Navigating Difficulty: The Role of Item Format in Norway's PISA 2018 Mathematical Literacy Scores

Erika Jassuly Chalén Donayre

UNIVERSITY OF OSLO

Master of Science in Assessment, Measurement and Evaluation

40 credits master thesis

CEMO: Centre for Educational Measurement
Faculty of Educational Sciences

Autumn 2023

Popular Abstract

The layout of achievement items classifies them in two main format groups, multiple-choice (MC), where the answer(s) is selected from alternatives, and constructed-response (CR) where the test-taker produces an answer 'from scratch'. The difficulty of an achievement item should stem from the achievement level required to answer the item, rather than from extraneous factors, like format. Otherwise, the uses given to the data would be invalid. Using the mathematics items of PISA 2018 answered by 3122 Norwegian students (50.54% males), we explored whether persons of the same mathematical achievement level have the same chance of answering a MC item and a CR item correctly and if this -potentially- different chance was the same for males and females. CR items were nearly 3-times harder compared to MC items for persons of the same mathematics achievement level and gender. This difference was slightly larger for males. The most difficult items were given in a CR format -with no equivalent MC counterpart and would exhibit a wider range of difficulty levels. Our study contributes to the relatively underexposed research on achievement items compared to research focusing on individuals. Among the implications, the reported relationship between format and how difficult an item is could jeopardise the interpretations given to this data. At the same time, it points to a bigger issue in the educational measurement field: the need for a more systematic approach when developing assessment items.

Acknowledgements

Education and development are two things I strongly believe in. They are also the main reasons why I joined this master's programme. I would like to express my gratitude to both of my supervisors, Johan Braeken and Kseniia Marcq, whose guidance was indispensable for this manuscript. Johan, thank you for sharing your knowledge, time, and patience, along with your attention to detail, and for setting a friendly environment. Kseniia thank you for your straightforward explanations and advice, I found them extremely helpful. I will be forever grateful to you both.

My sincere appreciation extends to the Centre for Educational Measurement staff at the University of Oslo; I deeply admire each one of you, and it has been a great honour for me to study with such dedicated professionals. Thanks to Angel for the conversations and to my classmates, the laughs made this period much more fun. My heartfelt thanks to my parents, siblings, and extended family for their unwavering support and encouragement, and to my husband Sergio for his unconditional support throughout this process. Cheers, God!

All real education is the architecture of the soul -Bennett

Navigating Difficulty: The Role of Item Format in Norway's PISA 2018 Mathematical Literacy Scores

Differences in performance can be related to the item format. This is a controversial issue that poses a problem for the validity of test-score interpretations. Using the mathematics items of PISA 2018, organized in rotated booklets answered by 3122 Norwegian students (50.54% males), explanatory IRT models (considering a response of a person on an item as the outcome measure) were used to focus on whether -and to what extent- people of the same ability had the same probability of answering correctly to a multiple-choice item and a constructed-response item; additionally, it was explored if this potential difference was the same for males and females. Format accounted for 11.5% of the difficulty differences, while gender accounted for <1% of the ability differences. A constructed-response item had approximately 3-times lower odds of being answered correctly than a multiple-choice item ($\beta_1 = -1.09(.37), p = .003; \text{OR} = .0.33$) when comparing people of the same ability and gender. The format difference was slightly larger for males compared to females. The hardest items would only be given in a constructed-response format with no equivalent multiple-choice counterpart, and the variance in difficulty for constructed-response items was 3-times larger than for multiple-choice items. We discuss potential explanations, and implications for the interpretations educational stakeholders give to this data and argue that these types of empirical findings stress the importance of improving item development.

Keywords: item format, mathematics, explanatory IRT, PISA 2018, gender

The total variance of a response can be regarded as a composition of two main sources of variance: the person and the item (Briggs & Wilson, 2007). The sources have received uneven attention, with more literature focusing on person variance compared to item variance. However, the evidence presented by Marcq and Braeken (2022) revealed that, on average, the item variance can be approximately double the person variance. This means that it mattered

more for the correctness of a response which items were responded to by a person, than which person responded to the items. Drawing upon this evidence, the present study focuses on the item side of the total variance of the responses and how this item variance might be related to the item format.

The format (also called item type) refers to its layout. How a person interacts with the item, groups the formats in two broad classifications: if the person selects a response from given alternatives, the format will be called multiple-choice or selected-response. If the person produces a response, the format will be called constructed-response (Haladyna & Rodriguez, 2013). Each item format has garnered both proponents and detractors over the years. Multiple-choice items are highly cost-efficient and reliable. Their structure, when thoughtfully used, makes them a valuable diagnostic tool for identifying student's conceptual understanding as well as misconceptions (Olsen et al., 2001; Tamir, 1990). However, they have been criticized and accused of being limited to assessing low-level thinking (Cronbach, 1970). Constructed-response items have been praised for their authenticity (fidelity) in resembling learning tasks of the target domain, assessing complex-level thinking, and providing more granular information about the construct being measured. However, they have also faced criticism for their significantly higher development costs, and have been accused of eliciting responses based on common knowledge and non-scientific explanations (Haladyna & Rodriguez, 2013).

The possibility of multiple-choice and constructed-response formats having an 'impact' on the difficulty level of an item poses a critical issue in educational measurement. If the probability of a correct response on an item is related to its format instead of the target construct, we would run the risk of making oversimplified and non-valid interpretations (Olsen et al., 2001). An item needs to represent the target construct adequately and not distort its meaning. For example, we want an item to be more difficult due to its required mathematics achievement level, rather than due to extraneous factors unrelated to mathematics. In the latter case, we would be in the presence of construct-irrelevant variance, a threat to the validity of test scores interpretations (American educational research association, 2014). The extent of the validity of interpretations and inferences is of utmost importance for

researchers and policymakers as they deal with issues related to the functioning of educational processes, the identification of areas for improvement, and the implementation of evidence-based solutions. Invalid or oversimplified inferences could jeopardise this work (American educational research association, 2014; Olsen et al., 2001). Furthermore, knowing that format may be related to difficulty stresses the importance of item development theory and a proper and planned item and test design, which increases validity by removing or randomising construct-irrelevant difficulty (Ahmed & Pollitt, 2007; Le Hebel et al., 2017).

Differences in Difficulty Between Items of Different Format: Alleged Factors

Possible reasons underlying the potential differences in difficulty between items of different formats can be divided into two groups: (i) task complexity factors due to the intricacy of mental and physical actions required to answer an item and (ii) factors due to test takers' attitudes and behaviours during test administration.

Task Complexity Factors

Both the multiple-choice and constructed-response formats require a certain level of verbal abilities to be answered correctly. In multiple-choice items, it is necessary to be able to read and interpret the item stem and the alternatives correctly to have a correct response. This is not always easy for test-takers (Schoultz et al., 2001). In constructed-response items, in addition to reading and interpreting the item stem correctly, it is also necessary to express the response in written language. Furthermore, the instructions for some constructed-response items may be more complex than those for multiple-choice items, which would place even greater cognitive demands on the test taker. Therefore, it is plausible that the constructed-response format requires a higher degree of verbal abilities (Haladyna & Rodriguez, 2013). The level of mathematics achievement required to answer a constructed-response item, coupled with the potentially higher verbal abilities required, could increase the overall difficulty of the item and therefore decrease the probability of obtaining a correct response in a constructed-response item compared to a multiple-choice item.

Another potential source for the relation between format and item difficulty is the cognitive behaviours elicited by each format. Cognitive learning theory is scarcely a unified science (Haladyna & Rodriguez, 2013). However, whether one follows learning objectives

taxonomies based on Bloom et al. (1956) classification -composed of the categories of knowledge, comprehension, application, analysis, evaluation, and creation (Anderson & Krathwohl, 2001) or simplified taxonomies, such as that summarised by Haladyna and Rodriguez (2013), which considers three categories: knowledge, skills and abilities; both the multiple-choice and constructed-response formats have the potential to elicit each of the above categories (Haladyna & Rodriguez, 2013). This would suggest that the two formats are equivalent in terms of their difficulty. Thus, there would be no difference in the probability of getting a correct response if a person were given a multiple-choice or a constructed-response format.

