Null Baseline Modeling Approaches with Applications in International Large-Scale Educational Assessments

Saskia van Laar

Dissertation submitted for the degree of PhD

CEMO: Centre for Educational Measurement at the University of Oslo,

Faculty of Educational Sciences, University of Oslo

2023

Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 360

ISSN 1501-8962

Cover: UiO.
Print production: Graphic Center, University of Oslo.

Acknowledgements

December 2022, Oslo

When I started my PhD, my main goal was just to learn as many new things as possible. Now five years later, I do believe that I have been able to broaden my knowledge and skills, with this thesis being the final accomplishment. However, I could not have done this without the help of others.

I want to express my deepest gratitude to my supervisor Prof.Dr. Johan Braeken. Thank you for sharing your wisdom, for challenging me to do better, and for your continuous help and support until the end of the project. You were there to help me overcome difficulties during the process and were able to put me back on track when I got lost. Honestly, I do not know how I could have done it without your guidance. On a side note, I would also like to thank you for your efforts to boost productivity by sharing some of the best (and worst) music I have ever heard.

I am also grateful to Jianan Chen for her help with the project. Without your valuable contributions, we would not have been able to finish as many of our ideas as we did. And of course a big thank you to my (former) colleagues at CEMO, friends, and family — both faraway and nearby— for their encouragement and support throughout it all, and for enriching this time in Norway. I want to extend a final acknowledgment to the Norwegian Research Council, without their support this project would not have been possible.

Summary

This thesis explores two specific applications of null baseline model comparisons in the context of quantitative research in international large-scale educational assessments. Both applications make use of a null baseline model in which all observed variables are assumed to be uncorrelated.

Articles 1 and 2 focus on model fit evaluation with the Comparative Fit Index (CFI). Here, the fit of a model of interest is compared against the fit of the null baseline model. Two simulation studies clarified the meaning and behavior of the CFI, as well as the consequences for the commonly used rule-of-thumb for model fit evaluation, as a function of the null baseline model. Both articles end with the general reminder that incremental fit indices are relative measures with the null baseline model as a standardized metric, and thus their values should not be compared in an absolute sense nor should universal rules-of-thumb be adopted.

Articles 3-6 focus on random response behavior in the TIMSS 2015 student questionnaire. To assess the prevalence of and to identify those students likely engaging in random response behavior, we adopted a mixture item response theory (IRT) approach. In this approach, a relative comparison definition for random responders is used based on a contrast between a measurement model (reflecting regular response behavior) and a uniform null baseline model (reflecting random response behavior). The articles investigated the prevalence, impact, and characteristics of random responders, as well as where and how often random response behavior occurs. Results showed that: (i) The average prevalence of random responders was estimated at 7.5% [0-38%]; (ii) The overall impact of random responders on aggregated-level was fairly limited; (iii) Scale position and scale character were important determinants for the prevalence of random responders; (iv) Certain groups of students (e.g., students in higher grades or male students) were more likely to be identified as random responders; and (v) Random responding is not necessarily a consistent behavior across the questionnaire.

This thesis was supported by a research grant [FRIPRO-HUMSAM261769] of the Norwegian Research Council and has been carried out at CEMO: Centre for Educational Measurement at the University of Oslo.

Sammendrag

Denne avhandlingen ser på to spesifikke anvendelser av nullmodell-sammenligninger. Sammenligningene utforskes ved å benytte data fra internasjonale storskalaundersøkelser innenfor utdanningsforskning. Begge anvendelsene bruker en nullmodell der ingen av de observerte variablene antas å korrelere.

Artikkel 1 og 2 fokuserer på evaluering av modelltilpasninger ved bruk av Comparative Fit Index (CFI). Her blir evaluering av en valgt modells tilpasning sammenlignet med tilpasningen for nullmodellen. To simuleringsstudier bidro til å få klarhet i betydningen og funksjonen til CFI i tillegg til å avdekke konsekvensene av å bruke den vanlige tommelfingerregelen for vurderinger av modelltilpasninger som en funksjon av nullmodellen. Begge artiklene avsluttes med en generell påminnelse om at trinnvise tilpasningsmål er relative mål når nullmodellen brukes som en standardisert enhet, som videre fører til at CFI-verdiene ikke kan sammenlignes i absolutt forstand. I tillegg advares det mot å benytte universelle tommelfingerregler.

Artiklene 3-6 fokuserer på tilfeldig svaratferd i elevspørreskjemaet til TIMSS 2015. For å måle utbredelsen av tilfeldig svaratferd og for å identifisere elever med slik atferd, brukte vi en tilpasset «mixture item response theory (IRT)» tilnærming. I denne tilnærmingen benyttes en relativ definisjon for tilfeldig svaratferd. Definisjonen tar utgangspunkt i kontrasten mellom en målingsmodell som gjenspeiler vanlig svaratferd og en enhetlig nullmodell som gjenspeiler tilfeldig svaratferd. Artiklene undersøkte utbredelsen, virkningen og egenskapene til elever med tilfeldig svaratferd, samt hvor og hvor ofte tilfeldig svaratferd forekommer. Resultatene viste at: (i) Gjennomsnittlig andel elever med tilfeldig svaratferd ble estimert til 7,5% [0–38%]; (ii) På et aggregert nivå er den generelle virkningen av tilfeldig svaratferd ganske begrenset; (iii) Skalaposisjon og -karakter var viktige faktorer i forekomsten av elever med tilfeldig svaratferd; (iv) Enkelte elevgrupper (f.eks. elever i høyere klassetrinn eller gutter) hadde høyere sannsynlighet for å bli identifisert som tilfeldig svarende; og (v) Å svare tilfeldig er ikke nødvendigvis en konsekvent handling gjennom hele spørreskjemaet.

Denne avhandlingen ble finansiert av forskningsmidler fra Norges forskningsråd [FRIPRO-HUMSAM261769] og har blitt utført ved CEMO: Centre for Educational Measurement at the University of Oslo.

Null Baseline Modeling Approaches with Applications in International Large-Scale Educational Assessments

> A man from Mars, asked whether or not your suit fits you, would have trouble answering. He could notice the discrepancies between its measurements and yours, and might answer no; he could notice that you did not trip over it, and might answer yes. But give him two suits and ask him which fits you better, and his task starts to make sense, though it still has its difficulties. (Edwards, 1965)[1]

With this example Edwards (1965)[1] illustrated that in order to draw proper conclusions one needs a point of reference to compare against. The same principle also holds within a measurement context. Consider for example a student who chooses the correct response option for 5 out of 10 yes/no test items. Is that a good or bad performance? In an absolute sense, the student got half of the items correct, yet if you realize that someone who randomly guesses is also expected to get half of the items correct, the student's performance is less than impressive.

This latter example shows that the interpretation of any measured quantity in an absolute sense is quite hard. Yet it also shows that providing a point of reference, to which a measured value can be compared, creates added value as it can help to make a better evaluation of the quality of the measured quantity. This point of reference is an essential condition for comparison and is key to meaningful interpretation of the measured values in a relative sense (Raivola, 1985)[2]. In this process of comparing measured values against this point of reference or baseline, further meaning can be gained. Some would even go as far as stating that something only has value and meaning in comparison to something else (e.g., Royall, 1997)[3].

[1] Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, *63*, 400–402.

[2] Raivola, R. (1985). What is comparison? Methodological and philosophical considerations. *Comparative Education Review*, *29*(3), 362–374.

[3] Royall, R. (1997). *Statistical evidence: A likelihood paradigm.* Chapman & Hall.

These comparative principles come naturally to most people, yet they also prove to be a useful tool in other more complex settings in research. This thesis explores two specific instances where comparison plays a crucial role. The first instance relates to the evaluation of the fit of statistical models through *model comparison*. The second instance relates to *model building* for evaluating response behavior of students on self-report questionnaires in international large-scale educational assessments. Overall, this thesis aims to examine and clarify how null baseline models can be effectively used as informative comparison grounds in research applications.

Outline of the Thesis

This thesis is article-based, meaning the basis consists of several journal articles (see overview below), supplemented with an extended abstract or so-called 'kappe'. The extended abstract, which started with the current introduction, provides room for a more general introduction to and reflections upon different aspects of the articles and the overarching theme.

The thesis is structured in such a way that it begins with the extended abstract and ends with the articles (i.e., Chapters 4-9). However, it is recommended to first read the articles and then the extended abstract. The extended abstract itself has begun with the more general part that provides some overarching reflections on the general theme of null baseline model comparison and ends in Chapter 3 with further reflections on the comparative use of models. In Chapter 1 and Chapter 2, the two different null baseline model comparison applications are discussed in more detail. As the two applications come from different perspectives, Table 1 provides a brief preview of what will be their main differences and common ground.

Application 1

The first application focuses on the meaning of null baseline model comparisons using the Comparative Fit Index (CFI) in the evaluation of structural equation models which are widely used in the social sciences. To this date, the evaluation of model fit remains a crucial, yet difficult topic. There have been many cautious notes and examples in the literature that show that fit indices and their rules-of-thumb —for determining whether

or not the fit of a model of interest is acceptable— do not always work as intended, as they have been shown to be sensitive to different data and model characteristics. Yet regardless of the warnings fit indices and their rules-of-thumb are still being used in a rigid and binary fashion. With this application, we specifically tried to re-establish what CFI stands for. Different simulation studies were included to provide a better understanding of the behavior and performance of CFI as a function of the null baseline and to clarify why current practice is not ideal.

Table 1

Overview of the Main Differences and Common Ground for the Two Applications of Model Comparison in this Thesis.

	Common Ground	
	Model-based approach in a measurement context	
	Comparison to provide meaning using a *Null Model* as baseline	
	Assessment of Fit	
	Complementary Perspectives	
	Application 1: Model Fit Evaluation with CFI	Application 2: Random Responders in TIMSS
Object of Comparison	Model-centered	Group-centered
Goal of Comparison	Justification	Classification
Data-Model Fit	Variable-based	Person-based

Application 2

The second application addresses the issue of random responders in in the Trends in International Mathematics and Science Study (TIMSS) 2015 assessment throughout different empirical studies. While the results of international large-scale educational assessments are widely used for research and educational policy, their low-stakes character (i.e., no direct consequences for the participating students) makes them vulnerable to invalid response behavior. Depending on the severity of this behavior, this can lead

to problems with the interpretation of the assessment results. The studies within this application are built around the development of a mixture model, incorporating a uniform null baseline model and a measurement model, to distinguish between individuals who are genuinely answering to the questionnaire and individuals who are choosing responses randomly throughout as if they disregard what is asked from them. This methodology can be especially useful in the context of measurement validity and data quality.

Contributions

By showing how null baseline models are effectively used in practice, for model fit evaluation in structural equation modeling and for the detection of random responders, this thesis contributes to the provision of methodological understanding and tools for better research practice. In particular, by reminding quantitative researchers that make use of fit indices what the indices actually measure, this work hopefully provides a starting point for more deliberate decision-making when evaluating models in practice. At the same time, the thesis hopefully increases awareness among testing organizations and educational researchers about the need for data quality monitoring for the survey part of international large-scale educational assessments.

Thesis Articles[4]

(1) van Laar, S., & Braeken, J. (2021). Understanding the comparative fit index: It's all about the base! *Practical Assessment, Research, and Evaluation*, *26*, Article 26. https://doi.org/10.7275/23663996

(2) van Laar, S., & Braeken, J. (2022a). Caught of base: A note on the interpretation of incremental fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(6), 935–943. https://doi.org/10.1080/10705511.2022.2050730.

(3) van Laar, S., & Braeken, J. (2022b). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501. https://doi.org/10.1111/jedm.12317

(4) van Laar, S., & Braeken, J. (2022c). *Prevalence of Random Responders as a function of Scale Position and Questionnaire Length in the TIMSS 2015 eighth-grade Student Questionnaire*. Manuscript under review.

(5) Chen, J., van Laar, S., & Braeken, J. (2022). *Who are those Random Responders on your Survey? The case of the TIMSS 2015 student questionnaire*. Manuscript under review.

(6) van Laar, S., Chen, J., & Braeken, J. (2022). *How Randomly are Students Random Responding to your Questionnaire? Within-Person Variability in Random Responding across Scales in the TIMSS 2015 eighth-grade Student Questionnaire*. Manuscript under review.

[4]Braeken, J. (Johan) is my PhD supervisor. Chen, J. (Jianan) is a master's student in MSc Assessment, Measurement, and Evaluation at the University of Oslo, who we hired as a research assistant on the project towards the end of the PhD.

Table of Contents

1 Application 1:
Model Fit Evaluation with the Comparative Fit Index

1.1 Background

An important topic within the context of Structural Equation Modeling (SEM) is model fit. Model fit refers to the ability of a model to reproduce (the observed relations between the variables in) the data. In practice, the data (i.e., the responses given by N individuals on p variables) is summarized in a sample-observed covariance matrix of size $p \times p$ (e.g., Bentler & Bonett, 1980). The model describes the expected relations between the variables and the outcome from estimating the model based on the data is a model-implied covariance matrix of similar size (i.e., $p \times p$). In the model evaluation process, this model-implied covariance matrix will be compared against the sample-observed covariance matrix and fit indices are used to summarize the discrepancy between them by some quantity.

1.1.1 Fit Indices: Quantifying and Summarizing Model Fit

The use of fit indices as a way to assess model fit can be traced back to the χ^2 test of exact fit. Comparing the sample-observed and model-implied covariance matrices, the specific null hypothesis being tested by this χ^2 test is one of no difference between both covariance matrices. Assuming that the model is correct and the assumptions are met, the χ^2-statistic is asymptotically distributed as a central χ^2-distribution based on the degrees of freedom of the model. This distribution can be used to determine the p-value for testing the null hypothesis with respect to the observed χ^2-value. In general, as the difference between the sample-observed and model-implied covariance matrices becomes larger, the χ^2-value will increase and will be less compatible with the tested hypothesis.

Unfortunately, the χ^2 test is also familiar with some problems related to significance testing. A review by Ropovik (2015) showed that, in practice, the most common reason for not reporting the χ^2 test statistic is its sensitivity to sample size. It is too liberal for large sample sizes, rejecting the null hypothesis of equal fit too easily. In addition, the null hypothesis of no difference might also be unreasonable, especially when the goal

is to find and accept a working model. In general, too much focus on statistical testing can also lead to disregarding or changing relevant and theoretical sound models without proper justification for it (Bentler & Bonett, 1980).

Nowadays, a lot of alternative fit indices are available to supplement the narrow perspective of null hypothesis testing provided by the χ^2 test of exact fit. In general, three broad classes of fit indices can be distinguished: absolute fit, parsimony fit, and incremental fit indices. In short, absolute fit indices use the data as a point of reference against which the fit of a model is compared, where fit is determined by the degree of discrepancy between the model en data (i.e., either sample-observed and model-implied correlation or covariance matrices). Lower values for the absolute fit indices indicate smaller discrepancies and thus better fit. The parsimony fit indices are characterized by incorporating penalties for model complexity (e.g., more parameters) when assessing model fit. The incremental fit indices compare the fit of a model against the fit of a more restricted baseline model. In general, higher values for the incremental fit indices indicate better fit (e.g., Brown, 2015; Kline, 2016).

1.1.2 Qualitative Model Evaluation

In practice, these alternative fit indices are often used for qualitative evaluation of a model in terms of good or bad model fit. With the main question being what rule to abide by to indicate good or bad fit? (Brown, 2015). For that, researchers often turn back to the different rules-of-thumb that have been proposed over time for determining whether the fit of the model is indeed acceptable. Nowadays, the most commonly used rules-of-thumb seem to originate from the simulation study by Hu and Bentler (1999), who evaluated the performance of various different fit indices.

In practice, problems arise as these rules-of-thumb seem to be universally applied, even though literature has shown the sensitivity of fit indices and their rules-of-thumb to different data and model characteristics (for a review see e.g., Niemand & Mai, 2018). As a consequence, the rules-of-thumb do not always lead to a correct evaluation of the model. This should not be unexpected, as it resonates well with warnings from Hu and Bentler (1999) with respect to the generalizability of their results beyond the conditions

studied and warnings against using single criteria and rigid rules. Yet, one potential explanation for the continued use of these rules-of-thumb across any and all situations is that "researchers need them because it is unclear how one can reach qualitative judgements in their absence" (Lai & Green, 2016, p.211). In addition, there have been concerns that the use and meaning of fit indices are not well understood in general (e.g., McDonald & Ho, 2002). Hence, it is not illogical to observe oversimplified rule-based behavior. Without good understanding of what the fit indices actually stand for it is difficult to put results into perspective and the most straightforward approach is to rely on fixed rules and procedures and justify choices by referring to some authority. If anything, this also suggests that it might be important to reassess the nature of the different fit indices and their role in model evaluation.

The general aim of the first two articles was to get a better understanding of fit indices. Yet given the number of fit indices available (i.e., Marsh et al. already evaluated 29 fit indices in 1988 and the number of fit indices still continues to increase), it is not feasible to address them all at once. Therefore, we focused on the class of incremental (also comparative or relative) fit indices, which is also the class most in line with the comparison principles in the introduction. The CFI, in particular, has been chosen because it is one of the most used indices in practice (e.g., Jackson et al., 2009; McDonald & Ho, 2002; Ropovik, 2015).

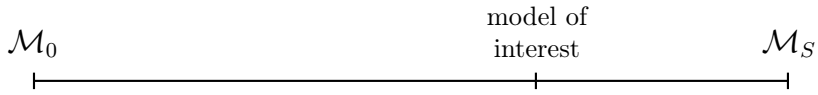
1.1.3 Rationale behind Incremental Fit Indices

Central to model fit evaluation with incremental fit indices is the comparison of a model of interest to a more restricted baseline model. In practice, a null baseline model in which all observed variables are assumed to be uncorrelated has taken off as the default baseline model. The incremental fit indices are also considered practical measures as they give an indication of the improvement in fit of a model of interest over a more restrictive baseline model and they provide information about the value of the model of interest in explaining the data (Bentler & Bonett, 1980). The relation between the models involved in the comparison can be depicted by means of a continuum of models, from the worst-fitting baseline model ($\mathcal{M}_0$) to the perfect-fitting or saturated model ($\mathcal{M}_S$). The role of

the incremental fit indices is to assess where the model of interest is located within this continuum (see also Figure 1). It follows that the closer the model of interest is located towards the baseline model, the smaller the improvement in fit (Bentler, 1990).

Figure 1

Continuum of Models and Incremental Fit.



Note. Considering a continuum of models, the improvement in fit of a model of interest over a more restrictive baseline is dependent on where the model of interest is located within the continuum.

1.1.4 The Comparative Fit Index

The Comparative Fit Index (CFI) is an index that describes the proportional improvement in fit of a model of interest compared to a baseline model by "[summarizing] the relative reduction in noncentrality parameter of two nested models" (Bentler, 1990, p.238). The noncentrality parameter λ_m of a model m can be seen as an indicator of model misspecification as it quantifies the discrepancy between the estimated fit for the model (i.e., χ^2_m value) and the expected value for the sample if the model is correctly specified (i.e., model's degree of freedom df_m). The value of CFI is based on the ratio of misspecification of both models and the sample estimator is given by:

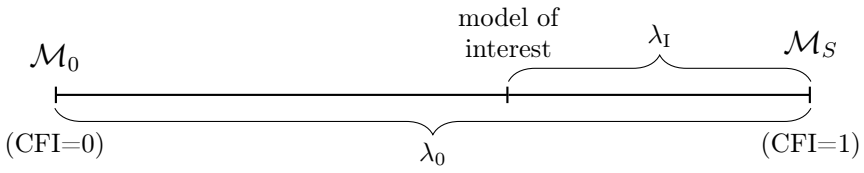
$$\text{CFI}_{(\text{I},0)} = 1 - \frac{\lambda_\text{I}}{\lambda_0} = 1 - \frac{\chi^2_\text{I} - \text{df}_\text{I}}{\chi^2_0 - \text{df}_0}$$

where the subscript indicates whether the statistics are of the model of interest I or the null baseline model 0. The continuum of CFI (see Figure 2) is usually normed to reflect a [0,1] interval[5], with higher values indicating larger improvement in fit, or higher correspondence between the model of interest and the data over the null baseline model. A CFI of .90 implies that the fit of the model of interest is 90% better than that of the baseline model.

[5]Normed value for $\text{CFI}_{(\text{I},0)} = 1 - \frac{\max(\lambda_\text{I},0)}{\max(\lambda_0,\lambda_\text{I},0)}$

Figure 2

Graphical Representation of the Comparative Fit Index.



Note. $\mathcal{M}_0$ = null baseline model; λ_0 = noncentrality parameter of the null baseline model; $\mathcal{M}_S$ = saturated model; λ_I = noncentrality parameter of the model of interest.

1.1.5 Origin of Rules-of-Thumb for Incremental Fit Indices

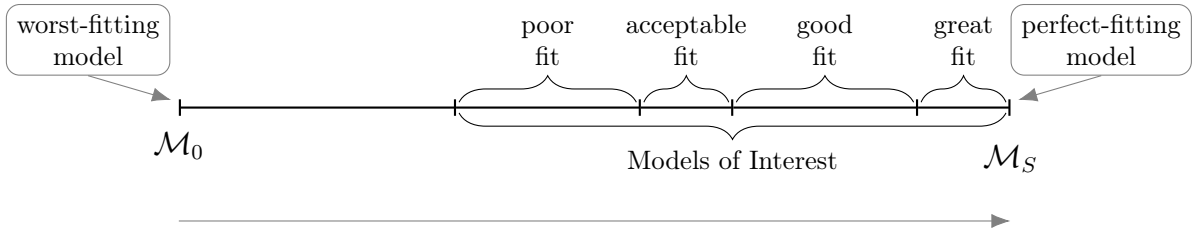
As mentioned before, the incremental fit indices are supposed to give an indication of the improvement in fit of a model of interest over a more restrictive baseline model. Yet early on, it had already been acknowledged that it might be hard to interpret the scale of incremental fit indices and that people probably need more guidance in using them. Therefore, Bentler and Bonett (1980) tried to attribute some practical meaning to the value of fit indices. They stated that "In our experience, models with overall fit indices of less than .90 can usually be improved substantially" (Bentler & Bonett, 1980, p.600). Not surprisingly, using values of at least .90 for indicating acceptable model fit seem to be found as the initial rule-of-thumb used in practice (see the review of e.g., McDonald & Ho, 2002). Yet the interesting part here is that, while McDonald and Ho (2002) found that in the late 90's most researchers used the .90 rule-of-thumb for incremental fit indices like CFI, CFI itself didn't even exist when this rule-of-thumb came to be in 1980. Over time, the performance of this and other rules-of-thumb has been questioned and more formal investigations of their adequacy followed. With that respect, Hu and Bentler's study (1999; see also Hu and Bentler, 1998) has become most influential for research practice and caused a transition in the model evaluation process. Based on their results, people have taken .95 as the new core threshold with CFI values $\geq$ .95 indicating good model fit, which, to this day, remains the most prevalent rule-of-thumb used in practice.

