The correlation between behavioral and self-reported effort was $\rho_s = 0.17$, and $\rho_p = 0.14$.

Table 2: Parameter estimates

parameter	posterior mean	95% Credible interval
a	114.99	$[104.98, 124.66]$
β_{QM}	344.91	$[334.95, 355.46]$
β_{SE}	85.70	$[81.37, 89.86]$
σ	97.94	$[97.33, 98.57]$

Regression results revealed that both questionnaire motivation and self-reported effort positively contributed to predicting plausible values, with questionnaire motivation being a stronger contributor. A 0.1 increase in θ (questionnaire motivation) correlated with a roughly 35-point increase in plausible value scores, while a 0.1 increase in self-reported effort corresponded to around a 9-point increase in plausible value scores. Therefore, questionnaire motivation was about four times more predictive of performance than self-reported effort. Nonetheless, both variables contributed to performance prediction when considered together.

Discussion

These analyses aimed to explore the relationship between self-reported effort and motivation measured using the approach in Study III. Correlation results suggest a weak connection between self-reported test-taking

effort and their questionnaire-taking motivation as measured by response times.

Regression modeling revealed that both self-reported effort and response process indicated motivation were positively related to performance when considered simultaneously. This indicates that both predictors capture different aspects of test-taking motivation, with motivation to answer the questionnaire being a stronger predictor. One possibility is that the response process more precisely captures test-taking motivation compared to self-reports since it uses information from nearly all questionnaire items, providing a more comprehensive view of test-taker behavior. However, as discussed in Study III, an alternative explanation might be that the questionnaire motivation estimate also captures aspects of reading ability, which explain parts of the stronger relationship.