Nevertheless, some authors have raised the possibility of non-equivalence of the formats due to potential differences in the mental processes underlying each, particularly concerning memory retrieval. In constructed-response items, the retrieval process is called 'recall' because the information is retrieved 'from scratch'. In multiple-choice items, the retrieval process is called 'recognition' because the information can be retrieved from alternatives. Whether or not recall and recognition are different retrieval processes remains a controversy to the present day (Goecke et al., 2022; Tulving, 1982; Uner & Roediger, 2022). If both retrieval processes are equivalent in both formats, this would suggest that we would not find differences in item difficulty. Conversely, if recall and recognition are different retrieval processes, this would suggest that constructed-response items are more difficult than multiple-choice items due to the absence of alternatives for aiding information retrieval. It has also been argued that constructed-response items require a mental assembly of a product from scratch (Martinez, 1999), which would make them more difficult to answer compared to multiple-choice items, where the mental assembly can rely on alternatives. Overall, if recall and recognition are different retrieval processes and the mental assembly of a product requires a higher cognitive demand, we would be faced with a scenario in which constructed-response items are more difficult than multiple-choice items and therefore, a person would have a lower probability of a correct response on a constructed-response item than on a multiple-choice item.

Test-Taking Attitudes and Behavioural Factors During Test Administration

It has been widely conjectured, especially in the high-stakes assessment literature, that the test-taking strategies - also known as problem-solving strategies or test wiseness - available for each format may be a potential source of differences in difficulty between items of different format (e.g., Katz et al., 1996). For constructed-response items, the approach is often to simply write down the answer. However, it is also possible to use certain test-taking strategies. For example, the individual may check his or her produced response against the cues given by the item stem (plug-in strategy) or use these cues when the answer is unknown (Katz et al., 1996). Furthermore, astute test-takers are likely to know how to craft responses that capitalize on their knowledge and hide any gaps (Martinez, 1999). Multiple-choice items are believed to offer even more opportunities for employing test-taking strategies. For instance, the plug-in strategy in the case of multiple-choice items might be enhanced by using the alternatives as aids in selecting the correct response (working backwards from the response). This could lead to receiving unintentional corrective feedback (Bridgeman, 1992; Katz et al., 2000; Katz et al., 1996). Risk-taking tendencies and guessing in competitive situations may be the test-taking strategy researchers have explored the most. Thorndike and Angoff (1971) discussed two main forms of guessing behaviour: random guessing and guessing based on partial knowledge or cues embedded in the item stem or the alternatives. Both guessing behaviours depend on contextual factors (such as risk-taking perception, and item difficulty) but may also be related to personality (Ben-Shakhar & Sinai, 1991).

As noted by Katz et al. (2000), there is limited empirical evidence regarding the connection between test-taking strategies and the relationship between format and item difficulty. Nevertheless, if test-takers do employ such strategies and if multiple-choice items provide more opportunities for using them, it is plausible that the difficulty of multiple-choice items could be reduced because of the assistance provided by the alternatives. Consequently, individuals might have a higher probability of a correct response in a multiple-choice item compared to constructed-response.

Another potential factor underpinning possible differences in difficulty between items of different formats is test-taking motivation. Wise and DeMars (2005) defined this factor as

the person's engagement and expenditure of energy in achieving the best possible test score. In the absence of risks and competition associated with the perceived personal benefit of a high-stakes assessment context, the test-takers intrinsic motivation often decreases in a low-stakes assessment context. Hence, test-takers might not put forth their best effort in a low-stakes assessment (Wise & DeMars, 2005), potentially decreasing their probability of getting a correct response. On this basis, it might be expected that both formats would elicit comparable levels of test-taking motivation. This would suggest that multiple-choice and constructed-response items would be equivalent in terms of their difficulty, thus both formats would have comparable probabilities of a correct response.

However, the effort invested in answering an item depends partially on its perceived difficulty: the greater the difficulty, the less effort is typically invested (Wigfield & Eccles, 2000; Wise & DeMars, 2005). Constructed-response items may be perceived as more mentally taxing and difficult than multiple-choice items, possibly due to their demand for higher verbal abilities and the need to retrieve and organise information from scratch (Goecke et al., 2022; Haladyna & Rodriguez, 2013; Martinez, 1999). While the inverse relationship between difficulty and effort is debatable, if it were to occur, test-takers would exert less effort when faced with constructed-response items (Wise & DeMars, 2005). Consequently, they would have a lower probability of a correct response in this format compared to multiple-choice.

Additionally, the literature has also mentioned that constructed-response items usually require more time to be answered compared to multiple-choice (Rodriguez, 2003). Consequently, they might be seen as more demanding due to the increased level of persistence required (Siegfried & Wuttke, 2019), which would make them more difficult and hence decrease the test-taker's probability of a correct response in this format. Also, constructed-response items have been accused of inducing more anxiety to the extent that it interferes with cognition, and, consequently, hinders a test-taker's ability to demonstrate proficiency (Martinez, 1999). However, the role of anxiety is debatable in a low-stakes assessment context. One could also argue that anxiety could help performance (e.g., Brady et al., 2018). Nevertheless, if constructed-response items elicit higher anxiety and anxiety is negatively related to performance, individuals would have a lower probability of correct

response on constructed-response items compared to multiple-choice items.

The Relationship Between Item Format and Gender

The literature suggests that the difference in difficulty of a multiple-choice item compared to a constructed-response item may not be the same for males and females of the same ability level. Due to the presumed higher risk-taking and guessing behaviours, males compared to females would guess more in multiple-choice items (Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994), which could make this format 'easier' for males. Due to the presumed lower verbal abilities (e.g., Halpern, 2004; Reilly et al., 2019) and less diligence when answering to a test especially if the stakes are low (DeMars et al., 2013), constructed-response items could be more difficulty for males, compared to females. The combination of these factors would suggest that the difference in the probability of a correct response between multiple-choice and constructed-response items would be larger for males compared to females.

Conversely, due to the presumed lower risk-taking and guessing behaviours, females compared to males would guess less in multiple-choice answers (Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994), which could make this format harder for females. Due to the presumed higher verbal abilities (e.g., Halpern, 2004; Reilly et al., 2019) and more diligence when answering to a test even if stakes are low (DeMars et al., 2013), it is reasonable to suspect that constructed-response items could be easier for females, compared to males. The combination of these factors would suggest that the difference in the probability of a correct response between multiple-choice and constructed-response items would be smaller for females compared to males.

The above factors are not intended to be an exhaustive record of the potential reasons for the relationship between format and item difficulty, but they undoubtedly demonstrate that differences in item difficulty based on format are important, possible, and remain a controversial issue to this day.

The Present Study

The present study explores the relationship between format and item difficulty using the Norwegian sample of an international large-scale assessment in education, the Programme

for International Student Assessment (PISA) 2018, mathematical literacy domain, as a working example. In this low-stakes assessment, data is gathered on the mathematical literacy achievement levels, along with other domains, of 15-year-old students from different countries (OECD, 2019). PISA, being one of the largest large-scale educational assessments (Hopfenbeck et al., 2018), is no stranger to the format controversy. This is suggested by their roughly balanced item pool, consisting of 30 multiple-choice items and 40 constructed-response items for the mathematical literacy domain.

Previous empirical findings addressing this debate show no consensus. Some point to an approximate format equivalency (e.g., Lissitz et al., 2012) while others suggest a relationship between format and difficulty (e.g., Katz et al., 1996). Furthermore, some studies have reported that the association between item format and item difficulty is not the same for males as for females (e.g., Beller & Gafni, 2000). In addition, this topic has been explored with different methodologies which poses a challenge for the comparability of the results. The domains studied diverged, the sample sizes were sometimes small, multiple-choice and constructed-response were - in many cases - disproportionately balanced and most of the research comes from high-stakes contexts. These circumstances not only make it difficult to anticipate what this study might find but also highlight the contribution of our study. To go beyond a mere summary in terms of descriptive statistics, and in line with previous recommendations (e.g., Beller & Gafni, 2000; Martinez, 1999), our study considers both person ability level and item difficulty level, utilizing item response theory models.