1.1.6 Room for Improvement vs. Is it Good Enough?

To me, the current use of fit indices seems to indicate a more important shift in interpretation with respect to the perception of the guideline provided by Bentler and Bonett (1980). Where Bentler and Bonett (1980) explicitly mention that certain models can be improved upon (see quote above), we see in practice that their guideline is perceived as a way to determine which models are good enough. More specifically, it has led to the concept that if the value of a fit index conforms to some rule-of-thumb then the model should be deemed acceptable (in this specific case values of at least .90). Thus, dividing the continuum of models we saw earlier in different sections and depending on the section in which the model of interest is located, a certain qualitative value is assigned to the model of interest (for an illustrative example, see Figure 3). This qualitative interpretation seems to be deeply rooted in practice. According to a review by Ropovik (2015) the most important question with respect to model evaluation is whether the fit of the model is good enough for further analysis and interpretation (Ropovik, 2015), like a box that needs to be checked without considering what the value for the incremental fit indices actually stands for.

Figure 3

(Qualitative) Model Evaluation with Incremental Fit Indices.



Note. Considering a continuum of models, the qualitative value assigned to a model is dependent on the section in which the model of interest is located. The figure is adapted from *Longitudinal structural equation modeling* by T. D. Little, 2013, *The Guilford Press*.

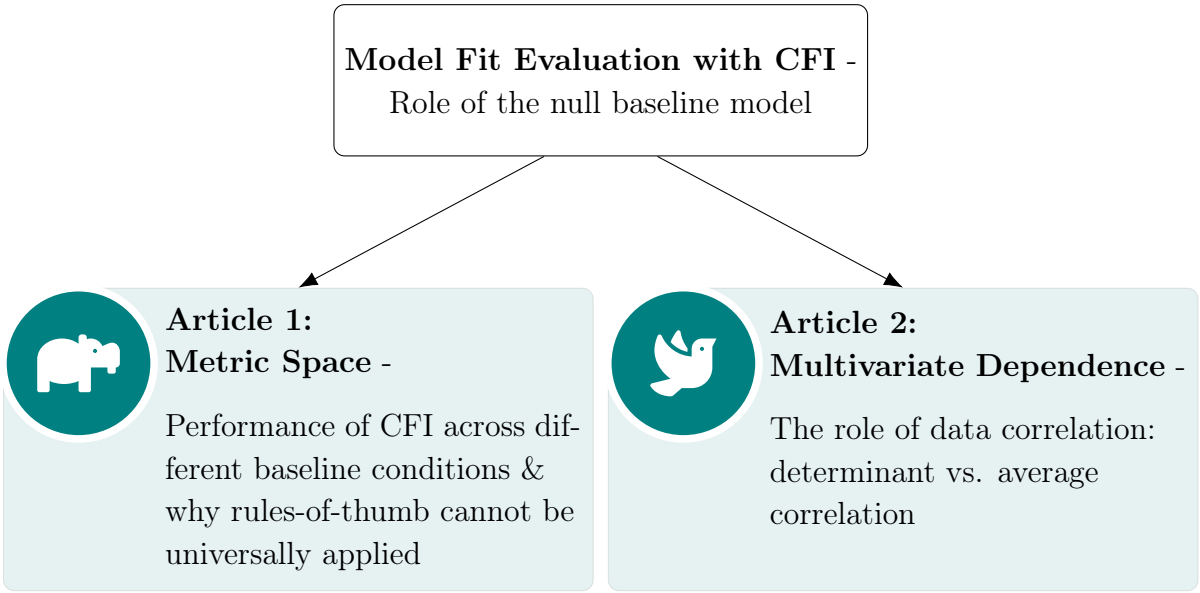
1.2 Conceptual Framework of Articles

As mentioned before, the aim of the first two articles was to get a better understanding of the CFI (e.g., what the index stands for, how we should use it, and why it behaves

as it does). In these articles, we tried to clarify the meaning and behavior of the CFI, as well as the consequences for the commonly used rule-of-thumb, as a function of the null baseline. The importance of the baseline had already been implied by the study of Marsh et al. (1988). Their study showed that the incomparability in the performance of fit indices was influenced by the fit of the null baseline. While both articles in this thesis include an explicit decomposition of the null baseline model as a way to get more insight into the specific components that are responsible for influencing the behavior of CFI, their main focus is slightly different (see Figure 4). 'Article 1: Metric Space' for example, discusses the so-called metric space principle (or baseline principle) which explains why the rule-of-thumb for CFI cannot work across any and all situations and in addition shows how/to what degree this principle, as a function of the components of the null baseline model, influences the behavior of CFI across different conditions. 'Article 2: Multivariate Dependence' on the other hand, concentrates on a specific element of the baseline and uses the alleged impact of model type as a way to discuss how it is the ability of a model to capture the most dominant correlation in the data, instead of the average correlation, that is crucial when considering the performance of CFI. The overall conclusion is that exactly because incremental fit indices are relative measures, they should be treated as such: Values should not be compared in an absolute sense across any and all situations, nor should universal rules be adopted.

Figure 4

Overview of the Different Articles about Model Fit Evaluation with CFI.



1.2.1 Metric Space Principle

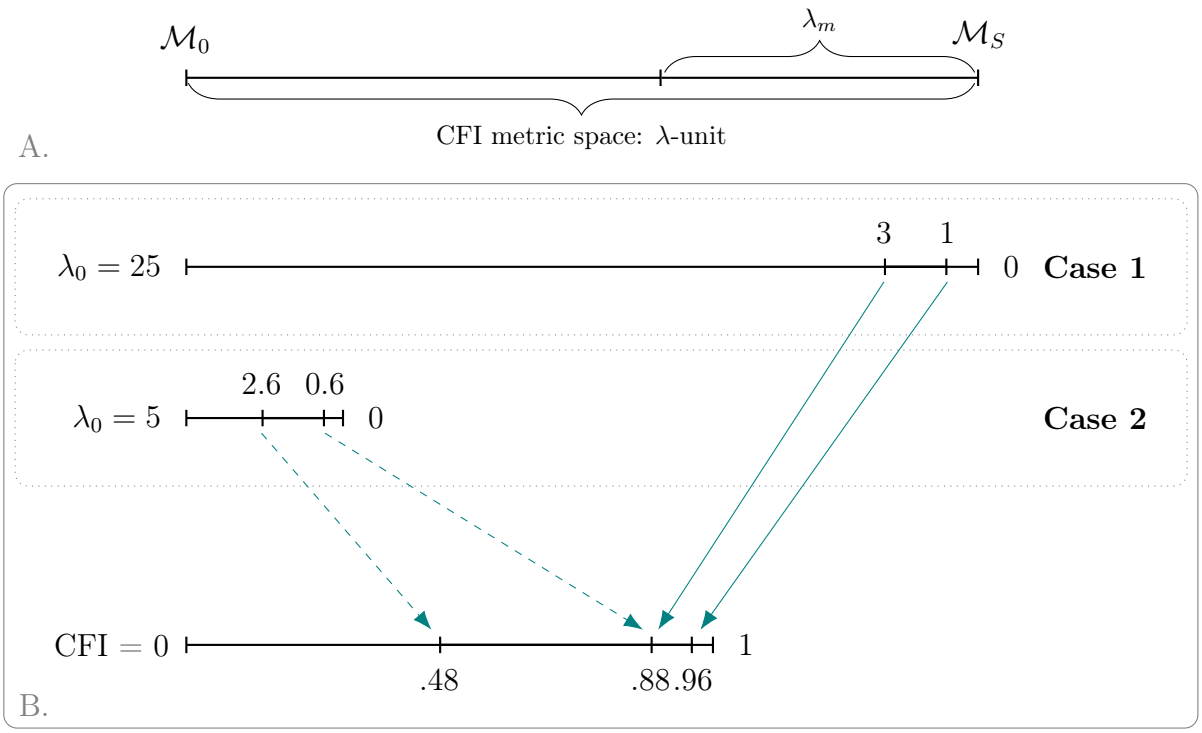
The value for CFI depends on both the fit of a model of interest as well as the fit of the null baseline model. Given the specific formulation, it follows that the fit of the null baseline serves as the standard for comparison. In 'Article 1: Metric Space', we refer to this standard as the 'CFI metric space', which can be visualized as a one-dimensional line with noncentrality as a unit (see also Figure 5: Panel A). The endpoints of the metric space are set by the null- and saturated model. Yet, the length of the metric space is determined by the noncentrality of the null model, with the noncentrality for the saturated model always being equal to zero.

The main idea here is that CFI and its rule-of-thumb become less useful when the metric space is shorter. The shorter the metric, the more similar all models are in the model comparison, making it harder to differentiate between the model(s) of interest, null model, and saturated model. Panel B in Figure 5 provides a visual representation to clarify the metric space principle (see also 'Article 1: Metric Space'). The figure shows two cases where the size of the metric space is different. Comparing the two cases, the metric space for case 1 could be considered relatively large with $\lambda_0 = 25$. Within this

space, we have two models with slightly different noncentrality values (i.e., $\lambda_1 = 1$ and $\lambda_2 = 3$). Translating these values to the CFI interval results in values of $\text{CFI}_{(1,0)} = .96$ and $\text{CFI}_{(2,0)} = .88$. In the second case where with $\lambda_0 = 5$ the metric space is much shorter, the two models are also only two noncentrality units apart (i.e., $\lambda_1 = .6$ and $\lambda_2 = 2.6$). However, translating this to the CFI interval results in values of $\text{CFI}_{(1,0)} = .88$ and $\text{CFI}_{(2,0)} = .48$.

Figure 5

Baseline Comparison for CFI.



Note. Illustration of the metric space principle. Where $\mathcal{M}_0$ = null baseline model; λ_0 = noncentrality parameter of the null baseline model; $\mathcal{M}_S$ = saturated model; λ_I = noncentrality parameter of the model of interest.

The first thing to notice is that although the difference between the models in terms of absolute misspecification is equal in both cases, the difference in the corresponding CFI-values for the two models is much lower in the case where we have the smaller metric space. This also implies that a reduction in the size of the metric space has a negative effect on the ability of CFI to differentiate between models. In situations similar to the second case, where small differences in noncentrality lead to much bigger differences in

CFI-values, interpretation of these values based on a fixed rule-of-thumb can become especially problematic as it could lead to very different conclusions with respect to model fit evaluation.

In addition, one could wonder whether these differences are meaningful in practice. For example, consider that in the second case the null baseline model already shows adequate fit. In such a situation, it might not only be more difficult for a model to do better, but more importantly, regardless of the CFI value, either being .48 or .88 it should reflect this adequate fit. Yet, interpretation based on rules-of-thumb does not take this 'base' fit into consideration.

Thus, what this hopefully shows is that the meaning we assign to CFI should be dependent on context and not some fixed rule-of-thumb. At the same time, this also implies that values for CFI cannot directly be interpreted and compared if we don't know what baseline we are dealing with. Look for example at the situation where both case 1 and case 2 contain a model of interest with an equal CFI-value (i.e., CFI = .88). Relatively speaking the models of interest show equal improvement in fit over the null baseline model. Yet, the misspecification as measured by the noncentrality parameter is lower in the second case than in the first. Thus in case 2, where the metric space is relatively small, one might say that there is hardly room for more improvement, while in case 1 some progress can potentially still be made. The main point here is that one CFI-value is not the other as they are not based on the same proportion. Thus, for fair (qualitative) interpretation of CFI we need to acknowledge that the CFI metric space has an influence on its behavior/performance and thus, we should take this baseline into consideration when using this fit index in the model evaluation process.

1.2.2 Key Components of the CFI Metric Space

As the CFI metric space serves as a standard for comparison, it is also important to know which specific components play a role here to further increase understanding of the behavior of CFI. Based on characteristics of the null baseline model, decomposition of this baseline for CFI can be reduced to $\lambda_0 = -\log|\mathbf{R}|(n-1) - p(p-1)/2$ (see Appendix A in 'Article 1: Metric Space' or 'Article 2: Multivariate Dependence') and thus the

three key components being: (i) the number of items p, (ii) sample size n, and (iii) the amount of correlation in the data as part of $-\log|\mathbf{R}|$ (with $|\mathbf{R}|$ being the determinant of the observed correlation matrix). For specific predictions regarding the influence of the different components on the behavior of CFI in practice, see 'Article 1: Metric Space'.

1.2.3 Degree of Multivariate Dependence vs. Average Correlation

While 'Article 1: Metric Space' showed that the data correlation is the most important factor with respect to the size of the metric space, 'Article 2: Multivariate Dependence' provided a more detailed evaluation of the role of the data correlation in the performance of CFI. More specifically, it focused on how the impact of this data correlation is defined. For example, at some point I was taught that the value for CFI depends on the average size of the pairwise correlations in the data, with the idea that the value for CFI will not be that high, if this average correlation is not that high. Similarly, it has been brought up that one could manipulate the value for incremental fit indices by artificially changing the average correlation (e.g., Rigdon, 1998a) While there is some value in these statements, as higher average correlations do help, the role of the correlation in the null baseline is actually represented by the determinant $|\mathbf{R}|$, which does not simply reflect the average correlation in the data. Yet, in some way it is understandable that people tend to talk about the average, because it is not necessarily clear how a determinant changes as a function of a single pairwise correlation (see also Lai & Green, 2016), while people probably have a better intuition with respect to the average.

Figure 6

Determinant vs. Average Correlation

$$\mathbf{R}_1 = \begin{bmatrix} 1 & .4 & .4 & .4 \\ & 1 & .4 & .4 \\ & & 1 & .4 \\ & & & 1 \end{bmatrix} \quad \begin{aligned} \bar{r}_1 &= 0.40 \\ |\mathbf{R}_1| &= 0.48 \\ -\log|\mathbf{R}_1| &= 0.74 \end{aligned}$$

$$\mathbf{R}_2 = \begin{bmatrix} 1 & .3 & .9 & .3 \\ & 1 & .3 & .3 \\ & & 1 & .3 \\ & & & 1 \end{bmatrix} \quad \begin{aligned} \bar{r}_2 &= 0.40 \\ |\mathbf{R}_2| &= 0.15 \\ -\log|\mathbf{R}_2| &= 1.91 \end{aligned}$$

Note. $\bar{r}$ = average correlation; $|\mathbf{R}|$ = determinant of the correlation matrix.

For example, take a look at Figure 6. Both correlation matrices do have an equal average correlation of .40, yet the determinant differs across the two situations. In matrix $\mathbf{R}_1$ all pairwise correlations are equal to .40, while in matrix $\mathbf{R}_2$ they are all equal to .30 except for one pair, where the correlation is much higher. In this case, it is the second situation that is characterized by a lower determinant. For the determinant of a correlation matrix it follows that $|\mathbf{R}| = 1$ if all $r_{ij} = 0$, otherwise $|\mathbf{R}| < 1$. Yet, with respect to the correlation matrices, this example indicates that it is the most dominant correlation that is most important for the value of the determinant, instead of the average correlation. However, that does not mean that the most dominant correlation always refers to a single highest correlation-pair, like in this example. It captures something broader, like dominant correlation dimension(s) where you can have group(s) of items with high mutual correlations. It is important that a model is able to capture all those dimensions. 'Article 2: Multivariate Dependence' elaborates these points and stresses that CFI evaluates fit in terms of the determinant as opposed to the average pairwise correlation.

1.3 Highlights of Articles

Article 1: Metric Space

Despite the sensitivity of fit indices to various model and data characteristics in structural equation modeling, these fit indices are used in a rigid binary fashion as a mere rule-of-thumb threshold value in a search for model adequacy. Here, we address the behavior and interpretation of the popular Comparative Fit Index (CFI) by stressing that its metric for model assessment is the amount of misspecification in a baseline model and by further decomposition into its fundamental components: sample size, number of variables and the degree of multivariate dependence in the data. Simulation results show how these components influence the performance of CFI and its rule of thumb in practice. We discuss the usefulness of additional qualifications when applying the CFI rule of thumb and potential adjustments to its threshold value as a function of data characteristics. In conclusion, we at a minimum recommend a dual reporting strategy to provide the necessary context and base for meaningful interpretation and even more optimal, a

move to using CFI as a real incremental fit index intended to evaluate the relative effect size of cumulative theoretically motivated model restrictions in terms of % reduction in misspecification as measured by the baseline model.

Article 2: Multivariate Dependence

This note serves as a reminder that incremental fit indices are a form of standardized effect sizes and hence, all reservations with respect to interpretations of standardized effect sizes also transfer to their interpretation. Such a realization has major implications for the interpretation and use of incremental fit indices, for the theoretical (im)possibility of default universal rules of thumb in their application, and for simulation studies mapping incremental fit indices as if their value is comparable in an absolute sense across any and all conditions. A small but illustrative working example centered around the alleged impact of model type will drive these points home.

1.4 Method of Study

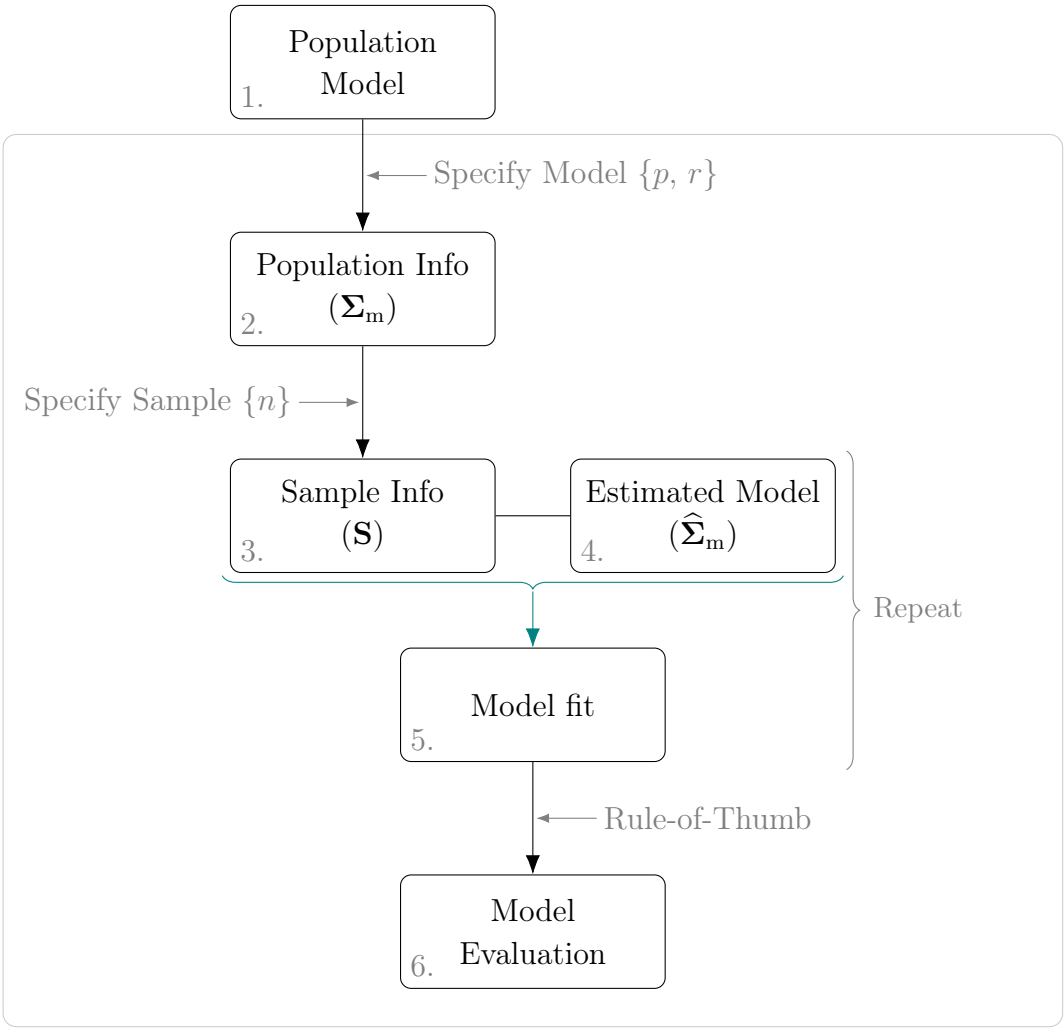
The common methodological feature of the two studies within Application 1 is that they are both simulation-based studies evaluating the performance of CFI and its rule-of-thumb across different conditions. Here, I will briefly introduce the advantage of adopting a simulation approach and discuss the concrete implementation used in the articles.

The first advantage of this simulation approach is that the 'truth' is known and this knowledge can be used in the evaluation process. Second, simulation studies are considered (empirical) experiments. They are set up in a well-controlled environment in which the 'truth' can be systematically varied to accommodate to different conditions. As a result, the effect of the systematic manipulations of potentially important design factors on performance can be easily observed without being dependent on existing data sets that meet certain criteria. Yet, while the flexibility of simulation studies does allow for the 'truth' to be manipulated in many different ways, the design of the simulation study should be in line with research objectives and should be kept feasible in practice and theoretically justified, instead of a mere random manipulation of anything and everything. Thus, given the goals for the first two articles, the experimental factors in the studies were chosen to reflect the key components of the CFI metric space: (i) the number of

items p, (ii) sample size n, and (iii) the amount of correlation in the data as part of the determinant of the observed correlation matrix $|\mathbf{R}|$.

Figure 7

Simulation Process Applied in the Context of Model Fit Evaluation.



Note. p = number of items; r = pairwise item correlation; n = sample size; $\boldsymbol{\Sigma}_{\text{m}}$ = model-implied population covariance matrix; $\mathbf{S}$ = sample-observed covariance matrix; $\widehat{\boldsymbol{\Sigma}}_{\text{m}}$ = model-implied sample covariance matrix.

1.4.1 Implementation of Simulation

Population-level. The general simulation process is reflected in Figure 7. The first step in the process was defining the overall structure of the true *population model* used for data-generation. In 'Article 1: Metric Space' for example, this was a one-factor model with p items and equal correlations of size r between all items. While the general

structure of the model is set, the data-generating process requires further specification of the population model to create a condition-specific population covariance matrix in step 2. In this part of the process, specification of the model is derived from manipulations to the number of items p, as well as $|\mathbf{R}|$ by changing to what degree the variables correlate. The model-implied population covariance matrix ($\boldsymbol{\Sigma}_{\text{m}}$) is then formed by combining the expected *model-implied population correlation matrix* and randomly generated variances from a uniform distribution (i.e., $s \sim \mathcal{U}(.75, 2)$).