Differences in Difficulty Between Items of Different Format

This scenario led to the formulation of our first and most important research question, in which we inquire whether the difficulty of a PISA 2018 mathematical literacy item relates to its format. Therefore, our research design sought to explore the existence and the extent of differences in difficulty between items of different formats.

> ***(RQ1)*** *Do Norwegian students of the same mathematics achievement level have a similar probability of responding correctly to a multiple-choice item and a constructed-response item in the mathematical literacy domain of PISA 2018? If not, to what extent is one format more difficult than the other?*

Although the empirical results do not point to a particular hypothesis, the literature review allows us to formulate certain expectations regarding our results. Responding to a constructed-response format requires a greater degree of verbal abilities (see, e.g., Haladyna & Rodriguez, 2013), as well as retrieving and organising information from scratch (see, e.g., Goecke et al., 2022; Martinez, 1999). Furthermore, the combination of these two circumstances may make the constructed-response format more mentally taxing than the multiple-choice format (Wise & DeMars, 2005). These task complexity factors would then suggest that the constructed-response format would be more difficult to answer correctly, compared to multiple-choice. The administration context factors point to the same hypothesis. The presence of a presumably greater number of test-taking strategies available in the multiple-choice format (e.g., Bridgeman, 1992; Cronbach, 1946; Thorndike & Angoff, 1971) would make this group of items easier to answer compared to constructed-response. The increased level of anxiety (Martinez, 1999) and greater effort due to the more demanding mental requirements of a constructed-response format (Wise & DeMars, 2005) would render them more difficult to answer compared to a multiple-choice format. Considering all the factors presented in the literature review, we expect a relationship between format and item difficulty, and that it would be easier to respond correctly to an item presented in a multiple-choice format than in a constructed-response format. However, we do not have an expectation as to the magnitude of this difference.

The Relationship Between Item Format and Gender

Drawing upon previous literature, which hinted that some alleged potential factors underlying the potential relationship between format and difficulty are not the same for males compared to females, (see, e.g., Beller & Gafni, 1996, 2000; Ryan & DeMark, 2002), the present study considered gender as a potential moderator of said relationship. This led us to our second research question:

> ***(RQ2)*** *Is the potential difference in the probability of responding correctly to a multiple-choice item and a constructed-response item in the mathematical literacy domain of PISA 2018 the same for Norwegian male and female students who have the same mathematics achievement level? If not, to what extent is the difference in*

difficulty between items of different formats larger (or smaller) for one gender compared to the other?

The literature suggests that the difference in difficulty of a multiple-choice item compared to a constructed-response item may not be the same for males and females of the same ability level. Due to the presumed higher risk-taking and guessing behaviours of males, males compared to females would guess more in multiple-choice items (Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994), which could make this format easier for males and, conversely, harder for females. Due to the presumed higher verbal abilities (e.g., Halpern, 2004; Reilly et al., 2019) and more diligence of females, compared to males, when answering to a test (DeMars et al., 2013), constructed-response items could be more difficulty for males and, conversely, easier for females. The combination of these factors would suggest that the difference in the probability of a correct response between multiple-choice and constructed-response items would be larger for males than for females.

Method

PISA 2018 contains both explanatory variables for this study: item format and person gender. PISA also offers a roughly balanced item pool (30 multiple-choice items and 40 constructed-response items) answered by a fairly large and random sample of persons (OECD, 2020a, 2020b). The previous reasons aligned with our focus to explore the item side of the total variance of the responses and its potential relation with the format. Consequently, we considered this secondary data to be well-suited to our purpose.

Sample

PISA 2018 Items

PISA 2018 was administered as a computer-based (CBA) assessment in Norway. The mathematical literacy domain item pool consisted of 70 items: 30 multiple-choice items and 40 constructed-response items. The multiple-choice items included single-selection (simple) multiple-choice and complex multiple-choice. The latter refers to a table with affirmations and a certain amount of yes/no or true/false options. All multiple-choice items were computer-scored. The constructed-response items required the person to show the steps taken

or some other form of extended written response. Roughly half of the constructed-response items were computer-scored and the other half were scored by human raters (OECD, 2020a, 2020b).

The 70 items covered four content knowledge areas: change & relationships, quantity, space & shape, and uncertainty & data. The items were assigned to six blocks or clusters (M01, M02, M03, M04, M05, M06), with each item appearing exclusively in one cluster. The clusters were roughly balanced in terms of their coverage of content knowledge areas, cognitive processes, situation or context of the question, number of items and item parameters. Each cluster was designed to take approximately 30 minutes to complete. After designing the clusters, they were rotated to generate a total of 72 booklets (test forms). Each booklet comprised four clusters. These booklets were then randomly assigned to test-takers. Each cluster appeared at least once in each of the four possible positions in a booklet. Each cluster pair appeared only once together in a booklet and not all possible cluster pairs appeared together (OECD, 2020a, 2020b).

PISA 2018 Persons

PISA 2018 Norway followed a two-stage stratified sampling design to sample students born in 2002 who were attending educational institutions in grades 7 and higher. In the first stage, a national list of schools with these students or schools with the possibility of having these students at the time of the assessment was created. 251 schools were sampled with probabilities proportional to the size (PPS sampling) of their 15-year-old students, considering explicit stratification variables. Once schools were sampled, each made a list of their 15-year-old students from which 35 were sampled with an equal probability of being selected. When the lists had fewer than 35 persons, all of them would be selected. The exclusion rate at school and student levels combined was 7.88% of the desired target population. Therefore, a sample of 5813 students was drawn from the target population of 15-year-old students (OECD, 2020b, 2020c). The total Norwegian sample was balanced in terms of gender, with a roughly equal proportion of males (50.46%) and females (49.54%).

As a result of the rotated cluster design used to structure the booklets randomly assigned to students, approximately 54% of the Norwegian sample received a subset of the

item pool (OECD, 2020b). The item distribution across booklets allowed a sufficient and efficient exposure of items to the person sample (Braeken, 2016) and reduced test burden. The 'planned missingness' design aligns with PISA's targeted inferences at the country level, where every single item doesn't need to be responded to by every single person. Therefore, the effective sample size was $n = 3122$ (50.54% males and 49.46% females) distributed across 250 schools. On average, each person responded, to a subset of 21 mathematics items ($\min = 1, \max = 24$) and 931 persons, on average, responded to each item ($\min = 599, \max = 971$).

Modelling Framework

A single response in our response data was considered a combination of a person answering an item. Responses, the lower-level data units were formed by pairs that resulted from crossing two higher-level data units: persons and items, reflecting the cross-classified data structure. Each person responded to several items, and each item was responded to by several persons. Consequently, responses were nested both within persons and items (Van den Noortgate et al., 2003).

In item response theory (IRT), responses are modelled as a function of two factors: person ability and item difficulty. If persons and items are considered as random samples drawn from a population of items and a population of persons respectively, two random residuals can be outlined, one for persons and one for items, leading us to a random-person random-item item response approach (De Boeck & Wilson, 2004; Van den Noortgate et al., 2003). We formulated a cross-classified mixed effects logistic regression model (a reformulation of the one-parameter-logistic IRT model), where we allowed the probability of a correct response to vary across persons and items and defined the descriptive null model as follows:

$$Logit(\pi_{pi}) = \beta_0 + \theta_p + \beta_i, \tag{1}$$

with

$$\theta_p \sim N(0, \sigma_\theta^2) \text{ and } \beta_i \sim N(0, \sigma_\beta^2), \tag{2}$$

where π_{pi} is the probability that person p will answer item i correctly. β_0 is the overall intercept (fixed effect), which corresponds to the estimated logit for the probability of a correct response for an average-ability student on an average-easiness item. θ_p is the person-specific deviation or person ability (random effect). β_i is the item-specific deviation or item easiness (random effect). Each specific deviation (also called varying intercept) was assumed to follow an independent normal distribution with mean=0 and variances σ^2_θ and σ^2_β, respectively. Note that the plus sign in β_i, implies that β_i should be interpreted in the equation as item easiness instead of item difficulty.

The model suggests that the total variance of the responses is partitioned into three parts:

$$\sigma^2_{total} = \sigma^2_\theta + \sigma^2_\beta + \frac{\pi^2}{3}, \tag{3}$$

where σ^2_θ and σ^2_β refer to the variances of the person and item varying intercepts (random effects), and $\frac{\pi^2}{3}$ is the residual variance, the item response variation that cannot be accounted for by systematic person and item differences and is due to more idiosyncratic elements or random events that play a role when a specific person responds to a specific item (e.g., unexpected relevant item-specific background knowledge or the occasional distraction) This residual variance is specific to the standard logistic distribution due to the applied link function which considers the binary responses: correct and incorrect.

Our first research question inquires about differences in difficulty between items of different formats for persons of the same mathematics achievement level. Our second research question explores if and to what extent these potential differences in difficulty between items of different formats are the same for males and females of the same mathematics achievement level. To address our research questions, the null descriptive model was extended by adding regression layers relating covariates of interest to item difficulty and person ability. The purpose of this approach, known as explanatory IRT, is to explain the responses in terms of other variables (De Boeck & Wilson, 2004). Therefore, the item difficulty was predicted by format and the person ability was predicted by gender. Similar to the null descriptive model, in this explanatory IRT extension, the probability of a correct response is a function of both

item difficulty and person ability. Therefore, this model accounts for the possibility that variation in item difficulty could be related to differences in item format and that variation in person ability could be related to a person's gender.

Specifically, to address our second research question, which explores if and to what extent the potential difference in the probability of responding correctly to a multiple-choice item and a constructed-response item is the same for males and females of the same mathematics achievement level, we tested an interaction effect between the item side predictor format and the person side predictor gender. This interaction effect was formulated with a cross-product second-order term composed of both predictors. This analysis would imply that the effect of format on item difficulty depends on the gender group.

We formulated our explanatory cross-classified mixed effects logistic regression model with an interaction effect as follows:

$$Logit(\pi_{pi}) = \beta_0 + \beta_1 format_i + \beta_2 gender_p + \beta_{12} format_i * gender_p + \theta_p + \beta_i, \tag{4}$$

with

$$\theta_p \sim N(0, \sigma_\theta^2) \text{ and } \beta_i \sim N(0, \sigma_\beta^2), \tag{5}$$

where β_0 is the general intercept (fixed effect). The added terms β_1, β_2 and β_{12} are also fixed effects. β_1 is the regression weight for item format $format_i$, β_2 is the regression weight for the gender of the person $gender_p$, β_{12} is the regression weight for the interaction between $format_i$ and $gender_p$. The person-specific deviation θ_p and the item-specific deviation β_i correspond to the residual variances in the person ability and item difficulty after accounting for gender and format predictors.

Measures and Statistical Analysis

Our outcome variable was defined as the response of a person to an item. 'System missing/blank' and 'not applicable' responses were treated as missing by design. From the remaining responses (valid responses), approximately 4% were 'Not reached' and regarded as

missing-at-random (Mislevy & Wu, 1996). Approximately 7% of the valid responses were regarded as 'no response' and scored as incorrect because it was reasoned that if the person had the opportunity to answer an item but did not do so, it was likely that the person did not know the answer. For 7 constructed-response items that allowed partial credit, partial credit was scored as incorrect. Therefore, all responses were scored as binary (0=incorrect, 1=correct), facilitating comparability across items. For the item side predictor format, we grouped the simple and complex multiple-choice items under a single group: multiple-choice, coded as 0. This was done because the amount of choices taken in a multiple-choice layout was not part of our research questions. Constructed-response items were coded as 1. For the person side predictor, gender, females were coded as 0 and males as 1. For all variables, the reference group was the one coded as 0.

We fitted several cross-classified mixed effects logistic regression models, including a heteroscedastic one, which allowed a specific variance for each of the item groups defined by format and for each of the person groups defined by gender. A marginal maximum likelihood estimation approach with the lme4 package (Bates et al., 2015) in R (Team, 2020) version 4.3.1 was used. To compare the models, Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the likelihood-ratio test were used. Lower AIC, BIC values indicated a better fit, while a significant result in the likelihood-ratio test indicated a better fit.

Results

Descriptive Null Model

Following our baseline model (model 0, see Table 1), the descriptive random-person random-item response model, a Norwegian 15-year-old student of average latent ability was estimated to have a probability of .43 (i.e., $\Pr(Y_{pi} = 1|\theta_p = 0, \beta_i = 0) = \frac{1}{1+\exp(-\beta_0)}$) to respond correctly to an item of average difficulty in the PISA 2018 mathematics achievement test. About 21% ($\sigma^2_\theta = 1.51$) of the item response variation was attributed to differences in ability between individual students. About one-and-a-half times more variance, 35% ($\sigma^2_\beta = 2.55$), was attributed to differences in difficulty among the mathematics item set. Thus, for the correctness of an item response, it mattered more to know which item was responded to by a student than which student was responding. The remainder 45% of item response variation

cannot be accounted for by systematic student and item differences and is due to more idiosyncratic elements or random events that play a role when a specific student responds to a specific item (e.g., unexpected relevant item-specific background knowledge or the occasional distraction). The latter residual item response variation might appear high at first sight, but keep in mind that PISA is a low-stakes assessment covering a broad range of mathematics contents. These variance component percentages are comparable to results in the literature on international large-scale assessments in education for other domains than mathematics and other countries than Norway (Marcq & Braeken, 2022).

Table 1

Overview of the Estimated Explanatory Item Response Models

	Model 0	Model 1	Model 2	Model 3	Model 4	Model 4c
Fixed effects						
Intercept β_0	-.28 (.19)	.34 (.27)	-.25 (.19)	.38 (.28)	.33 (.27)	.33 (.19)
Format β_1		-1.09** (.36)		-1.09** (.37)	-1.01** (.36)	-1.01** (.34)
Gender β_2			-.07 (.05)	-.07 (.05)	.01 (.05)	.01 (.05)
Format*Gender β_{12}					-.16*** (.04)	-.17*** (.04)
Random effects						
Student σ^2_θ	1.51	1.51	1.51	1.51	1.51	F:1.34 M:1.69
Item σ^2_β	2.55	2.25	2.55	2.25	2.25	MC:1.1 CR:3.13
Model comparison						
Deviance	66489	66481	66487	66478	66463	66442
AIC	66495	66489	66495	66488	66475	66458
BIC	66522	66525	66531	66534	66530	66531

Note. For the fixed-effect estimated parameters, standard errors are provided in between brackets. For the heteroscedastic model 4c, the variance parameter ability can vary between the female group (F) and the male group (M), and the variance parameter difficulty can vary between the multiple-choice group (MC) and the constructed-response group (CR). The other models have a homoscedastic variance applicable to both person groups and another homoscedastic variance applicable to both item groups. The *,**, and *** correspond to $p < .05, .01$, and .001, respectively; where p-values are connected to a default null hypothesis of parameter equals zero. See Equation 4 for the model formulation.

Explanatory Item Response Models

Main Effects Model

To study how a student's gender relates to their latent mathematics ability as measured by PISA, gender was brought in as a dummy predictor (0=female, 1=male) at the person side of the descriptive baseline item response model. Similarly, to study how an item's format relates to their difficulty in the PISA mathematics assessment, format (0=multiple-choice, 1=constructed-response) was brought in as a dummy predictor at the item side of the baseline model. Compared to the descriptive baseline model, the main effects model (model 3, see Table 1) was considered the better-fitting model ($LRT(\Delta df = 2) = 10.71,\ p = .005$).

When comparing people of the same ability and gender, the odds of responding correctly to a constructed-response (CR) item were estimated as being statistically significant 2.99-times lower ($\beta(\text{Format=CR}) = -1.09(.37), p = .003; \text{OR} = .0.33$) than for a multiple-choice (MC) item. An average individual on an average multiple-choice item has a .58 probability of a correct response, while an average individual on an average constructed-response item has a .32 probability of a correct response. Format differences among items were estimated to account for 11.5% of the differences in difficulty among items ($\sigma^2_\beta = 2.25$; $LRT(\Delta df = 1) = 8.54,\ p = .003$).