Sample-level. Manipulations to sample size n are taken into consideration going into step 3, where a *sample-observed covariance matrix* ($\mathbf{S}$) is drawn from a Wishart distribution: $\mathbf{S} \sim \text{W}(\boldsymbol{\Sigma}, n-1)$, where $\boldsymbol{\Sigma}$ is the condition-specific population covariance matrix. Manipulation of sample size n will lead to the sample-observed covariance matrix being more or less variable with respect to the true population covariance matrix. This sample-observed covariance matrix provides a summary of the data that is sufficient for model estimation in the next step such that generating item-level data was not required. In step 4, a model is to be fitted to the sample-observed covariance matrix, resulting in the *model-implied sample covariance matrix* ($\widehat{\boldsymbol{\Sigma}}_{\text{m}}$). Note that the model to be fitted can be the true population model or an alternative misspecified model. Subsequently, the sample-observed and model-implied covariance matrix are used to determine fit according to CFI in step 5. Obviously, one sample does not tell the whole story as a different sample could potentially lead to (slightly) different results. Therefore, steps 3 to step 5 are repeated, resulting in a sampling distribution of CFI-values reflecting the variation in fit values of the estimated model for data of sample size n from the population model. Based on this sampling distribution, in step 6 the performance of the rule-of-thumb for CFI (i.e., CFI $\geq$ .95 indicating acceptable fit) can be evaluated. In the end, steps 2 to step 6 are to be repeated for each experimental condition separately.

1.4.2 Added Value

One of the reviewers from 'Article 1: Metric Space' asked if it would not be sufficient to consider the population version for CFI, describe how CFI depends on both fit of the null model as well as the model of interest and just stick with the conclusion that

correlation has an impact on the null baseline model, instead of doing a whole simulation study.

One important consideration here is that in practice people tend to work at sample-level and not population-level. The example in Figure 5 showed how slightly different fitting models can have a huge impact on the values that we get for CFI. In a similar fashion, sampling variability can be expected to have a bigger impact on the estimated CFI values in situations where the metric space is shorter. To what extent exactly? That is hard to quantify a priori based on mere intuition. This is exactly where the added value of a simulation study comes into play as it can provide more insight into the degree of variability in CFI related to sampling variability as a function of changes in the size of the null baseline. Surely, the behavior of CFI is based on more than just its baseline. Where the formulation of CFI clearly expresses the relations, the simulation can directly address the impact of bias and variance in fit for both the null baseline as well as the model of interest and how they together influence the performance of CFI. If we would want to put a number on the degree of variability, the results in 'Article 1: Metric Space' showed for example how in certain situations sample CFI values between .57 and 1.00 would be realistic values to expect for a true model. Yet, this large variation in CFI values also suggests that using (the common rule-of-thumb for) CFI might not be informative in those situations. In addition, 'Article 1: Metric Space' also showed indications of when the general metric space principle might not hold (e.g., low-sample-size-low-correlation situations dealing with more severe bias and sampling variation for the model of interest). If anything, this shows how the simulation study can provide a more nuanced view on the performance and use of CFI in practice.

1.4.3 Ethics & Good Scientific Practice[6]

The data used for the studies within Application 1 is simulation-based. A benefit of using this type of data is that it doesn't require ethical approval (Sigal & Chalmers, 2016) as it has no impact on humans, animals, and environment (Ören, 2000). Yet, in the broader context of good scientific practice one thing to consider is that it is important to

[6]Part of this section is based on a course paper written for 'UV9010: Research Ethics' at the faculty of Educational Sciences, University of Oslo.

pay attention to deliberate choices in the research process (e.g., NESH, 2016). Although simulation studies offer the flexibility to randomly manipulate anything and everything, one should be mindful that the research remains meaningful and useful. In practice, this means that good research design for simulation studies should focus on formulating questions that are relevant from a theoretical perspective and that within the design of the study there should always be a link to applied research (Paxton et al., 2001). As will also be discussed later on (see the *'Experimental Factors'*-section), the experimental conditions considered in 'Article 1: Metric Space' for example were inspired by general theory and situations encountered in practice. In addition, one should make sure that the results of simulation studies are not deceiving. For instance, one should not only focus on the specific conditions where a theory or method is supported but also pay attention to the conditions where a theory does not hold or where a method is not the preferred one. We for instance highlighted conditions in 'Article 1: Metric Space' where the metric space principle did not hold (i.e., when increasing the number of variables in low-sample-size-low-correlation conditions).

1.5 Design Considerations

1.5.1 Justification of Choices and Alternative Decisions

With any simulation study, one can wonder how realistic the configuration has been. Therefore some of the decisions made in the different steps of the simulation process for the studies in 'Article 1: Metric Space' and 'Article 2: Multivariate Dependence' are highlighted here.

Population Model. Clearly, the data-generating population model used in 'Article 1: Metric Space' was quite simple, yet we decided against including more complex models. First, changing the used population model, for example by changing the factor structure or releasing the constraint on the correlation patterns, would not change the underlying metric space principle. It would however make it less transparent how the different components are moving. Specifically the expected behavior of $|\mathbf{R}|$ would become less clear as the size of the correlations as well as the number of variables is manipulated. With the focus being on understanding the role of the null baseline in model fit evaluation

with CFI, the structure was kept simple and constant to keep things straightforward. In Application 2 we will also see that the assumption of unidimensional constructs is not uncommon in practice.

Considering 'Article 2: Multivariate Dependence', besides the one-factor model, an orthogonal three-factor population model with independent cluster structure (i.e., the factors are unrelated and each of the items only loads on one factor) was also applied. The clusters were of similar size and the items within a cluster were expected to have equal correlations of size r between them. Here one could argue that in multi-factor models the factors are usually correlated, yet again a pragmatic choice was made here. For the example in 'Article 2: Multivariate Dependence' the specific type of model used doesn't matter. It actually shows that CFI is insensitive to model type given a constant degree of multivariate dependence as given by $|\mathbf{R}|$. From that viewpoint, the specific set of models considered in 'Article 2: Multivariate Dependence' does allow for easy calculation of the model-implied correlation that belongs to a specific value of the determinant; specifically due to the lack of between-factor correlations in combination with the constant cluster structure within each factor. Surely, translation between determinants and correlation matrices can also be done for other population models, but this will not be as straightforward as with the used models.

Experimental Factors. In 'Article 1: Metric Space', all three components related to the CFI metric space were manipulated and used in a full-factorial design. At the minimum, it was made sure that the foundation of the experimental manipulations was informed by general theory or applications in practice. Further extension of the chosen levels of the experimental factors would of course allow for an illustration of the effect of the experimental factors over an even wider range of settings. With additional or alternative factor levels, the value-specific recommendations in 'Article 1: Metric Space' might somewhat change. Yet again, the general underlying baseline principle does not change and the main ideas will still generalize quite well.

In 'Article 2: Multivariate Dependence' on the other hand, the simulation served for didactical purposes and experimental manipulations were kept to a minimum. The

conditions were set up in a way that there are two data-generating population models that have an equal determinant, equal within-factor item correlations, or equal average correlation (see also Table 1 in the corresponding article). The results of 'Article 1: Metric Space' were used to inform the selection of the two scenarios. Specific conditions were selected based on whether we could expect CFI to work well for the one-factor model (i.e., [relative] low model rejection rates given that the used model was correctly specified). In both scenarios 1 and 2, the number of items p, as well as sample size n were kept constant, and only the strength and/or correlation structure was adjusted to comply with the primary setup. The results do generalize to other conditions as well, although the pattern of results is slightly more or less pronounced.

Population Covariance Matrix. In creating the population covariance matrices, the population variances were obtained from a specific uniform distribution: $s \sim \mathcal{U}(.75, 2)$. The decision on using this specific interval has been an arbitrary choice. Yet it does not have any consequences as any other choice would have led to similar results (given estimated models of interest that perfectly reproduce the variances, as is the case here). In this part of the process, the correlations as specified in the model-implied correlation matrix are the most important element.

Estimated Model. For the study in 'Article 1: Metric Space', the true population model was refitted in step 4. The main consideration was that it seems to be common in practice for people to act on finding support for their model. Even if this means applying different types of adjustments to the model until it shows adequate fit (e.g., Ropovik, 2015). Therefore, the focus was on evaluating the performance of CFI in the ideal situation of a correctly specified model. This also indicated that there was no error related to the misspecification of the model and that only sampling variability will play a role in performance.

At the same time, this also points in the direction of considering alternative misspecified models. Some people actually argue that evaluating performance in non-optimal situations (i.e., model not being correct in the population) might be more valuable in practice (e.g., MacCallum, 2003). The difficulty that arises here is that misspecification

needs to be modeled in some way. What makes it even more difficult is that the impact of misspecification will be dependent on the strength of the other paths/correlations in the models, which directly implies that the impact will be different across different conditions; what might be ignorable in one situation, can be critical in another.

For the limited conditions considered in 'Article 2: Multivariate Dependence', we did try to address the impact of misspecification on the performance of different fit indices. To make sure that the misspecification would be sufficiently large to clearly show up in the results, not getting caught up in the differential impact of smaller misspecifications (e.g., omitting a certain number of paths), we fitted models in step 4 that were plain wrong (i.e., fitting a one-factor model when true population model is a multi-factor model and vice versa). Clearly, absolute levels of misspecification should still not be compared, with the misspecification being realized differently. Yet we hoped that this would give a very crude indication of what can be expected performance-wise in the extreme case where the estimated model is incorrect in the population. With the fit indices ideally indicating 'bad' fit with respect to the rules-of-thumb.

1.5.2 General Design Challenge

A general threat to more exploratory simulation studies is that they might not always reflect the process you have in mind. Consider how manipulating one experimental factor, might also impact other theoretical important factors that are related to the outcome. One should be aware that failing to address these confounding relations might give rise to misleading results and recommendations. In our case, the explicit decomposition of the null baseline model guided and created more awareness of how the different components are related and allowed us to shed light on examples of confounding relations found in the literature. In 'Article 2: Multivariate Dependence' this had particular emphasis as we showed how being aware of the underlying processes can help to address potential misrepresentations with respect to the role of model type on model fit evaluation and to prevent flawed conclusions. More detailed information about the different examples will follow below.

In 'Article 1: Metric Space' for example, the confounding relation revolved around

manipulating the number of variables. In general, the effect of increasing the number of variables does not stand on its own, as the determinant $|\mathbf{R}|$, which is a key element in the null baseline, is also dependent on the number of variables. This also implies that any effect found cannot be solely contributed to changes in the size of the model. Theoretically speaking it is possible for the determinant to not be affected by increasing the number of variables, but only in the very specific case when the added variables have zero correlations with the other variables and each other, which is not only highly unlikely in practice but also defeats the purpose of modeling those variables.

'Article 2: Multivariate Dependence' showed that there is a confounding relation between the determinant $|\mathbf{R}|$ and model type. To evaluate the alleged impact of model type on the performance of fit indices, we introduced different data-generating population models. Yet, by changing the structure of the population model, one also readily changes the degree of multivariate dependence in the population, as each model differently defines where the correlation in the population correlation matrix can be found (i.e., generating data based on different population models will impact the determinant of the sample-observed correlation matrix). This also implies that the baseline fit for each of the model types will be different as the underlying data will be different. With large differences in the null baseline, it is no longer possible to just attribute any difference in fit to the different types of models and that is something one should be aware of. Only if we take this confounding relation into consideration and adjust the population correlations for the difference in determinant, fair comparisons among different model types can be made. For the CFI this would in fact mean that sampling distributions would be entirely equivalent when estimating the correctly specified model.

1.6 Reflections on Model Fit Evaluation

1.6.1 Remaining Challenges

While addressing the basic mechanisms underlying CFI in 'Article 1: Metric Space' and 'Article 2: Multivariate Dependence' did shed some light on the performance of CFI and its rule-of-thumb, it is probably too ambitious to expect this to immediately change how people will use fit indices in practice. In general, it can be expected that changes

in rule-based procedures might be hard to accomplish. For quite some time now, calls have been made to promote more nuanced use of fit indices. Yet, often it seems, at least to me, as if they merely propose optimizing 'non-ideal' procedures such that users can maintain current practice of using rules-of-thumb for model fit evaluation.

One example can be found in 'Article 1: Metric Space', where we also looked at an additional qualification stating that "CFI should not be computed if the RMSEA of the null model is less than .158 or otherwise one will obtain too small a value of the CFI" (Kenny, 2015). The simulation results showed that even adopting this extra fixed criterium still does not work (i.e., correctly identifying the true model as a good fitting model, with a binary decision rule that works at least 95% of the time). Yet, the underlying idea might still have some merit if we take it as a more general indication of a situation when not to use fit indices (i.e., in case of low data correlation, incremental fit indices might not be that informative due to high variation in performance).

Another example can be found with McNeish and Wolf (2021, 2022) who also made an attempt to promote more nuanced model evaluation. They tried to improve the generalizability of rules-of-thumb by providing flexible rules better suitable for the specific context a researcher is working with (i.e., rules being adapted based on model and data characteristics). In theory, the new rules-of-thumb can be easily obtained with their web-based application. These rules are based on the ability of a specific fit index to distinguish between the researcher's model (treated as if the model was true) and different levels of misspecified models. One drawback is that currently only a limited number of situations are considered (i.e., set of models, data type, and estimation method) and researchers don't have an influence on the type of model misspecification considered.

While both examples strive for more nuanced decision-making and reporting, the adjusted procedures don't necessarily provide an indication of why this is important or how the fit indices work. Thus one important question that we should ask ourselves is how we can increase awareness about the nature of fit measures and incorporate what we have learned (i.e, the role of the null baseline in model fit evaluation and model comparison) into practical recommendations? In both articles, we urged for explicit reporting on fit

of the null baseline model. In absolute sense, reporting on the baseline does not provide a direct indication of whether or not the size of the metric space is sufficient for model differentiation, as this will still depend on context. Being able to use this information for better interpretation of the magnitude of incremental fit indices will require some insight about common values for the baseline within a specific research context, yet this will take time to establish. In the meantime, we hoped that, at the minimum, explicit reporting of the baseline would make the impact of the baseline visible when interpreting and/or comparing values of incremental fit indices. This should be a relatively straightforward addition when reporting results, with the necessary information being easily extracted from default software. Based on the review by Jackson et al. (2009) we also consider this a relevant call in practice, as they found that reporting on baseline statistics (i.e., χ^2 and degrees of freedom) is only sparsely done (i.e., only considered in 7.2% of the evaluated studies). For the studies that do report on this, it is however not clear if or how these values were being used, so maybe there is still more room for improvement here.

1.6.2 Change in Model Testing Strategy

In an ideal situation, we might not just want to optimize existing procedures, instead, we might need to strive for a change in strategy when interpreting incremental fit indices. The most important feature of model evaluation with fit indices is that it is not a binary process, but instead, fit indices are continuous measures of model-data fit (e.g., Hu & Bentler, 1998). With current practice, where people are holding on to rules-of-thumb and focusing on the 'absolute' fit of a single model, this feature is disregarded. Incremental fit indices are also considered valuable in comparing substantive competing models. So instead of using them for a single model in a search for model adequacy, it would be preferred to use them in a model testing strategy in line with original recommendations (e.g., Bentler & Bonett, 1980).

Comparing models for a single data set has the advantage of a similar baseline metric for interpretation and being able to compare the magnitude of the incremental fit indices more gradually. Conceptually, comparing models in this way should also provide more information about the practical importance of the specific parts that differ across the

models (Bentler & Bonett, 1980) in terms of deviations from the null baseline model. With current practice, this type of part-specific information ('local misfit') quickly disappears when summarizing the fit of a single model by means of a 1-number summary. In practice, it might still be hard for people to let go of the rules-of-thumb completely. It has for example been questioned how fit indices can be practically used to compare between competing models if there are no criteria that indicate what difference is considered meaningful (e.g., Marsh, 1998). So the difficulty here might be that there is still some personal intuition or reasoning required about what reduction in misspecification is considered meaningful in a specific situation.

Clearly, none of these recommendations will solve all issues at once. For some, the issues related to the use of fit indices and their rules-of-thumb are sufficient to warn against using any of them (e.g., Barrett, 2007). For others, the adequacy of a model should not be judged in isolation but together with sample size, model complexity, or more local measures of fit (e.g., Brown, 2015; Kline, 2016; Miles & Shevlin, 2007; Sobel & Bohrnstedt, 1985). Yet, if we eventually want to use fit indices in a way that does justice to what these measures represent, we might need to start by changing how we learn about incremental fit indices and their rules-of-thumb. Hopefully, 'Article 1: Metric Space' and 'Article 2: Multivariate Dependence' contributed to this by putting the meaning and behavior of incremental fit indices into perspective and reminding people of the underlying mechanisms that play an important role here, as a starting point to evoke more deliberate model fit evaluation in practice.

1.6.3 Choice of Baseline Model

It has been stated that 'the incremental fit indices depend critically on the availability of a suitably framed [baseline] model' (Bentler & Bonett, 1980, p.604). For selecting or specifying an appropriate baseline model, it has been recommended to choose the most restrictive model that would still be considered in practice (e.g., Bentler & Bonett, 1980). In practice, this led to a default use of the null baseline model. Yet, one can wonder whether this is a logical choice. What value can be assigned to a model of interest if it is compared against a model that expects zero correlation across all variables? It can be said

that the evaluation that is being made here is quite liberal and in general it is expected that not much is needed for a model of interest to demonstrate improvement in fit. Or at least one should hope that a theoretically justified model does better than 'nothing' (i.e., absence of any relation). Yet the fact that improvement might be perceived as relatively easy, also implies that this might not be a direct test for the strength of the model of interest. One might be able to ascribe more value to the model of interest if it was tested against a stronger competitor.

A potential stronger competitor can be found in the rather general but reoccurring theoretical concept of a so-called crud factor. The main idea here is that in psychological and behavioral research all variables are correlated with each other to some degree, even though clear theoretical justification for these correlations might be lacking (e.g., Meehl, 1990b). If we know that things tend to correlate, should we not take this knowledge into account when selecting a baseline model? Failing to do so might lead to comparisons that are misleading, as a model with zero correlation can no longer be considered a proper and theoretically defensible model in practice (Rigdon, 1996). In addition, the model of interest might otherwise also be valued for its ability to capture correlation in the data that has no theoretical support whatsoever (Rigdon, 1998a). Yet, how to use this prior knowledge in selecting or specifying a proper baseline model?

In the null baseline model, the correlations between the variables are constrained to be zero, while the mean and variances are freely estimated. Rigdon (1998a) for example proposed one way of accounting for the crud factor by promoting an *equal correlation* baseline model. In this model, the correlations between the variables are no longer assumed to be zero, but instead, they are constrained to be equal, without setting explicit expectations about the crud factor effect. In contrast, Sobel and Bohrnstedt (1985) argue that in order to make progress, it is not only about including prior knowledge, but the selection or specification of the baseline model should also be driven by explicit theoretical considerations regarding the relations. Building on this, an even stronger competitor might then be formed and more information might be gained by incorporating theoretical predictions about the direction of the effects (see also Rigdon, 1998b), or alternatively,

by putting a number on the relations that could be expected by default (see also Meehl, 1990b). In practice of course, one would first need to know what would be the appropriate size (and direction) of the crud factor. Based on experience, Meehl (1990a) expresses that it might not be unreasonable to expect estimates of $r = 0.30$. On the other hand, Ferguson and Heene (2021) are more cautious and give $r = 0.10$ as a lower-bound estimate. Orben and Lakens (2020) plea for a more structured estimation of the crud factor effect within different research areas, as a clear systematic overview is currently lacking.

Going forward, it would be important for researchers to reach a consensus about what would be an appropriate baseline model within a specific domain and given specific applications. In general, this can be any model as long as the baseline is nested within the model of interest (e.g., Widaman & Thompson, 2003). However, at the same time, we should not forget that changing the baseline also has an influence on the model evaluation process. First, in default software, the calculation of fit indices is based on the default null baseline model. Even though it is possible to set the baseline manually[7], it does require researchers to adapt their standard procedures (e.g., Widaman & Thompson, 2003).

Second, rules-of-thumb for model evaluation are even less applicable in this situation, as the simulation studies on which they are based used the default null baseline model in the calculation of the fit indices. While some plea for a re-examination of the existing rules-of-thumb (e.g., Rigdon, 1996; Widaman & Thompson, 2003), others have stated outright that "if there is not sufficient generality across different applications in the use of M_0 as a worst fitting model that can be used to anchor the lower end of incremental fit indices, then $M_=$ [the new baseline], for which no guidelines are even offered, must be even less useful in this respect" (Marsh, 1998, p.81). If anything, this restates the idea that people need guidance in making qualitative judgments about models.

This also brings up the question if this approach will ever really take off in practice. In the late 90's Rigdon (1996, p.377) stated that "it is unlikely that there will be a movement toward an alternate baseline model anytime soon" and so far it still does not seem to be

[7]For example with R::lavaan (Rosseel, 2012), calculation of fit indices can be based on an alternative model, by providing this alternative model to the `baseline.model` argument in the `fitMeasures` function.

common practice. Yet, despite the complications, changing the baseline is not something that should be disregarded by default. In general, there will be situations in which the null baseline model is considered an inappropriate comparison model (e.g., due to nesting issues), and as a result, the fit indices can no longer be interpreted as a valid measure of improvement of fit (for more guidelines on how to specify a baseline model suitable for a specific research context, see Sobel & Bohrnstedt, 1985; Widaman & Thompson, 2003). Consequently, adjustments to the baseline should and cannot always be avoided.

Additionally, Sobel and Bohrnstedt (1985) argue that a comparison to a stronger, more meaningful baseline model is needed to contribute to scientific progress of the current state of knowledge (see also Chapter 3). Following Sobel and Bohrnstedt (1985), one could say that the value of a model of interest is in its ability to better capture the data that could be done based on the current state of affairs. Yet, substantive improvement cannot be proven in comparison to a null baseline model where there is 'nothing', no theory nor prior knowledge. Therefore, Sobel and Bohrnstedt (1985) believe that the baseline model should not be the worst-fitting model unless the research is completely exploratory, but instead, it should be one that incorporates the current state of knowledge and theory within a given domain. The application of the crud factor effect is just one illustration.

References

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd). Guilford Press.

Ferguson, C. J., & Heene, M. (2021). Providing a lower-bound estimate for psychology's "crud factor": The case of aggression. *Professional Psychology: Research and Practice*, *52*(6), 620–626.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424.

Jackson, D. L., Gillapsy, J. A., & Purch-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23.

Kenny, D. A. (2015). Measuring model fit. http://davidakenny.net/cm/fit.htm

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*(2-3), 220–239.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139.

Marsh, H. W. (1998). The equal correlation baseline model: Comment and constructive alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 78–86.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391–410.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82.

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000425.

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods.* Advance online publication. https://doi.org/10.3758/s13428-022-01847-y.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244.

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, *42*(5), 869–874.

NESH. (2016). Guidelines for research ethics in the social sciences, humanities, law and theology. www.etikkom.no

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*, 1148–1172.

Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247.

Ören, T. (2000). Responsibility, ethics and simulation. *Transactions of the Society for Computer Simulation International*, *17*(4), 165–170.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(4), 369–379.

Rigdon, E. E. (1998a). The equal correlation baseline model for comparative fit assessment in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 63–77.

Rigdon, E. E. (1998b). The equal correlation baseline model: A reply to Marsh. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 87–94.

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, Article 1715.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with monte carlo simulation. *Journal of Statistics Education*, *24*(3), 136–156.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. *Sociological Methodology*, *15*, 152–178.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37.

2 Application 2: Characterization of random responders in the TIMSS 2015 student questionnaire

2.1 Background

Since the late '50s, there has been ongoing interest in comparing student achievement and its determinants across countries. Over the years, the number of participating countries has grown and more and more topics have been covered in different international large-scale assessments (ILSA) in education (Gustafsson, 2008). Not only does the data collected by these international large-scale assessments provide us with a wide variety of research opportunities, it is certainly used as well. Over the years there has been a tremendous increase in the number of studies that use these assessment results to answer a broad range of research questions (e.g., Hernández-Torrano & Courtney, 2021; Hopfenbeck et al., 2018). Yet, paying attention to the quality of these assessment results is crucial (e.g., Gustafsson, 2008) as the quality of the conclusions that are drawn is dependent on the quality of the corresponding data.

One major consideration is that these types of assessments are typically low-stakes for the students (e.g., there are no consequences for performance and no explicit benefits for participation). This in itself already makes different people wonder whether or not students' responses can still be trusted in a sense that they still reflect true knowledge, abilities, or opinions related to the assessment content. If students would not respond accurately or thoughtfully, this could potentially lead to problems with the use and interpretation of the assessment results. Thus from a quality assurance viewpoint, it is important to pay attention to undesirable or invalid response behavior.

2.1.1 What Has Been Done?

While a lot of research addresses concerns about how genuine students are responding, there is still a lot of ground to clear to determine the prevalence, generality, and impact of this invalid response behavior within the context of ILSAs. Within this context, the two most common approaches to address the validity of item responses are using either self-report measures or response times. Both approaches have some distinct features related

to measurement and can be differentiated from the approach we adopted in our articles, for a general overview see Table 2.

Self-report Measures. Using self-report measures can be described as an indirect measurement approach. The measures are used to collect information about the quality of item responses on the assessment, often in relation to achievement, yet there is no focus on the actual responses given on the assessment by the students. With this approach, students are merely asked to rate their behavior on the assessment. For example, students can be asked to indicate to what degree they 'put in good effort throughout the ... test' (e.g., Hopfenbeck & Kjærnsli, 2016). While these measures are generally easy to use and flexible in design, one major drawback is that they are often very general or global measures, in a sense that they try to say something about a test or questionnaire as a whole. Yet working on such a global level also implies there is no direct link with the actual responses that people want to draw conclusions about. In addition, people need to be aware that self-report measures are potentially biased as they themselves also rely on students responding genuinely as expected to this measure (cf. circularity). If they don't, this might already distort conclusions about the quality of the actual responses of interest.

Response Times. The response time approach can also be described as indirect. This approach is applied in achievement testing and seems to focus more on *how* students responded than on the actual response itself. Here, reaction time information is collected and used to identify responses that are given too fast (i.e., 'rapid-guess') for students to properly process the question and thus, for the response to be considered reflective of a student's true knowledge or ability (e.g., Wise & Kong, 2005). In contrast to self-report measures, the response time approach does work on item-level, meaning that the quality of each item can be addressed separately. Yet a difficulty here is that the conclusions that can be drawn are to some extent dependent on an arbitrary cutoff to distinguish between students showing valid or "too fast" invalid response behavior. In addition, response time indices can be said to be general measures, as they do not pick up on a specific type of response behavior, but can rather pick up multiple response patterns. However currently,

the biggest disadvantage is that item-level response times are not always available. For example, assessments are either not computer-based or response times are only available for the achievement part of assessments and not for the surveys. In addition, there are some questions about how ethical this approach is, as response time data can be collected and used without informing the respondents (Leiner, 2019).

Response-based Approach. There are some instances where response-based approaches are used in the context of ILSAs (for an example of the application of person-fit indices see e.g., Hopfenbeck & Maul, 2011). Yet, like the response time-based approach these measures can be very generic/aspecific in a sense that they pick up on multiple response patterns. In our adopted response-based approach, we tried to create a direct link between the conceptual definition of invalid response behavior and the quality of the item responses. More detailed information about this approach will be provided later, but for now it should be noted that, like the response time approach, our approach is more locally oriented, using information related to the items of interest. The difference is that while our approach uses the actual item responses as provided by the students, the response quality is addressed at scale-level. With respect to the implementation of the approach, a sufficient number of items and response options per item are required to enable the detection of invalid response patterns.

Table 2

Comparison of Different Approaches to Invalid Response Behavior.

Features	Approach		
	Self-report	Response-time	Response-based
Direct vs. Indirect	Indirect	Indirect	Direct
Measurement level	Global	Local: Item-level	Local: Scale-level
Implementation	Easy & economical, yet potentially biased	Not always available	Sufficient number of items / categories required

Note. Our approach: response-based

2.2 Conceptual Framework of Articles

2.2.1 Random Responders & Random Responding

In our studies, we specifically focused on random responding, one type of response behavior often associated with the low-stakes context and that is generally perceived to be harmful. In the different articles, we conceptually describe random responders as students who at times provide "responses without meaningful reference to the test questions" (Berry et al., 1992, p.340, see also 'Article 4: Where' & 'Article 5: Who') or "unrelated responses ... as if (s)he was not even reading the items and choosing a response option randomly throughout" (van Laar & Braeken, 2022, p.4, see Article 3). Either way, random responding can be situated as a type of non-response, where the responses that are provided by the students do not contain valid information with respect to the assessment content (see Article 6).

Labelling Controversy. The 'Random Responder' label has been proven to be a rather controversial choice with regular reactions by reviewers indicating that one of several alternative terms would be more proper. I would like to explain here why our adopted definition/labeling is not an unfortunate mistake, but an intentional choice.

With respect to alternative terms for random responding, there is for instance a large literature base in social science survey research that talks in terms of careless or insufficient effort responding (e.g., Huang et al., 2012; Meade & Craig, 2012). This formulation emphasizes the underlying causes of invalid responding and individuals' underlying intentions. I personally feel strongly that these causes/intentions cannot be directly verified as you would somehow be expected to successfully regulate post hoc introspection of the individual's thinking while responding to the questionnaire. This seems hard to accomplish based on the limited information we have at hand: the pattern of given item responses. Furthermore, the terms careless/insufficient effort responding are umbrella terms in a sense that they could accommodate many different response patterns, both random as well as more systematic.

The random responding formulation is not our own invention. It seems to have its roots in psychology/personality assessment (e.g., Baer et al., 1997; Berry et al., 1991)

and is still in use today (e.g., Credé, 2010; Kim et al., 2018). Even with this formulation, it is in practice quite easy to speculate about intentions as well. Yet the more technical papers emphasize how the resulting item response pattern appears to an objective outside observer. It is this latter perspective that we also have adopted, without putting forward any requirements on why the response pattern emerges. Thus, in this sense the chosen label perhaps also immediately signals my research perspective on the issue of invalid random responding.

An alternative terminology that could perhaps be considered more neutral, but also much broadly applicable, can be found in the person-fit literature, where originally Levine and Rubin (1979) introduced the term measurement appropriateness to discuss persons who responded in line or counter to a measurement model of focus.

2.2.2 Operationalization of Random Responders

In our own approach, we use a relative comparison definition for random responders based on a contrast between a measurement model and a null baseline model. We make the implicit assumption that there is at least one group of students that respond in a regular fashion to the questionnaire of interest such that a measurement model does approximately apply to at least part of the population. We expect there to be another group of students for which this measurement model is less fitting and whose response patterns are more compatible with a uniform null baseline model reflecting more random response behavior.

In our articles, we considered there to be two different groups of responders, regular responders and random responders, and both groups are expected to show different response behavior. The underlying idea is also represented in Figure 8 and provides a direct link with the observed item response patterns given on a survey scale.

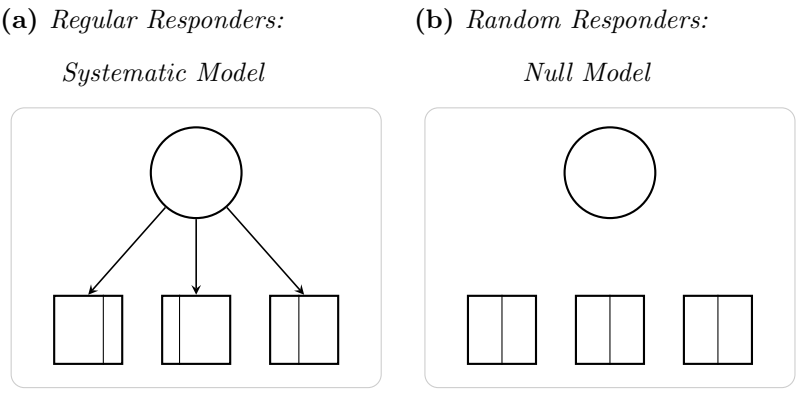
Regular responders were expected to respond consistently according to their own opinion and beliefs. In terms of model specification, these students follow a systematic measurement model (see Figure 8a) where there is a common latent trait (i.e., circle) which can be seen as the common cause underlying the students' item responses (i.e., squares) as indicated by the arrows going down. Another group of students, labeled the

random responders, were expected to provide unsystematic random response patterns across the items, such that the responses no longer reflect a student's own opinions or beliefs. The latter model can be seen as an application of the null model, where all observed variables are assumed to be uncorrelated, but now with the additional restriction that observed variables are uniformly distributed (i.e., each response category has an equal probability of being selected). In Figure 8b this is represented by the latent trait no longer being connected with the item responses and all squares being divided into equal (category) parts.

Random responders are students who have response patterns that are more similar to what can be expected under this null model (i.e., students providing unrelated responses) compared to what could be expected from the other group of students under the measurement model. One important consideration here is that we do not imply in any way that random responding is something that can be regarded as an attribute of students or something that they do deliberately. The operationalization of random responding does not in any way indicate how or why random response patterns came to be.

Figure 8

Framework to Define and Operationalize Random Responders.



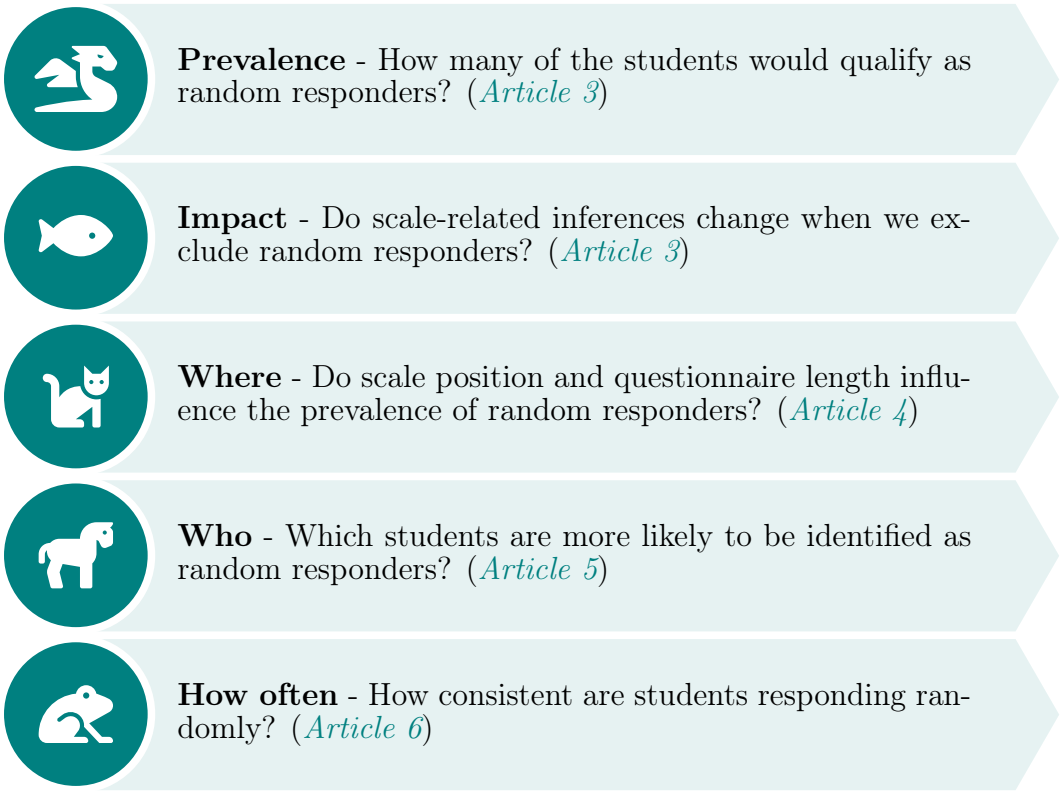
Note. Symbols follow standard path diagram conventions, with squares representing observed variables (i.e., item responses); circles, latent variables (i.e., trait to be measured by the scale of items); arrows indicating dependence relations; vertical lines, response category thresholds. Adapted and reprinted under the terms of CC-BY-NC from "Random responders in the TIMSS 2015 student questionnaire: A threat to validity?" by S. van Laar and J. Braeken, 2022, *Journal of Educational Measurement*.

2.2.3 Research Questions

While all studies within Application 2 revolve around random responding, they focus on different aspects of this behavior/phenomenon. For example, questions that are considered logical to answer are related to the following pointers: (i) prevalence, (ii) impact, (iii) where, (iv) who, and (v) how often. For an overview of the specific questions being answered in each of the articles see Figure 9. One question we did not address is *why* students are random responding. As indicated earlier this is due to the lack of data or means to study this within what is given in the current research context. Yet, I will return to this open question in a later section.

Figure 9

Random Pictograms Representing the Different Articles about Random Responders.



2.3 Trends in International Mathematics and Science Study (TIMSS)

In the different articles, we have tried to address the aforementioned questions using the TIMSS 2015 student questionnaire. In general, TIMSS is outlined as an international large-scale assessment used to monitor mathematics and science achievement

among fourth- and eighth-grade students across different countries (e.g., Mullis & Martin, 2013). In 2015, the sixth round of the TIMSS assessment was conducted. Besides the main focus on student achievement, TIMSS 2015 also contains a Context Questionnaire Framework which is to provide additional background information about the different contexts for learning (e.g., community, home, school, or classroom) mathematics and science as reported by students, parents teachers and/or principals (e.g., Mullis & Martin, 2013).

The student questionnaire is one of the contextual questionnaires that are part of the TIMSS 2015 assessment. Besides some basic demographics and background information about the home and school context, the student questionnaire's main focus is on students' attitudes towards learning mathematics and science (Mullis & Martin, 2013). This type of contextual information is often used in relation to student achievement and to compare differences in educational learning outcomes across countries (Mullis & Martin, 2013). With more than 580.000 participating students in TIMSS 2015 (Mullis et al., 2016), this provides a huge amount of potentially valuable data.

2.3.1 Why TIMSS?

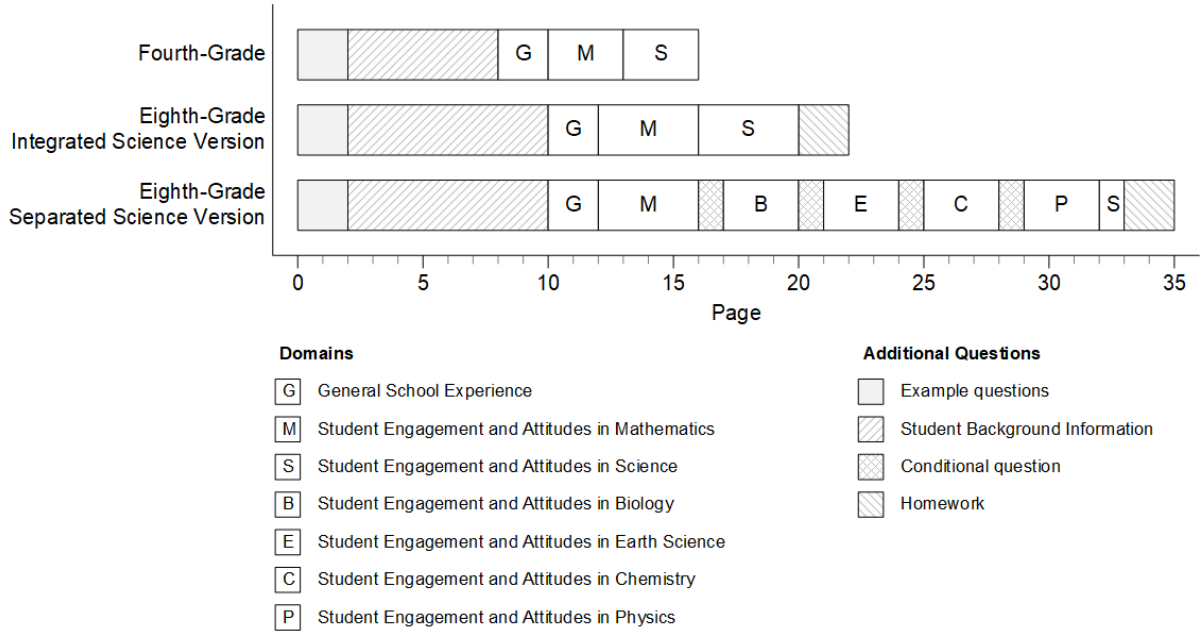
Clearly, data from other international large-scale assessments are also available. Yet, what I personally appreciate in TIMSS is how well information regarding the scales in the student questionnaire is documented. In addition, the TIMSS eighth-grade student questionnaire comes in two versions, allowing for the investigation of the role of questionnaire characteristics on response behavior.

Figure 10 shows an overview of the scale structure of the student questionnaire(s) in the different grades. Following the documentation, the fourth-grade student questionnaire contains eight standalone scales, while the student questionnaire for the eight-grade contains either 10 or 19 standalone scales (for information on the specific scales see Figure 10). In the latter case, the specific number of scales is dependent on the structure of the science program for a given country. In the integrated science version, science is treated as a single subject, while in the separated science version each science domain is addressed separately. For the latter version this means that the "Students Like Learning

Science", "Students' View on Engaging Teaching in Science Lessons" and "Student Confident in Science" scales are available for every science subject separately (i.e., in order of appearance: Biology, Earth Science, Chemistry, and Physics). The science scales in both student questionnaires do have the same structure. For the items in the separated student questionnaire, it is just the word 'science' that is replaced by the name of the specific science domain (e.g., 'I enjoy learning science' vs 'I enjoy learning Chemistry').

Figure 10

Structure of the Different Student Questionnaires in the TIMSS 2015 Assessment.



Note. With respect to the structure of eighth-grade student questionnaires, in the integrated science version science is treated as a single subject, while in the separated science version each science domain (i.e., biology, earth science, chemistry, and physics) is addressed separately. With respect to the number of scales within each domain, each page corresponds to one unique scale. This comes down to eight scales for the fourth grade, 10 for the eighth-grade integrated science version, and 19 for the eighth-grade separated science version. Within the context of General School Experience, the scales are 'Students' sense of school belonging' and 'Student bullying'. Within the context of Student Engagement and Attitudes, each of the student questionnaires contains the 'Students like learning *subject*', 'Students' views on engaging teaching in *subject* lessons', and 'Students confident in *subject*' scales for both mathematics and science (or biology, earth science, chemistry, and physics separately). In addition, both eighth-grade versions also contain the 'Students value mathematics' and 'Students value science' scales.

2.3.2 Why the Student Questionnaire?

In practice, the student questionnaire is used by a lot of researchers to answer a wide variety of research questions and it serves an important role in putting the achievement results into context. In addition, the student questionnaire also benefits from larger sample sizes, more scales, and lower non-response rates than for example the teacher and principal questionnaires. Thus, more data to systematically examine our research questions under a larger and wider scope.

Remarkably, up to this point, the student questionnaire has received less attention (time- and resource-wise) compared to the achievement part of the assessment (e.g., Rutkowski & Rutkowski, 2010), not only from the organizational side but also with respect to validity research. Yet, with the student questionnaire playing such a prominent role in research, it is important to closely and thoroughly inspect the quality and potential limitations of this data. As a start, the articles within Application 2 can hopefully contribute to this purpose by shedding some light on potential issues with the quality of the actual responses provided by the students on the student questionnaire.

2.4 Highlights of Articles

As mentioned before, all studies within Application 2 revolve around random responding, yet they focus on different aspects of this behavior/phenomenon. Table 3 provides an overview of the specific study characteristic across the different articles. In addition, in what follows I will provide a brief summary of the different articles and address the question *why* students might be random responding.

Article 3: Prevalence & Impact

Given the limited information surrounding random responding in international large-scale educational assessments, the aim of this study was to investigate the prevalence of random responders and their impact on scale-related inferences (i.e., scale score distribution, reliability, between-scale correlation, and correlations with achievement) in the TIMSS 2015 student questionnaire. To this end, a mixture IRT model approach was used to identify those students who would qualify as random responder. The results showed

that the prevalence, based on the number of students being classified as a random responder by the model, was non-zero across all countries and scales, with an average of 6%. The overall impact of random responders on aggregated-level results was fairly limited. Even though there were some differences in analysis results with and without random responders, these differences were not representing any qualitative changes.

Article 4: Where

It has been generally acknowledged that response behavior might change as students progress through a questionnaire due to changes in their subjective experience with the survey. Based on this idea, the aim of this study was to investigate the impact of two questionnaire characteristics, scale position and questionnaire length, on the prevalence of random responders. For this, we made use of the natural variation in questionnaire length in the two versions of the TIMSS 2015 eighth-grade student questionnaire (i.e., considering 10 scales in the integrated science version and 19 sales in the separated science version). The mixture IRT model approach was used to assess the prevalence of random responders for each scale and subsequently, a cross-classified linear mixed model approach was adopted to investigate how the prevalence of random responders varied as a function of scale position and questionnaire length. The results showed no support for the effect of questionnaire length, yet we did find a positive effect for scale position, with an increase of 5% in random responding over the course of the questionnaire. However, scale character turned out to be an unexpected but more important determinant. Scales about students' confidence in mathematics or science showed an increase of 9% in random responding, which is double the impact of scale position.

Article 5: Who

The low-stakes character of international large-scale assessments is often considered a contributing factor to invalid response behavior. At the same time, one might wonder if there are certain students that might be more prone to providing invalid responses in such a context. Specifically, the aim of this study was to examine which students are more likely to be identified as random responders across six different scales, related to students' attitudes and beliefs in mathematics and sciences, in the TIMSS 2015 fourth-

and eighth-grade student questionnaire. First, a mixture IRT model approach was used to assess the random responder status for each student in 22 different countries at each of the scales. Subsequently, we examined whether the prevalence of random responders was a function of grade, gender, socioeconomic status, spoken language at home, or migration background and summarized the results by means of a random effects meta-analytic model. In general, the results showed that being a student in higher grades, being male, reporting to have fewer books, or speaking a language different from the test language at home were all considered risk factors for random responding.

Article 6: How often

At the individual-level, one can wonder when someone is identified as a random responder on one scale whether their results on the other scales should be deemed invalid as well. At the same, this can be expected to be different for different students as they might respond to questionnaires in a different fashion. The aim of this study was to investigate and compare how consistent different students are in their random responding across the TIMSS 2015 eighth-grade student questionnaire. The mixture IRT model approach was used to assess the random responder status for each student in 7 different countries at each scale and subsequently, a latent class model was adopted to identify different types of random response profiles. Overall, the results showed four distinct profiles of random responding that we described as: A majority of consistent non-random responders, intermittent moderate random responders, frequent random responders, and students that were exclusively triggered to respond randomly on the confidence scales in the questionnaire. These profiles generalized quite well across countries.