When responding to an item of the same difficulty and format, the odds of responding correctly for male students were estimated as being 1.07 times, but not statistically significantly lower ($\beta(\text{Gender=Male}) = -.07(.05), p = .14; \text{OR} = .93$) than that for female students. The probability of a correct response for females and males is fairly equivalent: an average-ability female student responding to an average item has a .44 probability of a correct response, while a male student responding to an average item has a .42 probability of a correct response. Gender differences among students were estimated to account for less than 1% of the individual differences in ability among students
($\sigma^2_\theta = 1.51$; $LRT(\Delta df = 1) = 2.17,\ p = .141$).

Interaction Model

Our second research question inquired if (and to what extent) the difference in probability of responding correctly to a multiple-choice item and a constructed-response item

in the mathematical literacy domain of PISA 2018 was the same for Norwegian male and female students who have the same mathematics achievement level (same ability). Specifically, we explored the functioning of the two item groups -multiple-choice and constructed-response- for males versus females. This involved testing if the differences in difficulty between items of different formats were larger (or smaller) for one gender compared to the other, subsequently influencing the probability of a correct response.

To answer the second research question, we ran model 4 (see Table 1) which included an interaction between the item covariate (format) and person covariate (gender) to test if the potential difference in the probability of responding correctly to a multiple-choice item and a constructed-response item is the same for males and females of the same mathematics achievement level. Gender had the role of moderator, following the literature review previously exposed.

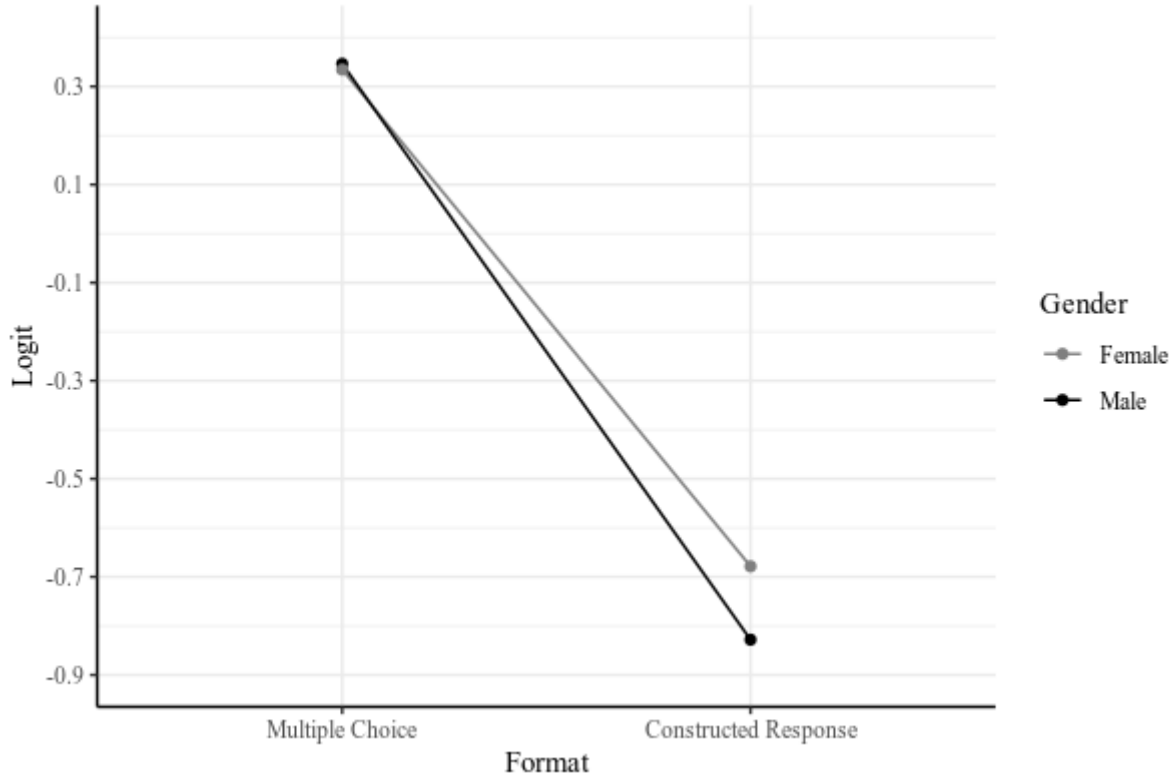
Compared to the 'main effects' model 3, model 4 fitted the data better ($LRT(\Delta df = 1) = 15.2,\ p < .001$). The significant interaction term β_{12} = -.16 suggested that the difference in one covariate depends on the value of the other covariate. The expected average difference in the logits of the probability of a correct response due to differences of an item being constructed-response, compared to multiple-choice, is larger for males versus females (see figure: steeper black line compared to the grey line); because constructed-response is a little more difficult for males. Whereas for multiple-choice the difficulty remains comparable across genders.

The simple slope of the format suggests that constructed-response is harder to answer than multiple-choice, regardless of which same-ability gender group is being compared. This means that an average-ability female responding to a constructed-response item would have significantly 2.75-times lower odds to respond correctly than an average-ability female responding to a multiple-choice item ($\beta_1 = -1.01(.36), p = .005; \text{OR} = .0.36$). An average-ability male responding to a constructed-response item would have significantly 3.24-times lower odds to respond correctly than an average-ability male responding to a multiple-choice item ($\beta_{1\cdot} = -1.17(.36), p = .001, \text{OR} = .0.31$).

The results did not support a gender difference in responding correctly to

Figure 1

Difference in the Logit of the Probability of a Correct Response



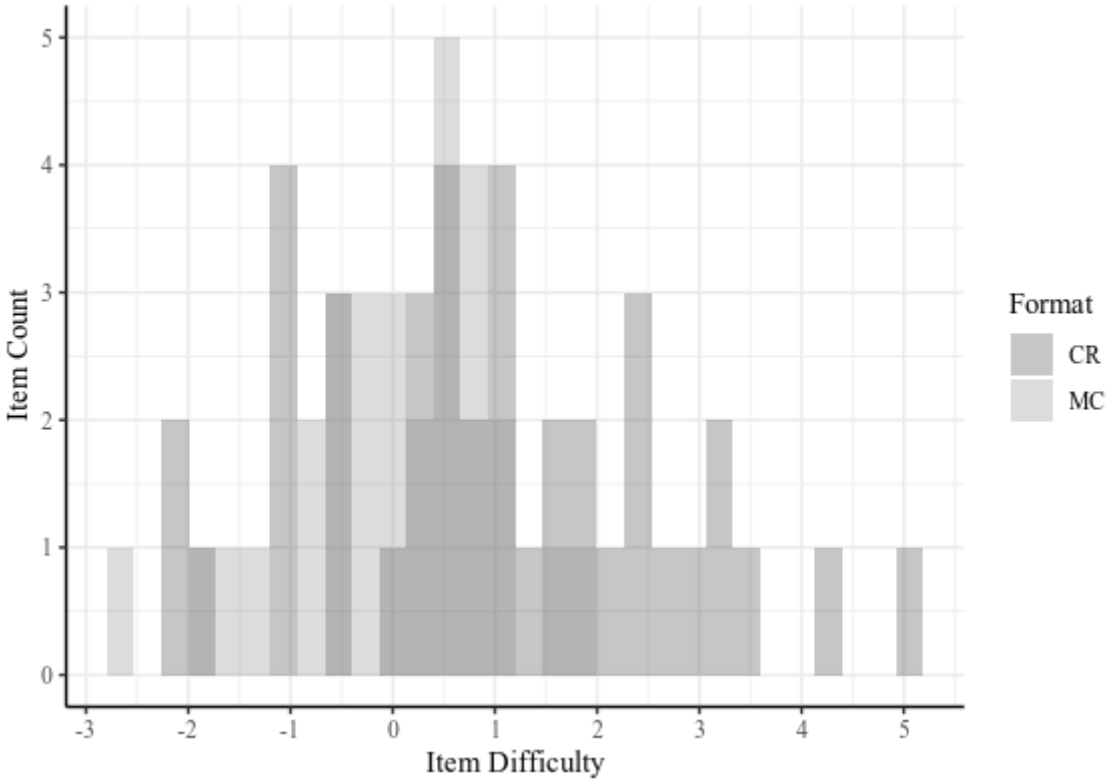
Note. The dots represent four pseudo data points: an average-ability female answering a multiple-choice item, an average-ability female answering a constructed-response item, an average-ability male answering a multiple-choice item, and an average-ability male answering a constructed-response item.

multiple-choice items of the same difficulty level ($\beta_2 = .01(.05), p = .825; \text{OR} = 1.01$). Yet for constructed-response items of the same difficulty level, there was a small gender difference that favoured females ($\beta_{2.} = -.15(.05), p = .004; \text{OR} = 0.86$).