Table 3

Overview of Study Characteristics in the Context of Random Responding.

	Title	Research Question	Design	Data Source: TIMSS 2015	Measures	Sample	Analysis
Paper 3	Random Responders in the TIMSS 2015 Student Questionnaire: A Threat to Validity?	• How many of the students would qualify as random responders? • Do scale-related inferences change when we exclude random responders?	Impact Study	Eighth-Grade Student Questionnaire: Integrated Science Version	• 4 scales	5 countries	• Mixture IRT model • Sensitivity analysis: Impact with and without
				Eighth-Grade Achievement Data	• Plausible values: Mathematics & Science		
Paper 4	Prevalence of Random Responders as a function of Scale Position and Questionnaire Length in the TIMSS 2015 eighth-grade Student Questionnaire.	• Do scale position and questionnaire length influence the prevalence of random responders?	Quasi-Experiment	Eighth-Grade Student Questionnaire: Integrated Science Version	• 9 scales	29 countries	• Mixture IRT model • Cross-classified linear mixed model
				Eighth-Grade Student Questionnaire: Separated Science Version	• 20 scales	11 countries	
Paper 5	Who are those Random Responders on your Survey? The case of the TIMSS 2015 student questionnaire.	• Which students are more likely to be identified as random responders?	Across-Countries Synthesis	Fourth-Grade Student Questionnaire: Integrated Science Version	• 6 scales • 5 background variables	22 countries	• Mixture IRT model • Group differences in prevalence (OR) • Random effects meta-analytic model
				Eighth-Grade Student Questionnaire: Integrated Science Version			
Paper 6	How randomly are students random responding to your questionnaire? Within-person variability in random responding across scales in the TIMSS 2015 eighth-grade student questionnaire.	• How consistent are students responding randomly?	Within-Person	Eighth-Grade Student Questionnaire: Separated Science Version	• 20 scales	7 countries	• Mixture IRT model • Consistency indices • Latent Class Analysis

2.4.1 Not Addressed: Why

One question that we have not addressed in our articles, but seems to interest many people, is *why* students tend to respond randomly. Especially the low-stakes character of the assessments is considered a contributing factor to invalid responses where scores are no longer representative of the construct under study (e.g., Wise & DeMars, 2005). In this context, speculations about the reasons underlying random responses, or invalid response behavior in general, are usually in terms of a lack of motivation/engagement/effort by the individual.

A popular theoretical framework for motivation within the field of education is the expectancy-value model (e.g., Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). In this model, a person's response behavior on a task (e.g., task choices, invested effort, persistence, and performance) is influenced by their motivational disposition resulting from the interplay between *expectancy beliefs* and *task-value beliefs* (see Figure 11). The former are a person's beliefs about being able to succeed in a certain task and the latter is more about the reasons a person puts forward for engaging with the task. A person's task-value beliefs are determined by four components: (i) how important is it for the person to do well; (ii) the level of interest in the task or how much enjoyment the person does get out of the task; (iii) the degree to which the task relates to the person's individual goals; and (iv) how big of an investment the person needs to make perform the task (e.g., time, pressure, missed opportunities).

Although an attractive general framework, I feel it falls somewhat short in adequately capturing the case of random responders in low-stakes assessments. Penk and Richter (2017), for instance, point out that there is mostly a one-sided focus on task-values in the interpretation of test-taking motivation in low-stakes testing, whereas the role of expectancy-related beliefs is hardly discussed. With no personal consequences for performance and no feedback on the correctness of the provided responses (Cole et al., 2008) in the achievement tests of the international large-scale assessments, it becomes indeed less straightforward to conceptualize what type of expectancy beliefs might still apply. To an even greater extent, this also applies to the student questionnaires, where

for most of the items there is no objectively correct answer. Thus what would 'ability' and 'succeeding in the task' look like for individual students in such a situation?

On the other side, the low-stakes character (i.e., lack of personal consequences for performance and no explicit benefits for participation for an individual student) is logically expected to lead to weak task-value beliefs which will contribute to lower motivation on the assessment, and as a consequence less effort and less valid responses. Although it is acknowledged that the degree to which the low-stakes context impacts task-value beliefs differs among students (e.g., Wise & DeMars, 2005), this simplified linear reasoning leads to quite speculative overly strong statements: "Without consequences for performance, many students will not give their best effort to such low-stakes tests; as a result, their assessment test scores may not serve as valid indicators of what they know and can do." (Wise & DeMars, 2005, p.1). These statements seem to imply that a person's motivational beliefs have a direct deterministic and stable influence on their test-taking behavior. Yet, seeing value or having positive expectancy beliefs is not a deterministic guarantee for being motivated, nor does a low motivation necessarily prevent people from responding in a regular valid fashion to the assessment. Typically only low to moderate positive correlations are found between self-reported motivation/effort measures and actual response behavior/achievement measures (e.g., Butler & Adams, 2007; Eklöf et al., 2014; Hopfenbeck & Kjærnsli, 2016). Hence, I think that it is important to acknowledge that, in contrast to the simplified interpretation of the Expectancy-Value model, there is not a deterministic one-to-one mapping between its components and that there is some uncertainty in the chain of relations. Otherwise, such statements make it too "easy for stakeholders ... to conclude that below-expected performance levels are due to lack of motivation, whether that be the true state of affairs or not" (Thelk et al., 2009, p.129).

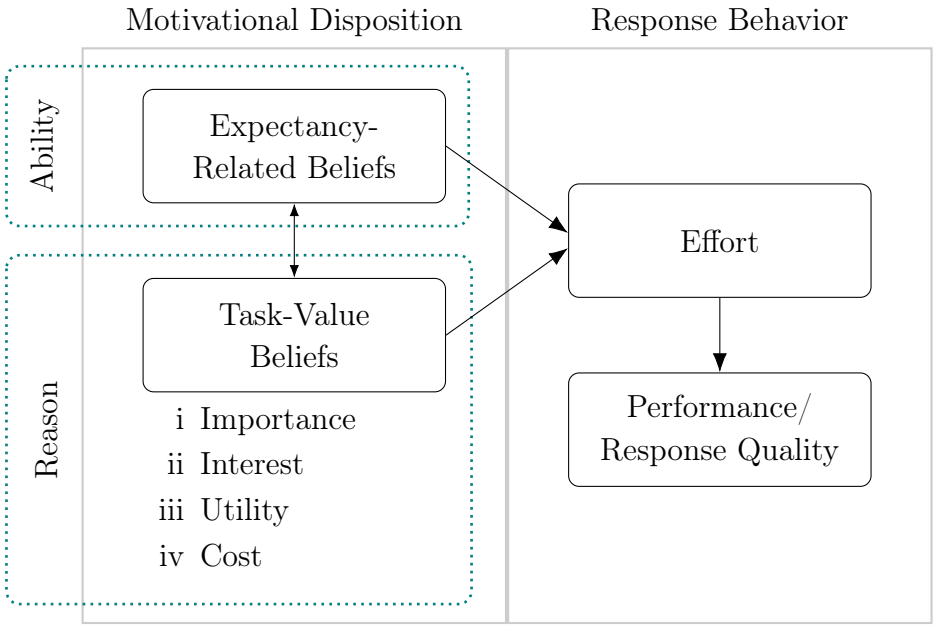
As a further complication, both a person's beliefs, motivation, and test-taking behavior are likely not a generally applicable phenomenon, but can potentially change throughout the assessment depending for example on the assessment content or duration. This would also imply that a single global measure of motivation might not be sufficient and other specific and/or dynamic dimensions need to be added. For example,

the study by Penk and Richter (2017) followed the structure of the Expectancy-Value model, and found a decrease in students' test-taking effort and perceived value within a single cognitive test at group-level, while expectancy beliefs remained rather stable. In a survey context, Galesic (2006) also reported changes in self-reported interest and testing burden depending on the varying content aspects in the blocks of the survey. General observations of response behavior being more consistent within assessment parts of similar nature go back to early classics such as Cronbach (1950) and interpretations of declining response quality as a function of test fatigue or boredom are widely represented within the literature. In our own work, 'Article 4: Where' and 'Article 6: How often' both expand on these issues and provide empirical support for the claim that random response behavior is non-constant throughout the survey. An interesting question could be whether these observed changes in random response behavior would coincide with a change in test-taking motivation and/or changes in motivational beliefs.

In conclusion, figuring out why students provided invalid random response patterns on the assessment could be very interesting, yet remains quite challenging as we currently lack a comprehensive theoretical framework and also proper measurement tools. Current attempts have mostly relied on self-report measures addressing motivation or effort on the assessment as a whole or on post hoc interviews (e.g., Butler & Adams, 2007; Eklöf, 2007; Hopfenbeck & Kjærnsli, 2016; Hopfenbeck & Maul, 2011). The former lacks specificity in their link to the actual responses provided by students, ignores the non-constant characteristic of the phenomenon, and is somewhat circular in nature (cf. asking someone by means of a survey whether they are motivated to answer such a survey). The latter relies on the ability of students to discuss one's own response processes during the assessment in retrospect, which might not always be that reliable. While these are good initial attempts, we will need to move a step further if we want a proper answer to the question of *why*. The good news is that other questions that are focusing more on actual response behavior can be answered with the current methodology and without the requirement of knowing exactly why a given response pattern was provided.

Figure 11

The Expectancy-Value Theory Model in the Context of Test-Taking Motivation.



Note. Adapted from "Change in test-taking motivation and its relationship to test performance in low-stakes assessments" by C. Penk and D. Richter, 2017, *Educational Assessment, Evaluation and Accountability.*

2.5 Method of Study

The four studies within Application 2 are all empirical studies using TIMSS data and applying quantitative methods with a hybrid mixture model as common methodological feature. Here, I will briefly introduce the history behind this model and discuss the concrete implementation used in the articles.

Latent class models are part of the larger statistical class of mixture models (e.g., McLachlan et al., 2019). These models are often used when population heterogeneity is expected with two or more groups in the population that show distinct response behavior, but where group membership is in fact not observed. In the context of educational measurement, the idea of population heterogeneity was explored when discussing the possibility of different problem-solving strategies for cognitive tests, with initial work dating back to the 1990's (e.g. Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990). Initially, the focus was on two-class mixture models of common item response models for binary responses. Yet, extensions to multiple classes, polytomous responses,

and less common item response models followed soon after (Jin et al., 2018; Sen & Cohen, 2019; von Davier & Carstensen, 2007).

One specific mixture model proposed and labeled by Yamamoto (1989) as the HYBRID model can be seen as the parent model of the mixture IRT model we applied throughout all studies within Application 2. The HYBRID mixture model consisted of one class that followed a common item response model for binary responses and other classes where item responses were expected to be independently distributed following a Bernoulli distribution where the probability of correct response potentially varies across items and classes (i.e., in line with prototypical item response patterns that are assumed to reflect different response processes). It was labeled HYBRID as it is a mixture of two different types of models: IRT and non-IRT. As pointed out by for instance von Davier and Carstensen (2007), there is also an alternative perspective on the HYBRID model. From this perspective, both mixture component models are treated as latent variable measurement models, yet one of them is much more constrained than the other. This perspective allows this type of HYBRID models to be estimated in software such as Mplus. This is also the strategy followed in our implementation of the polytomous extension of Yamamoto (1989)'s HYBRID model (for an example of Mplus syntax, see 'Article 4: Where' or 'Article 6: How often', Appendix A).

Figure 12 summarizes the formulation of the adopted mixture IRT model. The probability of the vector of item responses $\mathbf{Y}$ (i.e., $\Pr(\mathbf{Y}|C = c)$) is formulated as a weighted sum across the $C = 2$ classes of the prior probability of belonging to a class c (i.e., $\Pr(C = c)$) and the conditional probability of the response vector given that you would be a member of class c (i.e., $\Pr(\mathbf{Y}|C = c)$; see left panel of Figure 12). The formulation for the conditional probability of the response vector depends on the specific class. For one class it follows a graded response model (Samejima, 1969); a common latent variable measurement model for polytomous item responses where the item responses are assumed to be conditionally independent given the latent variable. For the other class the formulation is in line with a null baseline model; following a multinomial distribution where each response category has an equal probability of occurrence and the item

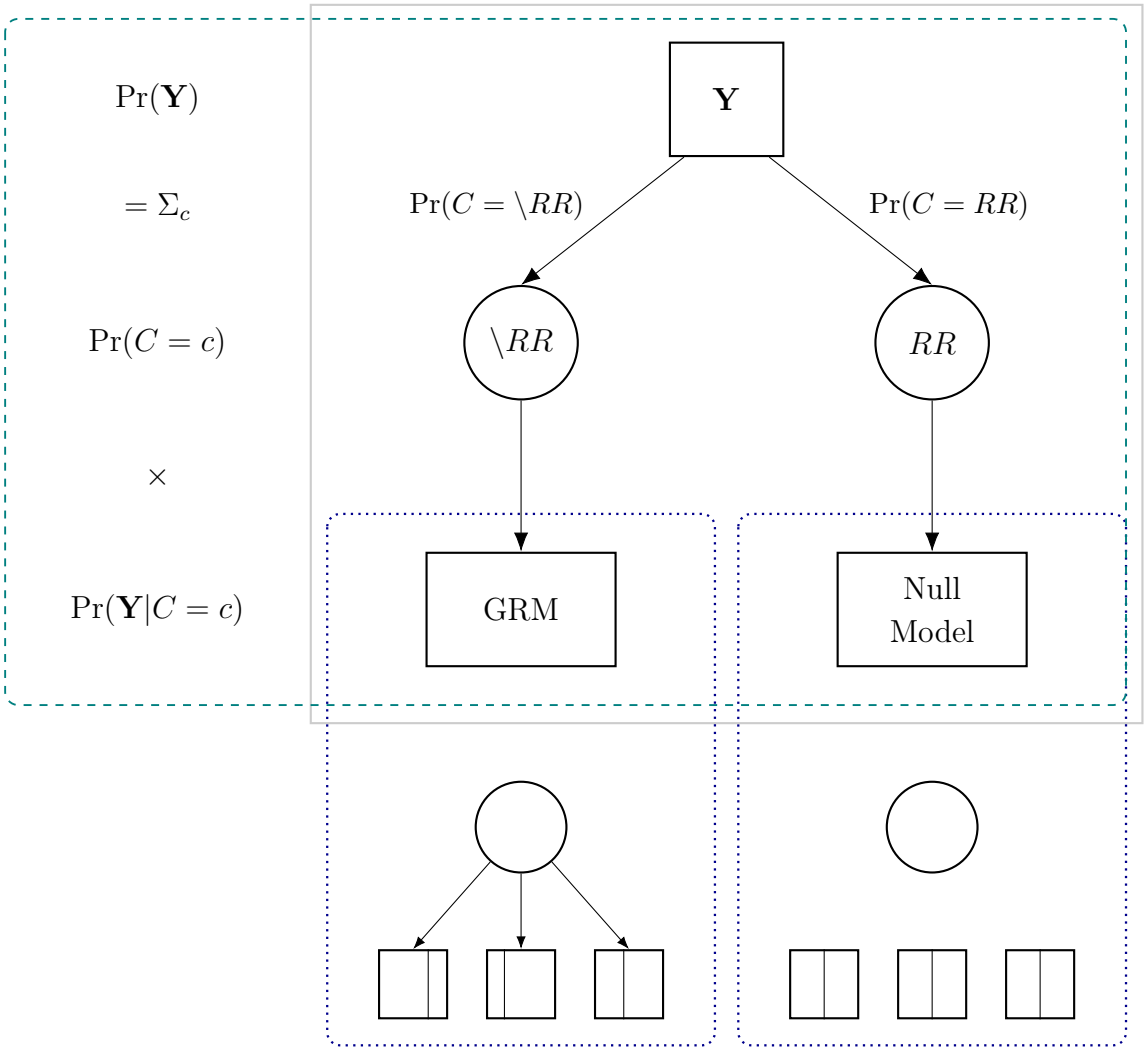
responses are assumed to be independent (see middle panel of Figure 12). The regular measurement model is obtained when the prior probability for the second class is fixed at 0. The null baseline model is obtained when the prior probability for the second class is fixed at 1. In the mixture model, the prior class probabilities by definition sum up to 1 (i.e., $\Pr(C = RR) = 1 - \Pr(C = \backslash RR)$).

For model estimation, typical algorithms for mixture models such as the expectation-maximization algorithm can be applied in combination with a multi-start procedure to counter the usual concerns of local maxima. For classification purposes, Bayes theorem can be applied to obtain the maximum a posteriori membership classification in which a person is assigned to the class for which their item response pattern has the highest posterior probability given the estimated mixture model. The latter posterior probabilities are obtained as:

$$\Pr(C = c|\mathbf{Y}) = \frac{\Pr(C \cap \mathbf{Y})}{\Pr(\mathbf{Y})} = \frac{\Pr(C = c)\Pr(\mathbf{Y}|C = c)}{\sum_{j=1}^{2}\Pr(C = j)\Pr(\mathbf{Y}|C = j)}$$

Figure 12

Graphical Representation of the Adopted Mixture Model Approach.



Note. $\mathbf{Y}$: vector of item responses; $C = c$: membership in class c; $\Pr(C = \backslash RR)$: mixture component weight for the regular responders; $\Pr(C = RR)$: mixture component weight for the random responders; GRM: graded response model.

2.5.1 Considerations for Valid Use of the Model.

While mixture IRT models have been around since the 1990s, their application beyond methodological papers is still limited. At the same time, real empirical data is often messier than any simulation design and corresponding data example. Thus, some caution is needed when using these mixture models in practice in order to secure their appropriateness and utility value.

Initially, we for instance also considered a variant of our mixture model in which the

uniformity restriction on the response distribution in the null baseline model was omitted and only independence of the item responses was assumed. Yet, the application of this non-uniform model raised quite a few conceptual questions by one of the reviewers for 'Article 3: Prevalence & Impact'. The most difficult feature of this variant is when one needs to consider the substantive interpretation of the estimated non-uniform category thresholds. Here, with this non-uniform variant, it is much more difficult to clearly demarcate the interpretation of this mixture component and consequently, this variant was abandoned and only the uniform variant was considered.

Some of the challenges of empirical data were made clear by the, for me infamous, case of Botswana where somewhat odd results appeared. Botswana showed a very high prevalence of random responders and close-to-zero or even negative factor loadings for some items in the measurement model. Yet in theory, the scale was supposed to be unidimensional. These anomalies were a signal that the measurement model was not very applicable to Botswana. Such findings lead to the quality criteria we imposed for use of the mixture model results: When (1) two or more standardized item discrimination parameters (i.e., factor loadings) were below .40 and/or (2) the classification entropy was not at least .70, the particular case and model were disregarded for further analyses. Without a strong measurement, the distinction to and meaning of the null baseline model also becomes too blurred and any further classification in random and regular responders seems unwarranted. With the transition to more digital assessments, things like response times might become more readily available, also for the survey part of the international large-scale assessments. In the future, the inclusion of such auxiliary information can potentially be used to further strengthen the model's application and the resulting classification.

2.5.2 Ethics & GDPR[8]

The data used for the studies within Application 2 comes from the TIMSS 2015 assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA). IEA has made anonymized data files publicly available in the TIMSS

[8]Part of this section is based on a course paper written for 'UV9010: Research Ethics' at the faculty of Educational Sciences, University of Oslo.

2015 International Database[9]. In line with Section 10 of the guidelines for research ethics from NESH (the National Committee for Research Ethics in the Social Sciences and the Humanities), using this type of data as part of the research project does not require further approval (e.g., from students or the Norwegian Centre for Research Data [NSD]). It has been stated that while re-using anonymized data files it is important to pay attention to the source of the data with respect to ethical processes (European Commission, 2021). In the background, all countries participating in the TIMSS 2015 assessment gave permission for releasing their country-data (Foy, 2017) and the IEA Data Processing and Research Center followed standardized procedures to protect anonymity and integrity of the data and arranged secure storage of the original data. For the newer cycles of TIMSS, more explicit information about the implementation of the General Data Protection Regulation (GDPR) rules are documented within their data protection declaration[10].

Because the data we used had already been previously collected, we had no influence on the study design or data collection itself. Yet, there are also more general norms and values that researchers should take into account (NESH, 2016). Concepts such as data integrity, reproducibility, and communication of the results are acknowledged as important factors related to research ethics (e.g., ASA, 2018; NESH, 2016). For example, in line with section F of the Ethical Guidelines for Statistical Practice from the American Statistical Associations (2018) to promote reproducibility, we provided code for the general model we used to identify random responders (see Appendix A in 'Article 4: Where' or 'Article 6: How often').

But maybe more importantly, being honest and open about results is key to good research practice (e.g., ASA, 2018; NESH, 2016). We were not in any way invested in any particular outcome related to international large-scale assessments, nor were we out to prove certain predetermined biases or hypotheses; instead we let the data speak for itself. Based on encounters with an occasional reviewer who objected to the method and the article because not enough random responders were identified and the impact was minimal, this might not always be self-evident in practice. In any case, we tried to discuss

[9]TIMSS 2015 Database: https:// timssandpirls. bc. edu/ timss2015/ international-database/
[10]IEA & GDPR: https:// iea. nl/ publications/ timss-2019-data-protection-declaration/

all the relevant decisions being made in the process, such that others can make their own conclusions.

2.6 Validity. Observable Consequences and Unobserved True Responses of Random Responders

As mentioned before, data from ILSAs are used to answer a wide variety of research questions. To ensure the trustworthiness of the conclusions or decisions that are being made based on this data, one should pay attention to the validity of the assessment results (e.g., Taylor, 2013). From one perspective it can be said that validity indicates the extent to which the evidence supports or refutes the proposed interpretations and uses of the assessment results (e.g., AERA et al., 2014; Taylor, 2013). What stands out here, is that focus is not on the assessment itself, but on its outcomes (e.g., AERA et al., 2014). In general, validity threats can then be seen as those factors that question the trustworthiness of the conclusions or decisions we make (Taylor, 2013). It is important to be aware of and collect information about these types of factors to help guide sound interpretation of assessment results (e.g., AERA et al., 2014).

The 'Standards for Educational and Psychological Testing' provide guidelines for assessing validity - and contains a specific section on the 'use and interpretation of educational assessment'. One specific threat to validity highlighted by the Standards is the influence of 'students' motivation to do well'. As we saw before, lack of motivation is one reason often brought forward for explaining why assessment results might not be reflective of what students really know and can do. This also relates to the main idea that the validity of the assessment results is to a large extent dependent on getting valid responses. Students providing random responses on the questionnaire scales would clearly be in opposition to that idea.

In what follows, the focus will be on scale score interpretation, where this will be limited to what the scores mean (e.g., scores describe the current level of attitude or belief of students). Lacking valid interpretation, the use of scale scores for further inferences and decisions will be impaired. First, I will discuss the consequences of having responded randomly on scale score interpretation at individual and more aggregate levels of inferences.