So far, the models assumed the residual variation in item difficulty to be equivalent in both format item groups (from the item side of the total variance of a student's response) and the residual variation in math ability to be equivalent for both gender groups (from the person side of the total variance of a student's response). A heteroscedastic model 4c (see Table 1) was also fitted. This model allowed the residual variation of item difficulty to be different between multiple-choice and constructed-response items and the residual variation of person ability to be different between males and females. Compared to the interaction model 4, the heteroscedastic model 4c fitted the data better ($LRT(\Delta df = 1) = 21,\ p < .001$) and the fixed

Figure 2

Differences in Difficulty for Multiple-Choice Items Compared to Constructed-Response Items



Note. Based on heteroscedastic model 4c, which allowed a specific variance for each of the item groups defined by format. The darker grey colour indicates an overlap between the variance in difficulty for the two formats.

effects remained robust.

Model 4c suggested that the residual variance of multiple-choice items is three times smaller in magnitude ($\sigma^2_\beta = 1.1$) compared to the residual variance of constructed-response items($\sigma^2_\beta = 3.13$), which translates in multiple-choice items being more alike in difficulty than constructed-response. Figure 2 shows that for every multiple-choice item, we have a constructed-response equivalent-difficulty item. However, the reverse does not apply, as there are only constructed-response formats for the hardest items on the right side of the ability latent dimension, ranging from 2 to 5. For gender, the model suggested that the residual variance of males is slightly larger ($\sigma^2_\theta = 1.69$) than the residual variance of females ($\sigma^2_\theta = 1.34$).

Discussion

Using the Norwegian sample from the PISA 2018 for the mathematical literacy domain, the focus of the present study was to explore if -and to what extent- the difficulty of an item relates to its format. As reviewed in the introduction, the format 'effect' debate is fueled by factors that may underlie the potential differences in difficulty between multiple-choice and constructed-response items. Which led us to our first research question:

> ***(RQ1)*** *Do Norwegian students of the same mathematics achievement level have a similar probability of responding correctly to a multiple-choice item and a constructed-response item in the mathematical literacy domain of PISA 2018? If not, to what extent is one format more difficult than the other?*

The greater degree of verbal abilities (see, e.g., Haladyna & Rodriguez, 2013) and retrieving and organising the information from scratch in constructed-response items (e.g., Goecke et al., 2022; Martinez, 1999) compared to multiple-choice items, could potentially make constructed-response items more difficult. Moreover, the combination of said factors could make the constructed-response format more mentally taxing than the multiple-choice format (Wise & DeMars, 2005). In addition, factors related to test administration, such as the presumably greater number of test-taking strategies available in the multiple-choice format (e.g., Bridgeman, 1992; Cronbach, 1946; Thorndike & Angoff, 1971) would make this group of items easier to answer compared to constructed-response. The increased anxiety (Martinez, 1999) and greater effort due to the more demanding mental requirements of a constructed-response format (Wise & DeMars, 2005) would render constructed-response items more difficult to answer compared to multiple-choice items. Therefore, we expected a relationship between format and item difficulty and that responding correctly to a multiple-choice item would be easier than responding correctly to a constructed-response item, for persons of the same achievement level. Consequently, persons of the same mathematics achievement level would have a lower probability of responding correctly to items presented in a constructed-response format. However, we did not have an expectation about the magnitude of this difference.

Given that previous literature suggested that some alleged factors underlying the potential format 'effect' on item difficulty are not the same for males compared to females, (e.g., Beller & Gafni, 1996, 2000; Ryan & DeMark, 2002), we considered gender as a potential moderator of the relationship between format and difficulty. This led us to our second research question:

> ***(RQ2)*** *Is the potential difference in the probability of responding correctly to a multiple-choice item and a constructed-response item in the mathematical literacy domain of PISA 2018 the same for Norwegian male and female students who have the same mathematics achievement level? If not, to what extent is the difference in difficulty between items of different formats larger (or smaller) for one gender compared to the other?*

The literature suggests that the difference in difficulty of a multiple-choice item compared to a constructed-response item may not be the same for males and females of the same ability level. This is due to the presumed risk-taking and guessing behaviour, verbal abilities and diligence (Ben-Shakhar & Sinai, 1991; DeMars et al., 2013; Gafni & Melamed, 1994; Halpern, 2004), which would make multiple-choice items easier for males and, conversely harder for females, and constructed-response items more difficult for males and conversely easier for females. The combination of these factors would suggest that the difference in the probability of a correct response between multiple-choice and constructed-response items would be larger for males compared to females.

We utilized an explanatory item response modelling approach to address our two research questions, where we sought to relate format to difficulty and gender to ability. An interaction effect between format and gender was included to test if the differences in difficulty between items of different formats were larger (or smaller) for one gender compared to the other.

Our expectation that it would be easier to respond correctly to an item presented in a multiple-choice format (compared to a constructed-response format) was supported and well aligned with previous high-stakes and low-stakes achievement research (e.g., Beller & Gafni, 2000; El Masri et al., 2017; Le Hebel et al., 2017). However, we did not expect the magnitude

of this difference to be so prominent, with the odds of answering correctly to an item presented in a constructed-response format being almost three times lower than the odds of answering correctly to an item presented in a multiple-choice format. In addition, it was also not expected that the more difficult items would only be given in a constructed-response format with no equivalent multiple-choice counterpart, nor that the group of multiple-choice items would have less variation. Below we provide an integrated discussion of our results.

Norwegian students are well versed in constructed-response items compared to multiple-choice items, as a quick look at the assessment tools of the Norwegian education system would show (e.g., Utdanningsdirektoratet, 2023). However, it appears that factors other than familiarity are more important for the probability of responding correctly to items of different formats. Arising from high-stakes achievement literature, a potential explanation for the higher probability of a correct response on an item presented in a multiple-choice format could be rooted in the availability of varied test-taking strategies that this format offers (e.g., Cronbach, 1946; Katz et al., 1996; Thorndike & Angoff, 1971). These strategies may compensate for the lack of achievement level required to respond to an item of a planned difficulty level, consequently reducing the item difficulty level during the test situation. This poses a challenge for test design, by making it harder to develop difficult multiple-choice items that tap complex mathematical literacy content in comparison to constructed-response items, which some authors argue are more suited to capture complex content (Haladyna & Rodriguez, 2013; Hancock, 1994; Rauch & Hartig, 2010).

Perhaps more research is needed on plausible distractors in multiple-choice format and how students interact with them, as noted by Olsen et al. (2001) and Twist and Fraillon (2020). Similarly, the verbal abilities required in a constructed-response item could potentially increase the difficulty level of an otherwise easier item (Haladyna & Rodriguez, 2013). The above factors could explain why items presented in a constructed-response format have such a lower probability of a correct response and also why Figure 2 presents the group of the most difficult items exclusively in a constructed-response format. A question that lingers is whether the use of test-taking strategies successfully transfers from high-stakes to low-stakes contexts, or if we are merely dealing with strategies that are generic to solving achievement tests.

It is therefore plausible that the reasons for the observed format 'effect' in item difficulty may be attributed to the test design, the test-taking attitudes and behavioural factors during test administration, or a combination of both. PISA item developers may favour, intentionally or unconsciously, the presentation of complex mathematical content items with a constructed-response format (because it is 'easier' to design difficult items with this format) and the presentation of simple mathematical content items with a multiple-choice format (because it is 'easier' to design easy items with this format). However, the format 'effect' may also be generated in the context of a person responding to the items: the difficulty of an item presented in a multiple-choice format may be reduced by the strategies available and/or the difficulty of an item presented in a constructed-response format may be increased by the verbal abilities required. Both would create a more pronounced difference between the two formats. Whether the reasons for the observed format 'effect' are due to the design, the behavioural factors during the test administration, or both, PISA item developers - and other stakeholders using these data - should be aware of the presence of construct-irrelevant variance in the data due to a format 'effect' that contaminates differences in the achievement levels of test-takers.