Second, I will address the complex question of what the unobserved true responses might have been if random responding had not occurred. Note that I will assume that the items used in TIMSS are representative of the corresponding constructs, thus the content and structure of these constructs will not be discussed here in this context (but are of course also up for debate when considering the larger validity question).

2.6.1 Observable Consequences of Random Responding

Scores for Individuals. Whereas there is no direct problem for scale score interpretation for individual students identified as regular responder, being identified as random responder, in contrast, threatens by its very definition a valid score interpretation for such an individual. The responses given are considered to not be consistent with their own opinions and beliefs related to the questionnaire content, and as such their scale score would just reflect random noise and be nigh non-interpretable. While we analyze the individual response patterns, we don't know why students might have responded as they did. Yet, if we in general consider them to be true random responders, a consequence would be that the scores of these students cannot be accurately interpreted as representing the students' current level of attitudes or beliefs.

Of course, there is always some uncertainty in classification/identification and there might be a chance that the apparent random response pattern is actually the true response pattern of a student (see e.g., the tweet below by Payne, 2022).

> What if the low effort, random responding we see on internet surveys is not an aberration, but a good reflection of how people really are? Maybe most people, most of the time, are just as shallow-thinking and random-responding in daily life. It would explain some things, no? (Payne, 2022)

Aggregate Scores. In practice, the assessment results in international large-scale assessments are not used to draw conclusions about individual students, but are used to draw conclusions about groups of students or differences between groups. Consider there is a group of students for which we want to say something about their average

level of attitudes or beliefs. A complication might arise as this group could potentially include some random responders. The corresponding aggregated scale scores would be the result of mixing valid scale scores of regular responders with invalid scale scores of random responders. If this is the case, can the average level we find for a construct still be interpreted as the true average for the group? One thing we could do is to investigate how scores would change if we excluded those random responders since their individual responses cannot be interpreted.

The impact of excluding random responders will depend on different factors. One could consider the size of the random responder group relative to the size of the regular responder group as for example reflected in prevalence statistics. While non-ignorable prevalence rates can indeed add to the distortion of aggregate scores, the presence of random responders itself is not a sufficient condition for finding differences in statistics and related inferences at the aggregate level. By definition, the random responders on a scale are expected to score around the midpoint of the scale. Logically this would imply that random responders have more impact on the scale mean, the further away the regular responders as a group score from this midpoint (i.e., higher score separation). At the same time, more heterogeneity within the regular responder group will decrease the impact of the random responder group on the mean and other univariate scale statistics. Thus, within a group that has rather homogeneous scores around the midpoint of the scale, random responders will only function as an additional unbiased noise factor, not substantively distorting the group mean nor related inferences. Keeping everything else equal, the impact of excluding random responders will increase as the difference in scores between the two groups increases (Credé, 2010).

2.6.2 Unobserved True Responses

Yet, this is not the end of the story with respect to score interpretation, as the validity also depends on what they would have answered if the students did not respond randomly. Only under the assumption that the true responses of the random responders would be in line with the rest of the students, removing the random responders from the data would provide an average score that can be interpreted as the average level on the construct

for the group as a whole. On the other hand, this also implies that, if this assumption is not applicable in practice, scores for the group without random responders might still not provide a valid indication of the average levels of the attitudes or beliefs within the group.

An alternative mechanism to consider is that the random responders are in fact all belonging to a specific subgroup. For example, in 'Article 5: Who' we saw that being a student in a higher grade, being male, reporting to have fewer books, or speaking a language different from the test language at home were considered risk factors for random responding. To the extent that these risk factors relate to the attitudes or beliefs under investigation, this could also potentially impact score interpretation. For example, consider a situation where only males are identified as random responders and where males and females score differently on the construct. If we would exclude the males in this situation aggregate scale score statistics might be affected by the different ratio of males and females in the group. As a consequence, problems with score interpretation might still arise depending on the severity of this imbalance.

If we would assume that the true responses of the random responders in this situation would be in line with the rest of the males, we could potentially replace their scores with plausible values from the score distribution among the other males as a solution to provide a valid indication of the average levels of the attitudes or beliefs within the whole student group. Of course, the fact remains that we don't know what responses they would have given if they had not responded randomly. However, in this situation it might still be the safer option to treat the random responders like the other males, than removing them and considering them to be like everyone else (cf. imbalance issue).

One huge caveat remains, those more problematic situations where we don't have any indication of who the random responders are and where unknown but systematic mechanisms elicit random response behavior in select individuals with specific (unobserved) true scale scores. Without further insight or data, none of the above working assumptions can support valid score interpretation in such a case.

Thus, next to the two preconditions of non-ignorable prevalence and score separation

(from midpoint), the actual consequences of having random responders present on aggregate scale score interpretation will also depend on the underlying mechanisms similar, as is the case for more traditional non-response (e.g., Groves & Peytcheva, 2008; Hedlin, 2020), that are commonly framed in terms of Rubin's (1976) framework of missing completely at random (MCAR), at random (MAR), or not at random (MNAR). Notice the similarities to these three formal mechanisms in the preceding paragraphs.

2.6.3 Relations among Scale Scores in the Presence of Random Responders

In practice, people tend to not only look at average scale scores (and other summary statistics), but often also relate scores on one scale with scores on other scales. Everything discussed thus far applies to scores on each of the scales separately, but the extent to which random responders systematically co-occur across scales adds another layer to the problem.

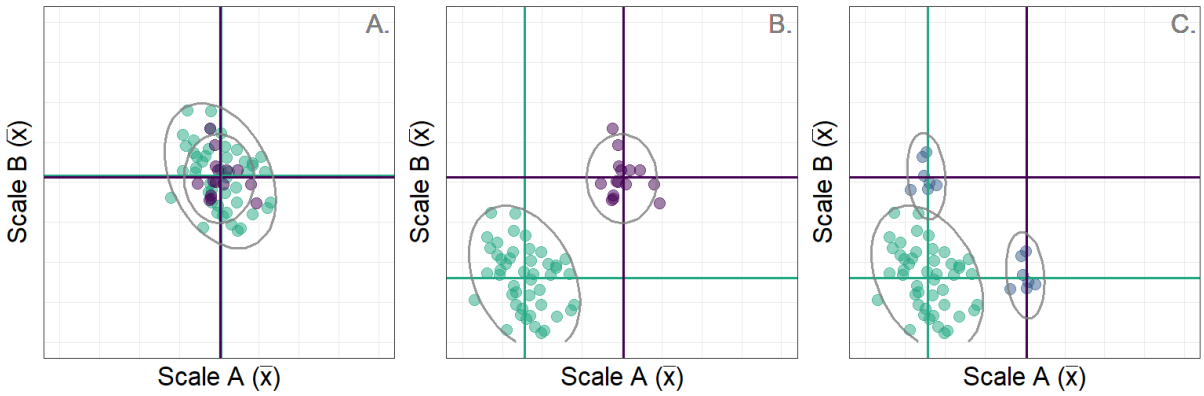
Some scenarios are sketched in Figure 13. When the regular responders respond on average across the midpoint on both scales (i.e., low score separation with the random responder group), the exclusion of random responders will have a mostly ignorable impact on the estimated correlation between scale scores (see Panel A). In contrast, when there is a clear score separation between the regular and the random responder group on both scales, the extent of the overlap between the random responder groups is crucial for the type of impact that will occur upon the exclusion of the random responders. When random responders are consistent across scales, they form a highly influential leverage point pulling any correlation between the two scales, and the removal of random responders from the data can potentially dramatically impact the estimated correlation (see Panel B). In contrast, when the random responder group on the first scale is mutually exclusive to the random responder group on the second scale, random responders would mostly disturb variance estimates (given score separation on both scales between random and regular responders) and thereby upon removal change the estimated correlation between the two scales (see Panel C). Hence, finding a noticeable prevalence and score separation between the random responder group and the regular responder group are not sufficient preconditions for finding inferential impact with respect to correlations, and variances

and correlations within the regular responder group as well as the type of overlap in the random responder groups across the scales will be additional moderating factors. In practice, it can be even more complicated if selection effects are at play and only people with a certain score range on one scale, end up random responding on the other scale. Thus, in the end correlations between scales can be impacted in either direction or remain unaffected, and are all still depending on the non-response mechanisms at play underlying the unobserved true responses for the random responders.

In 'Article 3: Prevalence & Impact' we did not find substantial qualitative differences in a set of correlations (and other descriptive statistics) after excluding random responders in our case study four scales of the TIMSS student questionnaire. This particular result is linked to the relatively low prevalence of random responders, in combination with sizeable variation in scale scores among the regular responder groups, low score separation, and low consistency in random responding across scales. Note that also 'Article 6: How often' points at random responders not being very consistent in being random responder across scales, which excludes the most impactful scenario of Panel B. While these results give reason to be cautiously optimistic with respect to the anticipated impact of in/excluding random responders for inferences at the aggregate level, they should not be taken for granted. Furthermore, as before, we should keep in mind that the full consequences of random responders on the interpretation of scores also depend on the unobserved true scale scores of the random responders. The implicit assumption, made by us and others, that random responders —if they had not responded randomly— would have scored like everyone else, is not per se a natural given. Hopefully, this also makes it clear that it is hard to make general predictions about the impact of random responders on the overall correlation between scale scores and the validity of score interpretations, as it does depend on features of the particular study setting and non-response mechanisms underlying random responding.

Figure 13

Potential Consequences of Random Responders.



Note. The green dots represent the regular responders. The purple dots represent those students who are identified as random responders on both, scale A and scale B. The blue dots represent those students who are identified as a random responder on either scale A or scale B. Depending on the situation, the impact of the random responders on the overall correlation between the scale scores on scale A and scale B differs. In panel A, removing the random responders (i.e., 25% of total) does not influence the correlation (i.e., $r = -.31$ for the whole group and $r = -.32$ for the regular responders). In panel B, removing the random responders (i.e., 25% of total) does decrease and change the direction of the correlation (i.e., $r = .67$ for the whole group and $r = -.32$ for the regular responders). In panel C, removing the random responders (i.e., each random responder group contains 12% of the total) does result in a stronger, yet negative correlation (i.e., $r = .11$ for the whole group and $r = -.32$ for the regular responders).

2.7 Reflections on Random Responders

2.7.1 Prevention of Random Responding

The main goal of any questionnaire is to get valid information from the respondents. Random responders are clearly not providing the desired responses. One question we can ask ourselves is whether prevention is possible? This question, not limited to random responding, has gained a lot of attention within many research contexts.

Incentives. Prior research has for example investigated whether the presence of incentives or increasing the stakes (e.g., academic grades, course credits, feedback, financial reward, or other prizes) could encourage individuals to provide more valid responses and to complete the questionnaire (e.g., Baumert & Demmrich, 2001; Galesic, 2006; Gibson & Bowling, 2020; Huang et al., 2012). Yet, the literature shows mixed results in this respect, making it difficult to identify a generic solution. Furthermore, some solutions

commonly used in an achievement test context might already not be feasible with survey questionnaires. For example, performance-based incentives will be difficult to implement as there are no clear right or wrong answers on a survey question about students' opinions, beliefs, or attitudes. While financial rewards might be ethically questionable and come with no guarantee for achieving their goal (i.e., more truthful answers) (Finn, 2015).

In the TIMSS assessment, a booklet design, as well as pauses in-between assessment parts, are implemented (Mullis & Martin, 2013) such that the test burden and testing time for students are reduced. Yet, to the best of my knowledge, there were no other official procedures in place with TIMSS 2015 to promote valid responding and data quality. Yet, what has been stated is that "In more than half of the cases (61% at the fourth grade and 65% at the eighth grade), school coordinators indicated that students were given special instructions, motivational talks, or incentives by a school official or the classroom teacher prior to testing" (Martin et al., 2016, p.9.14). Galesic and Bosnjak (2009) state that in situations in which students are expected to participate (as is the case for TIMSS), incentives and social pressure typically become more important than any personal interest. As such, it could have been interesting to see if there would be any difference in random responding across the different instruction-incentive-social-pressure conditions in practice. Even part of the between-country differences might be related to how international large-scale assessments are introduced to and regarded by the participating students. That this can differ quite heavily across countries is illustrated by the following quote:

> But I digress. My colleague, Sung-Ho Kim, observed the administration of the IAEP [International Assessment of Educational Progress] tests in Korea. He noted that, although the students chosen to take the test were selected at random, just as in all the other countries, they were not anonymous. No individual scores were obtained, but it was made quite clear that these chosen students were representing the honor of their school and their country in this competition. To be so chosen was perceived as an individual honor, and hence to give less than one's best effort was unthinkable. Contrast this with the performance we would expect from an American student hauled out of gym class to take a tough test that didn't count. (Wainer, 1993, p.13)

Unfortunately, we do not have information on such practices in the publicly available data to relate this to any of our results.

Warnings & Attention Checks. In addition to the provision of positive incentives, there are other approaches around which have a more negative connotation, like the use of warnings or the inclusion of bogus items or instructed response items. Consider, for example, the study by Gibson and Bowling (2020) where the respondents were provided with the following warning: "Please be aware that I will use sophisticated statistical control methods to detect the accuracy and thoughtfulness of your responses. If you do not provide accurate and thoughtful responses to today's survey, you will not receive course credit for completing the survey." (Gibson & Bowling, 2020, Supplementary Material 3; see also Bowling et al., 2021; Huang et al., 2012). In this case, the incentive comes in the form of punishment instead of a reward. The bogus or instructed response items on the other hand have somewhat of a double function. Initially, these types of items are supposed to give an indication if the provided data can be trusted. Yet, at the same time it is deemed good practice to include a warning at the start of the survey to inform respondents that such items can be included. But even without an explicit warning, by their irregular nature, these items might also give the respondents the impression that they are being monitored.

This feeling of being monitored, explicitly through warnings or implicitly by observing bogus items, hopefully encourages respondents to respond less randomly. Yet, the approach can become quite invasive and could actually be counterproductive. The bogus or instructed response items for example, are at risk of being perceived as 'trick' items (Meade & Craig, 2012), deliberately misleading to see if the respondents pay attention, which in addition might also be considered unethical by students and stakeholders (Hooper, 2022). Silber et al. (2022) showed that over 35% of the respondents feel 'controlled' or 'manipulated' by these types of items. These approaches might indeed not always result in more valid responses and easier detection of invalid response behavior. Galesic (2006) for example, has reasons to believe that monitoring respondents, by repeatedly asking them about how they experience the questionnaire (i.e., as an indicator of response quality), might have adverse effects, resulting in respondents not completing the questionnaire. This also complicates matters further, as it has been stated that for the best detection results of invalid responses it would be good to include multiple indicators throughout the questionnaire (Curran, 2016). In addition, Silber et al. (2022) for example also show that even if students are being aware of these indicators, they might actually choose not to respond to them, which also defeats their usefulness.

Despite the mixed results for both the positive and negative incentives in the literature, I do want to believe that it will be possible to reduce invalid responding in some way or the other, although I also realize that we won't be able to ever completely eliminate it. At the same time, we can also wonder whether it is really necessary to use countermeasures and monitor behavior of the respondents. Credé (2010, p.602) states that "the rate of random responding is nonzero for most populations and is likely to fall somewhere between 1% and 10%, although higher rates are certainly possible under certain circumstances." If we take this into consideration, general prevalence rates that might not be too bad, maybe the best thing we can do is make researchers aware of the different factors potentially influencing data quality and address them after data collection.

One example could be to follow our adopted approach in 'Article 3: Prevalence & Impact', where we assigned respondents into different groups based on their response be-

havior and performed sensitivity checks for the robustness of the results. Alternatively, Goldammer et al. (2020) also summarizes some other routes one could consider, such as incorporating levels of invalid responding as an additional covariate or method factor in the model or treating invalid responses as a missing data problem. But first, it is important to bring sufficient attention to invalid response behavior and its potential consequences. For some time there have been indications that, in certain research areas, these are underexposed topics that don't get enough credibility. As a consequence, standard procedures for the detection of invalid responses are often lacking (e.g., Hooper, 2022; Liu et al., 2013; McGonagle et al., 2016). In the end, regardless of the specific route chosen, all efforts to explore the presence of random responding (and other forms of invalid response behavior) (e.g., Cronbach, 1950; Curran, 2016; Huang et al., 2012) will hopefully result in increased awareness and a better understanding of the large amounts of data that are available for international large-scale educational assessment and the like.

2.7.2 Generalization of Results

The overall average prevalence rates (across different countries, scales and/or grades) among the different studies in this thesis ranged between 6% and 10%. At first sight, this is well in line with the general trends found among other studies. Of course, under specific conditions higher rates can occur, yet in the broader literature, common estimates for typical cases are often around 10% (e.g., Credé, 2010; Curran, 2016). Based on this there are no strong reasons to believe that these numbers wouldn't extend to other international large-scale assessments as well. Let's for example consider PISA (Programme for International Student Assessment from OECD) and PIRLS (Progress in International Reading Literacy Study from IEA). The PISA assessment is conducted among 15-year-old students, while PIRLS is conducted among fourth-graders. Most of our results are based on the eighth-grade students that participated in the TIMSS 2015 assessment, with some results present for fourth-grade students as well. What we saw in 'Article 5: Who' is that students in the higher grades have a higher odds of being identified as random responders. With that respect, it would be logical to assume that the results might be

more easily generalizable to results in PISA.

On the other hand, the PISA assessment is said to have a different focus, while the student questionnaire for PIRLS 2016 has the same setup as the TIMSS 2015 fourth-grade student questionnaire. It even has the same type of questionnaire scales and some similar items, although focused on reading instead of mathematics or science. Yet, as we saw in 'Article 3: Prevalence & Impact' and 'Article 6: How often', students don't seem to be very consistent in their random response behavior across the questionnaire. This makes us believe that random responding is likely not a generally applicable trait and that scale characteristics matter. We saw for example in 'Article 4: Where' that scale content seemed to be rather important, with the confidence scales showing higher prevalence rates of random responders than all the other scales. As there is likely less overlap in questionnaire scales and item wording between TIMSS and PISA, it would be logical to assume that the findings generalize less well to PISA than to PIRLS, and other cycles (assuming questionnaire content and formulation don't largely differ and that there are no huge cohort-differences). Alternatively, it could also be that the items that are different in PIRLS and the fact that they relate to reading, are also the specific cases that could actually make a difference. Overall, it seems wise to be cautious when considering the content of the questionnaire. Yet, in order to make more definitive statements about how and to what degree the different assessments might be impacted, it would probably be good to first figure out what the specific content or wording aspects are that would make a difference and continue from there.

In the end, it is complicated to say something about the actual differences that could be expected across the different assessments. While the general view does not directly give any reasons to worry about extreme random responding, if we want to be sure, we have to check it. There can always be that special case that stands out.

References

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

ASA. (2018). Ethical guidelines for statistical practice. Prepared by the committee on professional ethics of the american statistical association. www.amstat.org/

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI–A. *Journal of Personality Assessment*, *68*(1), 139–151.

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

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Widiger, T. A., Sumpter, J. C., Reynolds, S. K., & Hallam, R. A. (1991). Detection of random responding on the MMPI-2: Utility of F, back F, and VRIN scales. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(3), 418–423.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(2), 718–738.

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*(3), 279–304.

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.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*(3), 311–326.

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.

European Commission. (2021). Ethics and data protection. https://ec.europa.eu/

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *RR-15-19*.

Foy, P. (2017). *TIMSS 2015 User Guide for the International Database*. TIMSS & PIRLS International Study Center, Boston College.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313–328.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, *36*(2), 410–420.

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), Article 101384.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *The Public Opinion Quarterly*, *72*(2), 167–189.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, *7*(1), 1–17.

Hedlin, D. (2020). Is there a 'safe area' where the nonresponse rate has only a modest effect on bias despite non-ignorable nonresponse? *International Statistical Review*, *88*(3), 642–657.

Hernández-Torrano, D., & Courtney, M. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *PloS ONE*, *9*, Article 17.

Hooper, M. (2022). Dilemmas in developing context questionnaires for international large-scale assessments. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education: Perspectives, methods and findings* (pp. 721–747). Springer International Publishing.

Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal*, *27*(3), 406–422.

Hopfenbeck, T. N., Lenkeit, J., Masri, Y. E., 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.

Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing*, *11*(2), 95–121.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods*, *21*(1), 197–225.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, *27*(4), 307–327.

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A., & King, K. M. (2018). Detecting random responders with infrequency scales using an error-balancing threshold. *Behavior Research Methods*, *50*(5), 1960–1970.

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229–248.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*(4), 269–290.

Liu, M., Bowling, N., Huang, J., & Kent, T. (2013). Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *The Industrial-Organizational Psychologist*, *51*, 32–38.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College.

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology*, *65*(2), 287–321.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

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

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., & Loveless, T. (2016). *20 Years of TIMSS: International Trends in Mathematics and Science Achievement, Curriculum, and Instruction*. TIMSS & PIRLS International Study Center, Boston College.

NESH. (2016). Guidelines for research ethics in the social sciences, humanities, law and theology. www.etikkom.no

Payne, K. [@bkeithpayne]. (2022, January). *What if the low effort, random responding we see on internet surveys is not an aberration, but a good reflection of how people really are? Maybe most people, most of the time, are just as shallow-thinking and random-responding in daily life. It would explain some things, no?* [Tweet]. Twitter. https://twitter.com/bkeithpayne/status/1483827501210165252

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55–79.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, *42*(3), 411–430.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191.

Silber, H., Roßmann, J., & Gummer, T. (2022). The issue of noncompliance in attention check questions: False positives in instructed response items. *Field Methods*, *34*(4), 346–360.

Taylor, C. S. (2013). *Validity and validation*. Oxford University Press.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education*, *58*(3), 129–151.

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501.

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution rasch models: Extensions and applications*. Springer.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*(1), 1–21.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81.

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.

Yamamoto, K. (1989). Hybrid model of IRT and latent class models. *ETS Research Report Series*, *RR-89-41*.

3 Further Reflections on the Comparative Use of Models

In both applications, a selected null baseline model takes the center role in the comparison process. In the different applications of this null baseline model, its meaning is established in terms of a relative comparison to a measurement model of interest. By doing this, we hope to learn from the model comparison, either in terms of model fit improvement or in terms of measurement appropriateness. In both cases, the choice of baseline – for model comparison or for the mixture component – is crucial for inferences that follow their application.

3.1 Model Fit: Variable-based and Person-based

For Application 1, we explicitly looked at model fit by means of incremental fit indices. While it might not be directly apparent, the random responders in Application 2 could potentially also serve as a measure of model fit. For both applications, it can be said that the selected baseline model reflects some noise pattern and it is expected that any proper measurement model should be able to show better performance if the provided responses are in line with the measurement model. While both applications can give some indication about model fit, the underlying approach is different and model fit will be quantified differently.