This study has demonstrated that format plays a significant role in the level of difficulty of an item and that it can be associated with the probability of a person responding correctly to an item. Furthermore, the study has drawn our attention to factors that influence the correctness of a response, which is crucial for understanding the constructs being assessed and for ensuring the validity of test score interpretations by preventing the inclusion of construct irrelevant item attributes (De Boeck et al., 2016). Future research may help to identify and better understand variation in formats. Given the limitations of the PISA design, which prevents the isolation of individual effects of test-taking strategies (presumably more pronounced in multiple-choice items) and the separation of mathematical and verbal abilities (presumably more pronounced in constructed-response items), experiments could develop content-parallel format items aimed at measuring the same achievement level, manipulate each test-taking strategy and attempt to separate the verbal abilities from the mathematical abilities. Process data could prove to be a valuable tool in this endeavour, although particular care needs to be taken in its planning and integration into the assessment development process

(Twist & Fraillon, 2020).

The present study used variance component models which showed that which items were answered was more important for a correct answer than who answered those items. Our results pertain to the definitions and specific layouts of multiple-choice and constructed-response items used in the PISA 2018 mathematical literacy domain in Norway. In this context, the format accounted for a significant portion of the differences in item difficulty, with constructed-response items being harder than multiple-choice items. We considered our inferences to be, to some extent, supported by the reasonably large and format-balanced item sample (pool) answered by a somewhat random sample of persons and the fair robustness of our results after testing different approaches to handling missing data and scoring partial credit responses (see sensitivity analysis in supplemental material). Considering that a segment of prior research aligns with our results, particularly about constructed-response items being more difficult than multiple-choice items, it's conceivable that our findings could be replicated in future studies analyzing assessments across different countries, multiple-choice and constructed-response layouts, domains, stakes and other scenarios. Furthermore, given that the item format accounted for a relevant portion of the item variance, it is highly likely that other item properties (e.g. item content, length of text, context), which have been overlooked in the literature (Marcq & Braeken, 2022), could also further contribute to explaining the item difficulty.

Our second research question explored whether gender acts as a moderator of the relationship between item format and item difficulty. In other words, we inquired whether the difference in difficulty of a multiple-choice item compared to a constructed-response item would be the same for males and females of the same ability level. Which consequently, would suggest that the difference in the probability of a correct response between multiple-choice and constructed-response items would be larger for males than for females. Our expectation that the difference in the probability of a correct response between multiple-choice and constructed-response items would be larger for males than for females was met, however this 'gender difference in the format difference' was rather small. This was the case because for multiple-choice the probability of a correct response remained comparable across genders,

whereas for constructed-response there was a slight difference in favour of females.

The gender difference in the multiple-choice format was contrary to what the literature suggests (Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994). It is possible that due to the low-stakes context, males did not perceive any risk, and therefore their effort to guess the correct answer in multiple-choice items declined (DeMars et al., 2013; Wise & DeMars, 2005). It is also possible that both males and females have the same risk-taking and guessing behaviours when solving multiple-choice items, which would align with Norway's aim of promoting gender equality in education (Regjeringen, 2021). The gender difference in the probability of a correct response for constructed-response items was quite small and favoured females. The combination of a higher demand for verbal abilities and the low-stakes assessment context may potentially have reduced males' engagement, leading to a lower probability of a correct response in constructed-response items. Similarly, the greater verbal abilities and diligence in answering constructed-response items may have increased females' probability of a correct response (DeMars et al., 2013; Wise & DeMars, 2005). While there may be other possible reasons for the interaction between format and gender, our study showed that gender variations in moderating the relationship between format and item difficulty likely do not pose a problem for Norway.

The results of this study are a tangible demonstration of the importance of format equivalence for the validity of test-score interpretations. Ideally, there should be no differences in item difficulty due to the format employed, but we found that there are. Especially for researchers who collaborate closely with policymakers and delve into topics like the functioning of educational processes, identifying areas for enhancement, and suggesting evidence-based solutions, ensuring that inferences are drawn from precise data becomes crucial. This can only be achieved by eliminating possible sources of construct-irrelevant variance (American educational research association, 2014; Olsen et al., 2001).

Another implication of this study is the importance of proper and planned item design in assessments in general (Haladyna & Rodriguez, 2013), and large-scale assessments in particular. If a multiple-choice format decreases the difficulty level of an item and a constructed-response format increases it, then maybe, neither can maintain the level of

difficulty that the item originally aimed for (before it was presented in either format). To address this issue, a possible solution would be to balance the confounding effects of the format for the various item difficulty levels. This would include item developers avoiding the overuse of constructed-response formats for the difficult items and the overuse of multiple-choice formats for the easy items. Of course, practical concerns (time and financial resources) may arise because the item pools would have to be much larger than they are now. However, we believe that the advantages of a more systematic design outweigh the disadvantages. To simplify this process, a helpful approach would be to utilize the difficulty parameters acquired from previous cycles in order to restructure the item clusters. The reuse of certain items by PISA makes this even more pertinent (OECD, 2020b). Procedures like these could facilitate the development of test forms of equivalent difficulty, addressing the current lack of such uniformity. That being said, our intention is not to undermine the extensive item-level analyses carried out by PISA as a way to ensure the quality of items. Instead, we want to redirect the focus to a general issue that greatly affects the educational measurement field: it is the fact that we still have a significant amount of progress to make in the theory and practice behind the development of assessment items (Haladyna & Rodriguez, 2013). This provides a rich opportunity for future educational assessment research.

Conclusion

This study was motivated by the format 'effect' controversy that affects the educational measurement field to the present day. Potential factors underlying the format 'effect' could either reduce or increment the difficulty of an item. In either scenario, a relationship between format and difficulty would suggest a multiple-choice format and a constructed-response format, the measurement tools, are not equivalent. Which poses a validity issue for the test-score interpretations. Using the Norwegian sample of the PISA 2018 mathematical literacy domain as a working example, the present study provided evidence for differences in difficulty between items of different formats, suggesting that constructed-response items can be nearly 3-times harder compared to multiple-choice items for persons of the same ability level. This main finding should, at the very least, make education stakeholders more cautious about the interpretations and uses ascribed to these test scores. Furthermore, the most difficult

items were exclusively of a constructed-response format and the format predictor accounted for a meaningful amount of the difficulty differences among items. Taken together, these findings point to a larger issue: our field needs a more systematic approach to developing assessment items -or balancing their effects if these cannot be eliminated. Otherwise, we run the risk of making overly simplistic interpretations (Olsen et al., 2001).

References

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: An experimental investigation of focus. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 201–232. https://doi.org/10.1080/09695940701478909

American educational research association (Ed.). (2014). *Standards for educational and psychological testing* (3rd ed.). American educational research association. https://www.testingstandards.net/open-access-files.html

Anderson, L., & Krathwohl, D. (Eds.). (2001). *Taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (1st ed.). Longman.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beller, M., & Gafni, N. (1996). The 1991 international assessment of educational progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology*, *88*(2), 365–377. https://doi.org/10.1037/0022-0663.88.2.365

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1), 1–21. https://doi.org/10.1023/A:1007051109754

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23–35. https://doi.org/10.1111/j.1745-3984.1991.tb00341.x

Bloom, B. S., Engelhart, M. D., Furst, E., Hill, W., & Kratwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook 1: Cognitive domain* (1st ed., Vol. 1). McKay.