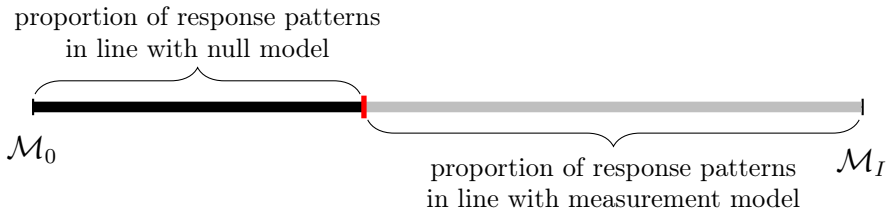
Application 1 addressed model fit from a variable-based perspective. Within this perspective, model fit is quantified by determining to what extent the observed relations between the variables are better explained by a measurement model than a null baseline model. In contrast, Application 2 approached model fit from a more person-based perspective. From this perspective, the focus is more on 'How many people in the study behaved or responded in a manner consistent with theoretical expectation?' (Grice et al., 2020, p.444). Where the percentage of people that (do not) fulfill this criteria can be seen as a person-centered quality measure (Grice et al., 2020). More specifically, for Application 2, we assumed there to be two different groups of students following different models describing the relation between their item responses. Building on the idea of a person-centered approach, one could say that the prevalence of random responders can

serve as a person-centered measure of badness-of-fit. In this case, model fit would reflect the 'quality' of the measurement model by indicating the percentage of students for which the response patterns are more in line with the measurement model than with the null baseline model.

Visually, as for CFI, this relation can also be depicted by means of a continuum, yet now ranging from the null baseline model to the measurement model of interest (see Figure 14). The continuum itself represents 100% of the response patterns. In this case, the prevalence estimate serves as an indicator dividing the continuum into two parts: the proportion of response patterns in line with a null model (left side) and the proportion of response patterns in line with a measurement model (right side). If the measurement model would be able to describe the response patterns of all students, the indicator would be located on the far-left side and the bar would be completely gray. Yet, as the indicator moves more to the right side, a larger part of the sample does not fit with the measurement model of interest, indicating worse fit with respect to the measurement model.

Figure 14

Person-Centered Approach to Model Fit.



Note. The red line represents the estimated prevalence of random responders.

In both applications, we could also identify different characteristics influencing model fit. Yet as the applications address model fit from a different viewpoint, the characteristics are also different in nature. In Application 1, the features that influence fit are based on more factual data characteristics such as the number of items, sample size, and the degree of multivariate dependence. Whereas in Application 2, we saw that the features that influence fit are based on person and assessment characteristics such as gender, grade, SES, and scale position.

3.2 Compared to What?[11]

3.2.1 Multiple Comparison Grounds of Interest: Accuracy versus Usefulness

As mentioned before, in both applications we considered the null baseline model as the standard for making comparisons. Yet in practice, many other comparisons could have been made instead. For example, as we shortly pointed out in Application 1, when using incremental fit indices the comparison being made does not necessarily have to involve this specific null baseline model. Sobel and Bohrnstedt (1985) for example believe that the baseline should be one that incorporates the current state of knowledge and theory within a given domain. At the same time, there is also the idea that the baseline should not be replaced, but alternatives should be added to model comparison practices instead (e.g., Marsh, 1998; Rigdon, 1998). The underlying idea is that more information will be available if a model of interest is compared to a set of alternative models, instead of just a single one. Building on this idea, one might want to consider comparing a series of models. For example, consider not only comparing the measurement model against the original and alternative baselines, but also comparing the different alternative models with each other, or comparing different substantive models like the model testing strategy described in Application 1. In the end, this would lead to a whole range of relative fit values.

Similarly, for Application 2 we used the null baseline model for representing an aberrant random response pattern. However, the concept of invalid response behavior is of course not limited to random responding and can be viewed as a much broader umbrella term including many different response styles, sets, or tendencies (e.g., Cronbach, 1950; McGrath et al., 2010; Messick, 1991). This also implies that we could have made many different modifications to the mixture model we worked with. We could for example have considered changing the baseline model, simultaneously including more than two groups of responders to take into account other possible response sets, or put other restrictions on the measurement model. Each mixture component and comparison would have provided

[11] Part of this section is based on a course paper written for 'UV9002: Philosophy of Science' at the faculty of Educational Sciences, University of Oslo.

us with different information.

At the same time, we also need to be aware that we can be at risk of taking the comparison process too far by differentiating between too many alternative models or too many groups of responders. For example, taking this to the extreme in the context of the mixture model, we could end up with a separate class for each person or alternatively, each response pattern could potentially be linked to a specific response style. Yet, what can then still be considered genuine or valid responses? In theory, a mixture model with that many classes will probably be perfectly accurate, yet in practice, the usability gets lost as the model no longer provides a useful summary. This point can also be related to a part of the story 'Sylvie and Bruno concluded' (see the excerpt below).

> Mein Herr looked so thoroughly bewildered that I thought it best to change the subject. 'What a useful thing a pocket-map is!' I remarked.
> 'That's another thing we've learned from *your* Nation,' said Mein Herr, 'map-making. But we've carried it much further than you. What do you consider the *largest* map that would be really useful?'
> 'About *six inches* to the mile.'
> 'Only six inches!' exclaimed Mein Herr. 'We very soon got to six *yards* to the mile. Then we tried a *hundred* yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a *mile to the mile*!'
> 'Have you used it much?' I enquired.
> 'It has never been spread out, yet,' said Mein Herr: 'the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.' (Carroll, 1983, p.403)

What this example shows me is that we will end up with an unworkable model if we take the comparison to the extremes and expect 1-on-1 similarity between all possible elements in the real world and the model. At the same time, when focusing on creating the

most accurate or perfect-fitting model, we should keep asking ourselves what the model actually symbolizes. In general, as capturing the phenomenon under study becomes more important than the phenomenon itself (Schouten, 1992) we are at risk of losing the meaning that can be attributed to the modeling process. Any perfect model that cannot be used, will not provide any new information about the world. In the end, it seems to come down to a trade-off between accuracy and usability when working with models to evaluate complex systems or phenomena.

3.2.2 Using Models: Attainability of Truth versus Progress

When comparing two models, the relative evidence for a model A over a model B, does not necessarily represent evidence for model A on its own (Royall, 1997). Relating this to model fit evaluation with CFI, the measurement model might show better overall fit than the null baseline model, yet this does not imply that the measurement model itself is a good model or even correct. In theory, there could still be non-ignorable misspecification even with 'good fit' (although chances get smaller with better CFI values). Similarly, in the context of random responding, the mixture model only indicates whose response patterns are more consistent with a null baseline model than the measurement model, by comparing the two options. It does not mean that students were consciously responding randomly or that there are no other ways to characterize the distinct response patterns.

In an ideal situation, one might want the model to be the true or the perfect representation of the phenomenon under study, yet in practice this is not attainable. In general, it seems to be acknowledged that by nature models are always approximations of the phenomena under study or to put it differently, they are always wrong to some degree (e.g., MacCallum & Austin, 2000; MacCallum, 2003; McDonald & Ho, 2002). Models can be defined as providing a "specification of a theory" (Hélie, 2006, p.1) by operationalization of a phenomenon under study. Yet, due to the complex nature of the real world, models will not be able to capture this completely as they are a simplified representation of the phenomenon. Models not being 100% true does however not mean that all models are useless nor that we cannot learn from them.

According to Morrison and Morgan (1999), models can be seen as instruments that

are used to incorporate and connect knowledge about theories and the world. One way of learning through models actually occurs by using them, for example by applying the model to data and making estimations (Giere, 2004; Morrison & Morgan, 1999) or by making adjustments to the model (Morrison & Morgan, 1999). Thus, the best we can do in order to make progress is to build our models around the theories that currently have the largest evidence base, while acknowledging there might still be some inconsistencies that we are not yet aware of. At least this will provide us with a functional working model and the opportunity to use and compare the model against a meaningful baseline, observe consequences, make conclusions, and continue learning.

This might also be related to the original idea of adopting a model-testing strategy when using incremental fit indices for model evaluation. CFI and other incremental fit indices provide a 1-number summary for describing the level of agreement between the model and the data. These summary values are used to provide a qualitative value judgement about a model and are generally expected to conform to some threshold for indicating 'good' model fit. Yet this is probably not how science should work and does not say anything about where the model could potentially be improved. Thus, instead of '[evaluating] a single model in isolation, it is often more informative and productive to compare a set of alternative models and possibly to select a preferred model from the set.' (MacCallum, 2003, p.130). At the same time, we need to remain open to make adaptations to our models as new knowledge becomes available and practice changes. But in the end, it is this whole process of learning through models that will enable us to say something about different phenomena in the world.

References

Carroll, L. (1983). Sylvie and bruno concluded. *The Complete Novels of Lewis Carroll: With All the Original Illustrations + The Life and Letters of Lewis Carroll* (pp. 327–490). e-artnow: eBook.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *71*(5), 742–752.

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455.

Hélie, S. (2006). An introduction to model selection: Tools and algorithms. *Tutorials in Quantitative Methods for Psychology*, *2*(1), 1–10.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201–226.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139.

Marsh, H. W. (1998). The equal correlation baseline model: Comment and constructive alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 78–86.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470.

Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach* (pp. 161–200). Lawrence Erlbaum.

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. S. Morgan & M. Morrison (Eds.), *Models as Mediators: Perspectives on Natural and Social Science* (pp. 10–37). Cambridge University Press.

Rigdon, E. E. (1998). The equal correlation baseline model: A reply to Marsh. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 87–94.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Chapman & Hall.

Schouten, F. (1992). The paradox of the map: Semiotics and museum education. *Museum Management and Curatorship*, *11*(3), 285–289.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. *Sociological Methodology*, *15*, 152–178.

4 Article 1: Metric Space

van Laar, S., & Braeken, J. (2021). Understanding the comparative fit index: It's all about the base! *Practical Assessment, Research, and Evaluation*, *26*, Article 26. https://doi.org/10.7275/23663996

Understanding the Comparative Fit Index: It's all about the base!

Despite the sensitivity of fit indices to various model and data characteristics in structural equation modeling, these fit indices are used in a rigid binary fashion as a mere rule of thumb threshold value in a search for model adequacy. Here, we address the behavior and interpretation of the popular Comparative Fit Index (CFI) by stressing that its metric for model assessment is the amount of misspecification in a baseline model and by further decomposition into its fundamental components: sample size, number of variables and the degree of multivariate dependence in the data. Simulation results show how these components influence the performance of CFI and its rule of thumb in practice. We discuss the usefulness of additional qualifications when applying the CFI rule of thumb and potential adjustments to its threshold value as a function of data characteristics. In conclusion, we at a minimum recommend a dual reporting strategy to provide the necessary context and base for meaningful interpretation and even more optimal, a move to using CFI as a real incremental fit index intended to evaluate the relative effect size of cumulative theoretically motivated model restrictions in terms of % reduction in misspecification as measured by the baseline model.

The evaluation of model fit remains a crucial yet controversial topic in the application of structural equation models. In line with concerns that a focus on mere statistical significance testing would lead to disregarding or changing relevant and theoretical sound models without proper justification for it (Bentler & Bonett, 1980), a whole range of alternative goodness-of-fit indices is currently available for model evaluation beyond the traditional chisquare significance test of exact fit. As part of the general trend to report multiple fit indices (e.g., Jackson et al., 2009; Ropovik, 2015), McDonald and Ho (2002) point out that "it is sometimes suggested that we should report a large number of these indices, apparently because we do not know how to use any of them" (p. 72). This

statement highlights a common concern about current model evaluation practices that are characterized as thoughtless routine applications of binary (good/bad) rules of thumb for fit indices.

Different cut-off criteria or rules of thumb have been proposed over time (e.g., Bentler & Bonett, 1980; Hu & Bentler, 1999; Schermelleh-Engel et al., 2003). In particular, Hu and Bentler's (1999) suggested criteria gained huge popularity. Yet, Hu and Bentler (1999) themselves stressed that "it is difficult to designate a specific cutoff value for each fit index because it does not equally well with various conditions" (p. 27). Their underlying simulation study was based on only a few conditions with either a simple or a complex structure with fixed values for a three-factor confirmatory factor analysis model with 15 manifest variables. Their note of caution resonates well with more recent findings in the literature where simulation studies have illustrated the sensitivity of fit indices and their rules of thumb to various data and model features such as sample size, model size and type, strength of relations within the measurement model, and violations of distributional assumptions (for a review, see e.g., Niemand & Mai, 2018). Nevertheless, people have been universally applying the rules of thumb regardless of their own specific context, study design, data, or model. The main point of concern is exactly this thoughtless default way of applying rules of thumb (Marsh et al., 2004). One reason given for abiding by such a thoughtless rule-based approach is that "researchers need them because it is unclear how one can reach qualitative judgements in their absence" (Lai & Green, 2016, p. 221).

Overall, one major point of concern with respect to the application of SEM in practice is the lack of deliberate decision making in all parts of the process (McDonald & Ho, 2002). In order to make more informed decisions with respect to the use of fit indices it is important to know how these fit indices work. Yet what 'good' fit means and how fit indices map onto this meaning is not well understood (Lai & Green, 2016). Hence, if we would desire not mere mindless rule-following but more deliberate practice when assessing model fit, we need to better clarify what type of fit each of the different indices stand for and to provide a better insight in their inner workings to understand why fit

indices behave like they do.

Here, we will try to make one step into that direction by focusing on the Comparative Fit Index (CFI) (Bentler, 1990), the most-used statistic among the class of comparative goodnes-of-fit indices (for reviews covering time periods in the interval 1995-2013, see e.g., Jackson et al., 2009; McDonald & Ho, 2002; Ropovik, 2015). A decomposition in the main components that play a role in the CFI's baseline comparison allows to clarify CFI's meaning and behavior, explain some of the mixed results in the SEM simulation literature regarding its sensitivity to model and data characteristics, and highlight the (limited) generalizability of common rules of thumb for CFI and factor analysis. We hope that this exposition can help guide the decision-making process in practice and lead to smarter, more deliberate inferences when interpreting the CFI for model fit evaluation.

A decomposition of the Comparative Fit Index

In contrast to absolute fit or parsimony fit indices (e.g., Brown, 2015), the class of comparative fit indices promotes comparison in fit between a model of interest and a more restricted baseline model. This fit assessment strategy has its foundation with Bentler and Bonett (1980) and involves a continuum of models from the worst fitting null model to the perfect fitting or saturated model. The role of the comparative fit indices is to assess where the model of interest is located within this continuum.

Within this class, Bentler's (1990) Comparative Fit Index (CFI) is an "index to summarize the relative reduction in noncentrality parameter of two nested models" (p. 238). The noncentrality parameter λ_m of a model m can be seen as an indicator of model misspecification as it quantifies the amount of deviation between the estimated χ^2 value and the expected χ^2 value (i.e., df_m, the model's degree of freedom) for the sample under the assumption that the model is correct: $\lambda_m = \chi^2_m - \text{df}_m$. The value of CFI is then based on the ratio of misspecification of both models:

$$\text{CFI}_{(m,b)} = 1 - \frac{\lambda_m}{\lambda_b} = 1 - \frac{\chi^2_m - \text{df}_m}{\chi^2_b - \text{df}_b} \tag{1}$$

where the subscript indicates whether the statistics are of the model of interest m or the

baseline model b. The one-minus-noncentrality-ratio is there to turn it from a relative misspecification measure into a relative goodness-of-fit measure. Note that the CFI is usually truncated to the $[0, 1]$ interval, although technically values higher than one can arise if the model of interest fits better in a noncentrality sense than the saturated model (e.g., perfect fit with less then full parameters) and values below zero can arise if the model of interest fits worse than the baseline model.

Null baseline. A so-called null model in which all observed variables are uncorrelated has taken off as the default baseline model for popular applications of CFI. Following the idea of Bentler and Bonett (1980), the $\text{CFI}_{(m,0)}$ can be referred to as an 'index of information gained' by the model of interest over the more restrictive null model. Hence, conceptually it is similar to an R-square, a relative reduction in 'unexplained' variance, whereas a $\text{CFI}_{(m,0)}$ could then be seen as a relative reduction in 'unexplained' variance-covariance. From here on we will drop the subscripts referring to the models being compared, if we talk about the CFI with the null model as default baseline.

Rules of thumb. For determining whether a model shows adequate fit according to the CFI, different rules of thumb have been proposed. Early on up to the late 90's, values of at least .90 for comparative fit indices were assumed to indicate decent model fit (for a review, see McDonald & Ho, 2002). This rule of thumb has been mostly motivated based on experience by expert users: At CFI origins, "In our experience, models with overall fit indices of less than .90 can usually be improved substantially" (Bentler & Bonett, 1980, p. 600) or more recently, "In my experience, models with .90+ values for the CFI ... can be quite acceptable models" (Little, 2013, p. 116). The currently most common CFI standard is based on the influential simulation study by Hu and Bentler (1999): "the results suggest that, for the ML method, a cutoff value close to .95 for ... CFI ... are needed before we can conclude that there is a relatively good fit between the hypothesized model and the observed data" (p. 1). As indicated earlier in the introduction, even about the core rule of thumb, stating CFI $\geq$ .95 for good model fit, there have been many cautionary notes and simulation studies have illustrated that its applicability varies depending on data and model characteristics.

If we would desire more deliberate practice when assessing model fit using CFI values, then knowing the inner workings of this measure is an essential requirement. So how does this CFI really work? Additionally, can knowledge of its inner workings indeed shed some light on the performance of the CFI rules of thumb under various data characteristics?

CFI as a relative measure with a variable metric space

Equation 1 clarifies that the CFI is a relative measure with its denominator set by the noncentrality of the baseline model. Now suppose there is a line that represents the CFI metric. The metric space endpoints are set by the null and saturated model. The length of the line is determined by the noncentrality of the null model, as the noncentrality for the saturated model is zero. Given the formulation of CFI, this metric space serves as standard for comparison. Conceptually, the length of the line, the CFI metric space, has an influence on the behavior of CFI. Having more space, will allow for a finer grained differentiation. Having less space, makes the CFI to become less useful. The rationale is that in general it is harder to differentiate between models as they are becoming more similar. When placing a model of interest in the metric space, it will always be closer related to both the null and the saturated model as the line becomes shorter. As a consequence, a comparison in terms of CFI values is no longer based on the same standard when the denominator, the baseline noncentrality, is different among the cases being compared.

As an example to drive this idea home, consider the following two cases for which the size of the CFI metric space is different. The baseline noncentrality in the first case is $\lambda_0 = 25$. Within this space two models with slightly different noncentrality values can be placed. Overall their values only differ by 2 units, with $\lambda_1 = 1$ and $\lambda_2 = 3$ being the noncentrality value of the first and second model, respectively. Translating this to CFI values, this results in values of $\text{CFI}_{(1,0)} = .96$ and $\text{CFI}_{(2,0)} = .88$. Now consider the second case in which there is a shorter metric space with baseline noncentrality $\lambda_0 = 5$. Here as well, we have two models that only differ by 2 noncentrality units, now with $\lambda_1 = .2$ and $\lambda_2 = 2.2$. However, translating this to CFI interval, values of $\text{CFI}_{(1,0)} = .96$ and $\text{CFI}_{(2,0)} = .56$ are obtained. This example demonstrates the impact of widely differing

metric spaces as defined by the baseline noncentrality. The difference in CFI-fit between the two models is huge between the two cases whereas the difference in terms of absolute misspecification as expressed by the noncentrality index is exactly the same. Sampling variability can also be expected to have a huge impact in the second case, a small difference in noncentrality value can lead to widely differing CFI values when baseline noncentrality is small. Thus, the main conclusion is that we cannot interpret a CFI-value of a model or differences in CFI between models without considering the fit of the CFI baseline model for the same sample data. This is similar advice as with any ratio or risk measure, you cannot ignore the numerator and denominator when interpreting a percent; Or more colloquially speaking, whereas a small percent of everything is a lot, a large percent of nothing, is still nothing.

Null model baseline noncentrality as key factor

For the default CFI with a null model as baseline, the null model noncentrality λ_0 is the key to CFI behavior and interpretation as it sets the metric space that serves as standard for comparison. With F being the ML discrepancy fit function (e.g., Bollen, 1989) between the observed and null-model-implied covariance matrices $\mathbf{S}$ and $\hat{\mathbf{\Sigma}}_0$, the null model noncentrality can be rewritten and simplified as follows to identify its key components:

$$\begin{aligned} \lambda_0 &= \max\left(\chi_0^2 - \mathrm{df}_0, 0\right) \\ &= \max\left(F(\mathbf{S}, \hat{\mathbf{\Sigma}}_0)(n-1) - \mathrm{df}_0, 0\right) \\ &= \max\left(-\log|\mathbf{R}|(n-1) - p(p-1)/2, 0\right) \end{aligned} \tag{2}$$

where $\mathbf{R}$ is the observed correlation matrix, n the sample size, and p the number of manifest variables (for the derivation, see Appendix A).

Equation 2 clarifies that the CFI metric space is a function of correlation (i.e., generalized variance as expressed by the determinant of the data correlation matrix), sample size, and number of variables. Notice that all three core components of the null model baseline noncentrality are completely data dependent. In an ideal situation with a lot of correlation in your data, large sample sizes and not too many variables, CFI would allow you to make a fine-grained differentiation between models in terms of relative non-

centrality. These ideal conditions are quite in line with common sense guidelines for the application of SEM. There are some more general intuitions that can be derived a priori from this decomposition that can be linked to findings in the SEM model fit literature.

Sample size n. Originally, comparative fit indices were conceptualized as 'indices of information gained' and should be independent of sample size (Bentler & Bonett, 1980). However, previous studies (e.g., Heene et al., 2011; Hu & Bentler, 1999; Marsh et al., 2004; Shi et al., 2019) as well as the decomposition show that CFI is clearly dependent on sample size. In this case, with higher sample sizes resulting in higher baseline noncentrality values and better expected performance.

Number of variables p. In the literature (e.g., Shi et al., 2019) a general trend has been reported that more variables complicate the use of CFI and its default rule of thumb. At first sight the decomposition supports this notion as more variables leads to lower baseline noncentrality making model differentiation more difficult. However there is a confounding factor that is easily forgotten, the determinant $|\mathbf{R}|$ is also a function of the number of variables p, and with more variables more non-zero correlations can in principle occur in the correlation matrix $\mathbf{R}$. Hence, the number of variables only has a clear negative effect on CFI if $p(p-1)/2$ the degrees of freedom of the null model outweighs the contribution by $-\log|\mathbf{R}|(n-1)$.

In the extreme theoretical situation in which only additional uncorrelated variables are added this will be always the case, as this has no impact on the latter factor. Yet the more correlation the added variables contribute the faster the negative effect of the number of variables disappears (i.e., the logdeterminant factor increases nonlinearly). Hence, it should thus not be surprising that Shi et al. (2019) found that, for correctly specified models, the effect of p on performance of CFI's rule of thumb was dependent on the size of the factor loadings they used. Hence, CFI also follows the general principle that having more signal in the data facilitates matters, whereas adding more noise further confounds matters.

Data correlation $\mathbf{R}$. As already indicated in the previous paragraph, the more the data is unlike the null model, the higher the baseline noncentrality and the easier CFI

can differentiate between models. The study by Heene et al. (2011) also showed that performance of CFI's rule of thumb is dependent on used factor loadings. It should also not be surprising that performance issues became more severe as the sample size decreased (Heene et al., 2011), as there is a synergistic interaction between n and $-\log|\mathbf{R}|$ as reflected by the prominent role of their product in the decomposition. Given the formulation, a decrease in both components will provide the smallest metric space, providing worse conditions for model differentiation.