Brady, S. T., Hard, B. M., & Gross, J. J. (2018). Reappraising test anxiety increases academic performance of first-year college students. *Journal of Educational Psychology*, *110*(3), 395–406. https://doi.org/10.1037/edu0000219

Braeken, J. (2016). International large-scale educational assessments: Elephants at the gate? In M. Nordengen & H. Thorsen (Eds.), *Northern lights on PISA and TALIS* (pp. 195–216). Nordic Council of Ministers 2016. https://doi.org/10.6027/TN2016-517

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, *29*(3), 253–271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131–155. https://doi.org/10.1111/j.1745-3984.2007.00031.x

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. https://doi.org/10.1177/001316444600600405

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

De Boeck, P., Cho, S.-J., & Wilson, M. (2016). Explanatory item response models. In *The Wiley Handbook of Cognition and Assessment* (pp. 247–266). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118956588.ch11

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach* (1st ed.). Springer New York. https://doi.org/10.1007/978-1-4757-3990-9

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

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: The case of key stage 2 assessments. *The Curriculum Journal*, *28*(1), 59–82. https://doi.org/10.1080/09585176.2016.1232201

Gafni, N., & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation*, *20*(3), 309–319. https://doi.org/10.1016/0191-491X(94)90018-3

Goecke, B., Staab, M., Schittenhelm, C., & Wilhelm, O. (2022). Stop worrying about multiple-choice: Fact knowledge does not change with response format. *Journal of Intelligence*, *10*(4), 102. https://doi.org/10.3390/jintelligence10040102

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (1st ed.). Routledge. https://doi.org/10.4324/9780203850381

Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science*, *13*(4), 135–139. https://doi.org/10.1111/j.0963-7214.2004.00292.x

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, *62*(2), 143–157. https://doi.org/https://www.jstor.org/stable/20152406

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. https://doi.org/10.1080/00313831.2016.1258726

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of Response Format on Difficulty of SAT-Mathematics Items: It's Not the Strategy. *Journal of Educational Measurement*, *37*(1), 39–57. https://doi.org/10.1111/j.1745-3984.2000.tb01075.x

Katz, I. R., Friedman, D. E., Bennett, R. E., & Berger, A. E. (1996). Differences in strategies used to solve stem-equivalent constructed-response and multiple-choice SAT® mathematics items. *ETS Research Report Series*, *1996*(2), i–20. https://doi.org/10.1002/j.2333-8504.1996.tb01698.x

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: Example of PISA science items. *International Journal of Science Education*, *39*(4), 468–487. https://doi.org/10.1080/09500693.2017.1294784

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*(3). https://eric.ed.gov/?id=EJ1001221

Marcq, K., & Braeken, J. (2022). The blind side: Exploring item variance in PISA 2018 cognitive domains. *Assessment in Education: Principles, Policy & Practice*, *29*(3), 332–360. https://doi.org/10.1080/0969594X.2022.2097199

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2

Mislevy, R. J., & Wu, P.-K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, *1996*(2), i–36. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

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

OECD. (2020a). Item pool clasification. In *PISA 2018 technical report - PISA*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2020b). *PISA 2018 technical report - PISA*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2020c). The PISA target population, the PISA samples and the definition of schools. In *PISA 2018 results (volume II): Where all students can succeed*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Olsen, R. V., Turmo, A., & Lie, S. (2001). Learning about students' knowledge and thinking in science through large-scale quantitative studies. *European Journal of Psychology of Education*, *16*(3), 403–420. https://doi.org/10.1007/BF03173190

Rauch, D., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*(4), 354–379.

Regjeringen. (2021). Women's rights and gender equality. Retrieved August 30, 2023, from https://www.regjeringen.no/en/topics/foreign-affairs/the-un/innsikt/womens_rights/id439433/

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, *74*(4), 445–458. https://doi.org/https://doi.org/10.1037/amp0000356

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational*

Measurement, *40*(2), 163–184. https://doi.org/https://doi.org/10.1111/j.1745-3984.2003.tb01102.x

Ryan, J., & DeMark, S. (2002). Variation in achievement test scores related to gender, item format, and content area tests. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity technical adequacy, implementation* (1st ed., pp. 67–88). Lawrence Erlbaum Associates, Inc. https://doi.org/https://doi.org/10.4324/9781410605115

Schoultz, J., Säljö, R., & Wyndhamn, J. (2001). Conceptual knowledge in talk and text: What does it take to understand a science question? *Instructional Science*, *29*(3), 213–236. https://doi.org/10.1023/A:1017586614763

Siegfried, C., & Wuttke, E. (2019). Are multiple-choice items unfair? And if so, for whom? *Citizenship, Social and Economics Education*, *18*(3), 198–217. https://doi.org/10.1177/2047173419892525

Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education*, *12*(5), 563–573. https://doi.org/10.1080/0950069900120508

Team, R. C. (2020). R: A Language and environment for statistical computing [Computer software]. https://www.R-project.org/

Thorndike, R. L., & Angoff, W. H. (1971). *Educational measurement* (2nd ed.). American Council on Education.

Tulving, E. (1982). Synergistic ecphory in recall and recognition. *Canadian Journal of Psychology / Revue canadienne de psychologie*, *36*(2), 130–147. https://doi.org/10.1037/h0080641

Twist, L., & Fraillon, J. (2020). Assessment content development. In H. Wagemaker (Ed.), *Reliability and Validity of International Large-Scale Assessment : Understanding IEA's Comparative Studies of Student Achievement* (pp. 37–59). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_4

Uner, O., & Roediger, H., III. (2022). Do recall and recognition lead to different retrieval experiences? *American Journal of Psychology*, *135*(1), 33–44. https://doi.org/https://doi.org/10.5406/19398298.135.1.03

Utdanningsdirektoratet. (2023). Eksamen og prøver. Retrieved November 5, 2023, from https://www.udir.no/eksamen-og-prover/

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*(4), 369–386. https://doi.org/10.3102/10769986028004369

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

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

Appendix A

GDPR Documentation & Ethical Approval

Based on the information shared in the PISA 2018 Technical Report (OECD, 2020b), the data used in the present study can be regarded as anonymous. EU General data protection regulation (GDPR) and the Norsk Senter for Forskningsdata (NSD) states that anonymous data is not subject to notification.

- PISA 2018 technical report https://www.oecd.org/pisa/data/pisa2018technicalreport/
- GDPR not applicable for anonymous data
  https://gdpr.eu/Recital-26-Not-applicable-to-anonymous-data
- NSD not applicable for anonymous data
  https://www.gdprsummary.com/anonymization-and-gdpr/

Recitals

Recital 26
Not applicable to anonymous data

The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

Sikt

Which personal data will be processed?

What are personal data?

What is processing?

General categories of personal data

- [ ] Name
- [ ] 11-digit personal identifier or other national ID number
- [ ] Date of birth
- [ ] Contact information
- [ ] Online identifiers
- [ ] People in images or video recordings
- [ ] Voice on audio recordings
- [ ] Location data
- [ ] Background information that, when combined, can be used to identify an individual
- [ ] Other personal information

Special categories of personal data

- [ ] Health data
- [ ] Ethnicity
- [ ] Political beliefs
- [ ] Religious beliefs
- [ ] Philosophical beliefs
- [ ] Sex life
- [ ] Trade Union Membership
- [ ] Genetic data
- [ ] Biometric data
- [ ] Criminal offences

> **If you will only be processing anonymous data you should not notify your project**
> Anonymous data are data where individual persons are not/no longer identifiable; not directly, indirectly or via email/IP address or scrambling key.

Continue to login

Appendix B

Data Management and Analysis Code

Data are publicly available and retrievable at https://www.oecd.org/pisa/data/2018database/.

R syntax for data management steps and analyses related to this master thesis can be found via the link https://drive.google.com/drive/folders/1M15NGrajC0V6XXBvSEWHX6zwxuApWht4?usp=drive_link or in the repository https://github.com/EJassulyCDN/thesisformat. The following parts can be found:

- Data management, descriptives and models: OverallCode.R
- Sensitivity analysis for missing data, partial credit and students nested in schools: Sensitivity.R

OverallCode.R should be run before the Sensitivity.R

Appendix C

Supplemental material

The robustness of the results when taking other approaches to missing data, partial credit response handling and students nested in schools were tested. The results were fairly robust after coding 'no response' as NA and taking different partial credit handling approaches (i.e. partial credit=NA, partial credit=correct). The significance pattern was the same and the difference in magnitude of the regression weights did not surpass an absolute value of 0.18 in the logit scale. The variance component of the school was .01%. These results can be found in the already listed Sensitivity.R file.