Now that we have identified the core components that play an integral part in the baseline comparison for CFI we will first zoom in further on CFI in relation to different data characteristics, by assessing the impact of sampling variability on the proposed metric space principle and the extent to which this relates to the general applicability of the common rule of thumb for CFI. Secondly, we will follow up on an additional qualification on when the general CFI rule of thumb can be used. We end the paper with a more general discussion on implications of these results and with recommendations for the use of CFI and its common rule of thumb in practice.

Sampling variability & CFI

At population level, CFI is determined by the population model noncentrality $\lambda_m^{(\Sigma)}$ and the population null baseline noncentrality $\lambda_0^{(\Sigma)}$. When the estimated model is the true population model, $\lambda_m^{(\Sigma)}$ shows perfect fit ($\lambda_m^{(\Sigma)} = 0$) and consequently the population CFI will always equal one. This means there is only systematic variation in $\lambda_0^{(\Sigma)}$, caused by variation in the components that make up the CFI metric space. Even though this does not have a direct influence on the CFI value at population level, it will set the basis for sample performance of CFI: a larger null baseline noncentrality $\lambda_0^{(\Sigma)}$ provides a more solid basis for model differentiation. In practice, the two noncentralities at sample level $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ will be prone to sampling variability and potentially also sample bias. Depending on the extent that both noncentralities are somewhat differently affected, this could lead to differences in results compared to our expectations.

Monte Carlo Simulation Design

We considered a simple one-factor data-generating population model with equal factor loadings implying equal correlations between all items. The focus was on the use of correctly specified models, as it seems that the goal of most people is not to falsify their model, but to find an adequate model as starting point for further analysis (e.g., Ropovik, 2015). Given this focus on adequate model fit, it would be good to know whether CFI's rule of thumb can meet its purpose in the ideal case of a correctly specified model.

Experimental Factors. The conditions studied are related to the three components of the baseline noncentrality provided by the decomposition of CFI: sample size n, number of variables p, and data correlation $\mathbf{R}$.

First, sample size is varied ($n \in \{100, 200, 500, 1000\}$). More information is present with increasing sample size, such that there is less uncertainty in making inferences about model fit. Minimum sample size requirements around 150-200 have been proposed for SEM (e.g., Barrett, 2007; Boomsma, 1985; Kenny, 2015; Muthén & Muthén, 2002), yet in practice about 1 in 5 studies uses sample sizes below 200 (MacCallum & Austin, 2000) and around 8-18% uses sample sizes below 100 (Jackson et al., 2009).

Second, the number of variables is varied ($p \in \{4, 8, 12, 24\}$), as previous research has shown that the number of variables does have an influence on model evaluation (e.g., Moshagen, 2012; Shi et al., 2019; Shi et al., 2018).

Third, the degree of data correlation as expressed by $|\mathbf{R}|$ is varied through the chosen data-generating population model. The use of the one factor homogeneous factor loading model as population model allows to make this determinant a direct function of one correlation number r, where $|\mathbf{R}| = [1 + (p-1)r][1-r]^{(p-1)}$ (e.g., Graybill, 1983) with $r \in \{.1, .2, .3, .5, .7, .9\}$. According to Brown (2015), in practice standardized factor loadings of at least .3 or .4 are considered the norm for a meaningful interpretation, which corresponds in our simulation setup to values of $r = .09$ and $r = .16$, respectively. Hair et al. (2006) are stricter and require factor loadings to be above .5 or even .7 in the context of validation studies, which corresponds to values of $r = .25$ and $r = .49$.

Experimental Design. These three experimental factors are combined into a full

factorial simulation design leading to $n(4) \times p(4) \times r(6) = 96$ experimental conditions. Within each condition, 1000 sample covariance matrices $\mathbf{S}$ were drawn from a Wishart distribution, $\mathbf{S} \sim \mathrm{W}(\mathbf{\Sigma}, \mathrm{df})$, where $\mathbf{\Sigma}$ is the model's population covariance matrix and df the model's degrees of freedom. The model was then refitted to each of the generated samples. The simulation and analyses were conducted in R (R Core Team, 2020) through custom scripts in combination with the lavaan package for R (Rosseel, 2012).

Outcome measures. For each sample, the sample non-centrality of the baseline model and of the fitted model – being the numerator and denominator of the CFI, respectively – are computed. The CFI of the fitted model is assessed and used to decide whether or not the fitted model is judged to be of good fit according to the .95 rule of thumb (i.e., CFI<.95 leads to rejection of the model).

Monte Carlo Simulation Results

Full results of the 96 experimental conditions of the Monte Carlo simulation study are reported in table-format in Appendix B. In what follows, we will report on general trends for the respective outcome measures and zoom into specific conditions when relevant.

Null baseline noncentrality $\lambda_0^{(S)}$***.*** Given that noncentrality parameters are shifted-versions of the chisquare statistic (i.e., $\lambda_0 = \chi_0^2 - \mathrm{df}_0$), the same sampling distributions would apply under asymptotical theory given regularity conditions (e.g., Steiger et al., 1985), implying a central or noncentral chisquare distribution depending on whether or not the model is correctly specified. Yet note that for the null baseline model it has been found that a noncentral chisquare distribution does not properly describe its sampling distribution beyond its central tendency (Curran et al., 2002). However, the sample null baseline noncentrality does follow nicely the population trends (see Table 1) that are function of the earlier identified three components of the metric space. Where an increase in either of the components has a positive effect on the baseline noncentrality. Notice that the sampling variation unaccounted for by the design factors is almost non-existing (i.e., $1 - \eta^2_{total} = .001$).

Table 1

Eta square (η^2) effect size patterns for the main components of the CFI metric space across different outcome measures in the main simulation study.

	η^2				
term	$\lambda_0^{(\Sigma)}$	$\lambda_0^{(S)}$	$\lambda_m^{(S)}$	CFI	$< .95$
p	.124	.134	.276	.010	.025
r	.268	.265	.000	.129	.364
n	.146	.144	.033	.076	.212
$p \times r$	.135	.134	.000	.011	.028
$p \times n$	.081	.080	.075	.022	.068
$r \times n$	.163	.160	.000	.081	.227
$p \times r \times n$	.082	.081	.000	.029	.076
total	1	.999	.384	.358	1

Note. $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality; $\lambda_m^{(S)}$ = sample noncentrality for the estimated true model; CFI = sample CFI value for the estimated true model (i.e., $CFI = 1 - \lambda_m^{(S)}/\lambda_0^{(S)}$); $< .95$ = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95. η^2's are based on the type-III sum of squares in a full factorial ANOVA.

Comparing the theoretically expected $\lambda_0^{(\Sigma)}$ with the sample average $\overline{\lambda}_0^{(S)}$ (see Table B1) indicates that a small upward sampling bias for $\overline{\lambda}_0^{(S)}$ is present. This bias tends to become more severe with additional variables p. The relative effect of this upwards bias is worse for the lower sample size conditions, but has less of an impact with increased correlation r as the corresponding increase in the absolute value of $\overline{\lambda}_0^{(S)}$ dwarfs the bias. One consequence of the upward bias is that all small-sample-with-limited-correlation conditions that had a similarly restricted non-optimal baseline at population level, now at sample level are ordered as a function of the number of variables p.

Model noncentrality $\lambda_m^{(S)}$**.** Under asymptotical theory given regularity conditions (e.g., Steiger et al., 1985), the χ_m^2 fit statistic when the true model is estimated, is expected to follow a central chisquare sampling distribution with mean df. Hence, the

sample noncentrality of the model $\overline{\lambda}_m^{(S)}$ should tend to its expected value 0.

However, some upward sampling bias in $\overline{\lambda}_m^{(S)}$ is present for almost all simulation conditions, although in absolute terms this is smaller than for $\overline{\lambda}_0^{(S)}$.The true model's noncentrality (and hence its sampling bias) is most affected by the number of variables p (see Table 1), and in contrast to its prominent role in the null model unaffected by the amount of correlation r. The most severe bias is observed in the low-sample-size-many-variables conditions ($p = 24, n = 100$). Overall, increasing sample size seemed to reduce the biasing effect of the additional variables. The finding of large sampling bias as a function of increasing number of manifest variables and moderated by sample size corresponds to earlier findings in the literature (e.g., Moshagen, 2012). Notice that the sampling variation unaccounted for by the design factors (i.e., $1 - \eta^2_{total} = .671$) is also much higher for the model noncentrality than for the null baseline noncentrality (i.e., $1 - \eta^2_{total} = .001$).

Comparative Fit Index (CFI). The asymptotically-derived sampling distribution of the CFI has not yet been established in the literature although logically it would conform to the sampling distribution of a ratio of two dependent shifted (non)central chisquare distributions, with the caveat that even a shifted noncentral chisquare is not fully applicable for the null baseline model. What we identified so far in the simulation study is that sampling affects the numerator $\lambda_m^{(S)}$ and denominator $\lambda_0^{(S)}$ of the CFI in a slightly different fashion. The resulting effect patterns on CFI in our simulation design (see Table 1) reflect this duality and lead to a mix of both λ-patterns, with the most central role for correlation r followed by sample size n, whereas the effect of the number of variables p has become negligible.

As we looked at CFI values for estimated true models, all observed CFI values should be indicative of the kind of sample values that can be expected to express good model fit. The 5% CFI quantile shows that the expected range of realistic CFI values actually varies greatly and covers a broad range across conditions (see Table B1). This difference becomes most prominent in those conditions where low sample size co-occurs with low correlation. In the most extreme situation (i.e., $n = 100, p = 24, r = .1$), 5% of the replications even have CFI values below or equal to .57. As reference to get the picture of

the whole range, 16% of replications in this condition still have CFI values above or equal to .95. At the same time, for some conditions (e.g., but not exclusively, the conditions where correlation $r = .9$) the range of realistic CFI values is much more limited as the 5% quantile was already as high as .99 or even 1.

Rule of thumb $CFI \geq .95$**.** The common rule of thumb for CFI states that CFI should be at least .95 to speak of acceptable goodness of fit, and otherwise if CFI $< .95$ one would reject the model. Given that the true model is fitted each time, the ideal outcome is of course a rejection rate of 0%. The results in Table B1 however, show that this is not accurate for all conditions. The median rejection rate is 0% but the average is 8% with a maximum of 84%. Of our 96 conditions, 43 had a non-zero rejection rate and 27 a rejection rate larger than 5%.

These results follow automatically from the observed ranges of CFI values for a true model not being consistent with the range implied by the rule of thumb $[.95, 1]$. The much wider or at times more narrower range of observed CFI for the estimated true model would imply that the rule of thumb should/could in fact be made more lenient or strict depending on the situation. A point to which we will return in the discussion.

Metric space principle $CFI|\lambda_0^{(S)}$**.** In line with our starting 'metric space' principle that the baseline determines differentiation power of CFI, the effect size patterns (see Table 1) for the model rejection rates given the rule of thumb follow the trends for the (sample and population) null baseline noncentrality yet with a diminished role of the number of variables p. Hence, increasing the metric space by increasing CFI's denominator through increasing either of the three design components has a positive effect on the size of $\lambda_0^{(S)}$, the size and range of CFI values, and the resulting model rejection rates according to the common rule of thumb (see also Table B1 for a detailed overview of results).

The observed diminished role of p is due to the set of conditions where low sample sizes are combined with low correlation in the data (i.e., $n = 100$ & $r \leq .5$ or $n = 200$ & r $\leq .2$) where a larger number of variables p leads to higher (see the excerpted conditions in Table 2) instead of the generally expected lower rejection rates. Sampling variability

and bias in those conditions destroy the regularity of the metric space principle. Focusing on one of the low-sample-size-low-correlation conditions, Figure 1 shows an example of how sampling variation in $\lambda_m^{(S)}$ relates to sampling variation in $\lambda_0^{(S)}$ as a function of the number of variables p. The horizontal and vertical line in the figure respectively show the average value of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ within a specific condition. Given the definition of CFI (see Equation 1), the diagonal line is the critical line representing the combination of $\lambda_m^{(S)}$ values and $\lambda_0^{(S)}$ values that result in CFI = .95. When replications are positioned in the area above this line, the corresponding CFI value will always be below .95, leading to rejection of the model. In other words, the values of $\lambda_m^{(S)}$ in these situations are becoming too large compared to their $\lambda_0^{(S)}$ counterpart to acquire good model fit according to CFI. While replications positioned on or below the diagonal line correspond to good model fit according to the .95 rule of thumb for CFI.

For both, $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$, their mean values increase with additional variables p as seen in their respective marginal distributions. However, the trend in $\lambda_m^{(S)}$ seems to be dominant over the trend in $\lambda_0^{(S)}$, as with additional variables p, $\lambda_m^{(S)}$ results in more extreme values relative to the $\lambda_0^{(S)}$ counterparts as seen in the heavier right tail in the distribution of the former. As a consequence, more replications are wrongly classified as showing inadequate model fit. In these specific conditions, problems in CFI performance are due to the strong sampling variation and bias in the numerator $\lambda_m^{(S)}$ that counteracts the positive effect of increased average size of the metric space reflected by the denominator $\lambda_0^{(S)}$.

Table 2

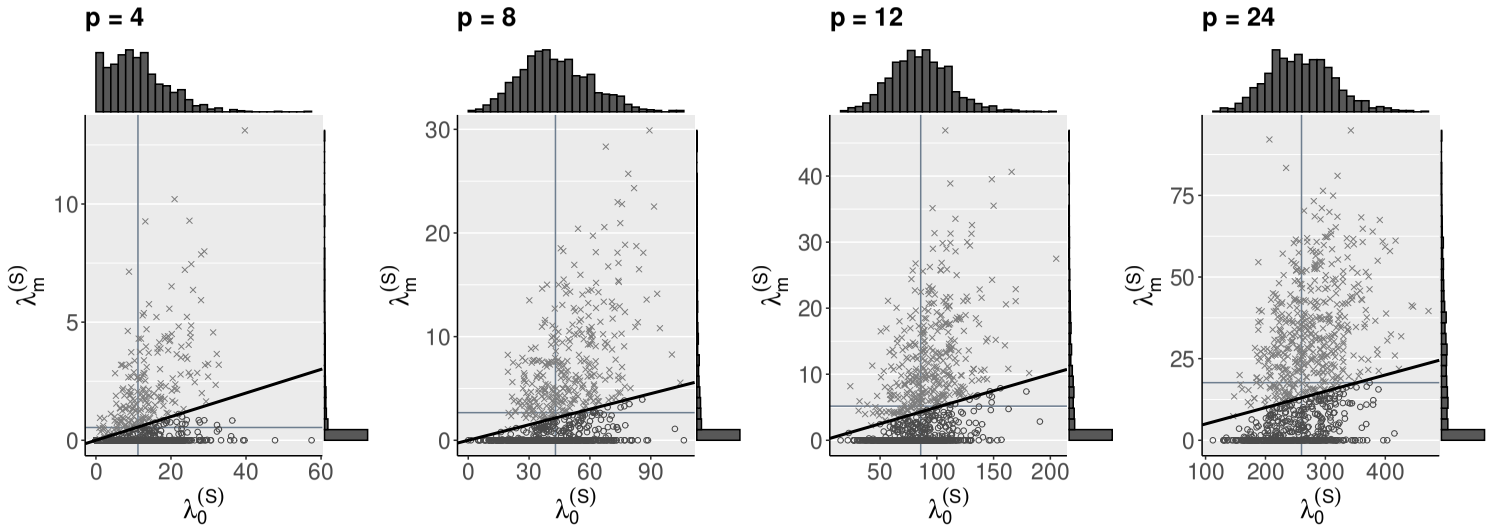
Contradicting the metric space principle: Negative effect of the number of variables p on the performance of CFI.

n	r	$p = 4$			$p = 8$			$p = 12$			$p = 24$		
		$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	$< .95$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	$< .95$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	$< .95$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	$< .95$
100	0.1	6.3	0.3	18.8%	22.7	2.7	41.6%	46.7	6.0	51.7%	152.6	32.2	84.2%
	0.2	19.7	0.6	20.1%	69.0	3.1	34.3%	133.0	5.7	34.4%	371.9	31.4	68.2%
	0.3	43.7	0.7	11.3%	137.0	3.4	19.2%	250.6	6.8	21.6%	653.9	32.3	45.8%
	0.5	116.2	0.8	1.6%	334.6	3.0	1.8%	583.2	6.6	2.6%	1372.4	32.3	7.6%
200	0.1	11.2	0.5	20.6%	43.0	2.7	31.8%	85.8	5.2	37.6%	260.1	17.6	47.4%
	0.2	40.5	0.7	12.9%	139.4	2.7	14.8%	258.9	4.9	12.4%	696.4	18.1	17.6%

Note. In general an increase in the size of the metric space is expected to have a positive effect on the CFI model rejection rates. However, the results, excerpted from Table B1, show those conditions where additional variables p result in increased rejection rates for CFI, even though $\overline{\lambda}_0^{(S)}$ increases as expected. It should however be noted that in some conditions the rejection rates are still close to zero (e.g., when $n = 100$ and $r = .5$). With $\overline{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\overline{\lambda}_m^{(S)}$ = average sample noncentrality for the estimated true model; $< .95$ = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95.

Figure 1

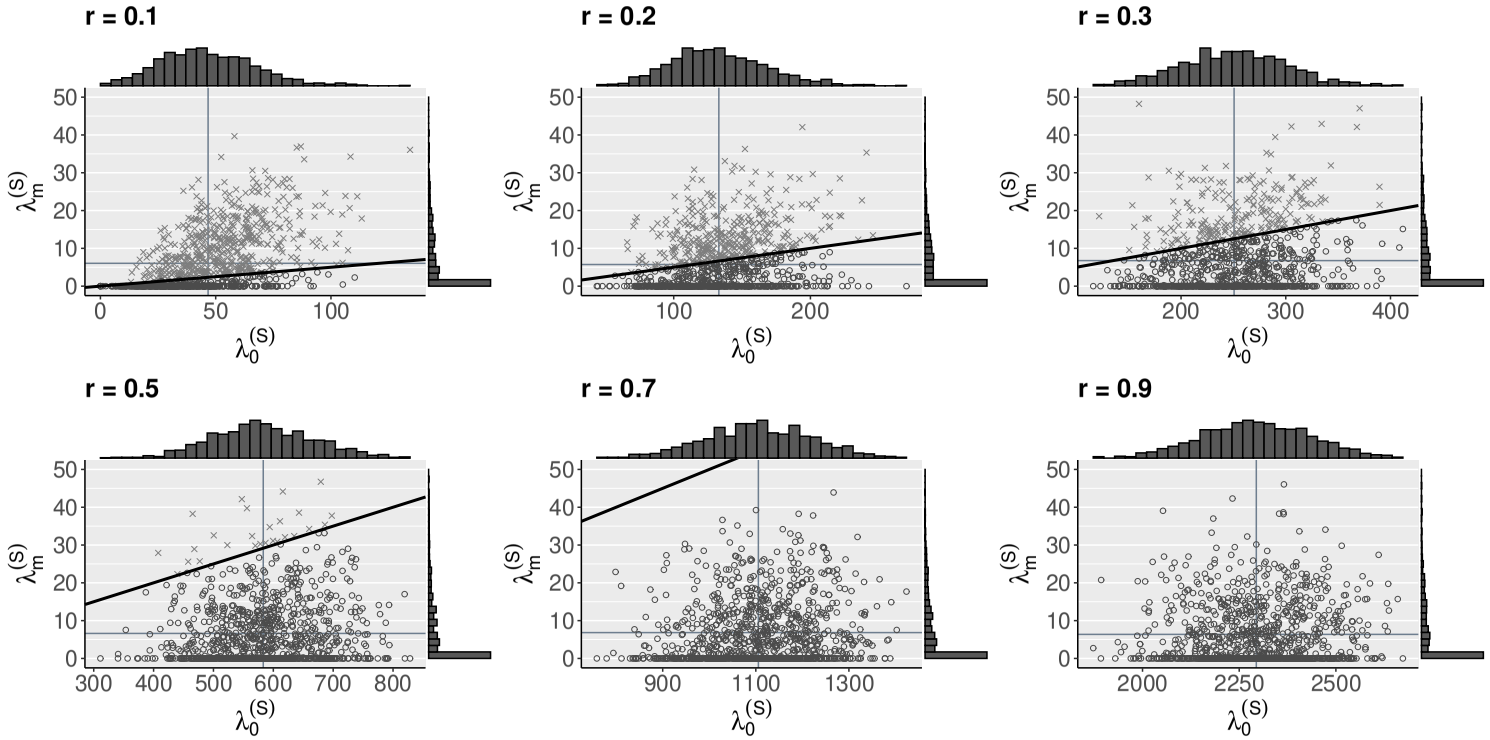
Scatterplot with marginal distributions of $\lambda_m^{(S)}$ *and* $\lambda_0^{(S)}$ *as a function of the number of variables* p *for the conditions where* $n = 200$ *and* $r = 0.1$.



Note. The horizontal and vertical line in the figure respectively show the average value of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ within a specific condition. With $\lambda_m^{(S)}$ = sample noncentrality for the estimated true model; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality. Given that CFI $= 1 - \frac{\lambda_m}{\lambda_0}$, the diagonal line is the critical line representing the combination of $\lambda_m^{(S)}$ values and $\lambda_0^{(S)}$ values that results in CFI = .95. Replications that are positioned in the area above this line will always result in CFI values below .95, leading to rejection of the model. While replications positioned on or below the diagonal line will result in good model fit according to the .95 rule of thumb for CFI. The pattern observed is for the low-sample-size-low-correlation conditions not conforming to the metric space principle, for which theoretically unexpected higher rejection rates occur with increasing number of variables (see also Table 2).

Figure 2

Scatterplot with marginal distributions of $\lambda_m^{(S)}$ *and* $\lambda_0^{(S)}$ *as a function of the correlation* r *for the conditions where* $n = 100$ *and* $p = 12$.



Note. The horizontal and vertical line in the figure respectively show the average value of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ within a specific condition. With $\lambda_m^{(S)}$ = sample noncentrality for the estimated true model; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality. Given that CFI $= 1 - \frac{\lambda_m}{\lambda_0}$, the diagonal line is the critical line representing the combination of $\lambda_m^{(S)}$ values and $\lambda_0^{(S)}$ values that results in CFI = .95. Replications that are positioned in the area above this line will always result in CFI values below .95, leading to rejection of the model. While replications positioned on or below the diagonal line will result in good model fit according to the .95 rule of thumb for CFI. In contrast to Figure 1, the pattern seen here is the dominant pattern conforming to the metric space principle, instead of the exception to the rule.

In the majority of the cases, this bias-interference is not applicable and the general metric-space principle works out despite sampling variation and bias in CFI's numerator and denominator. Figure 2 serves as an illustration of this principle. Whereas the distribution of $\lambda_m^{(S)}$ remains relatively constant across increasing correlation, the distribution of $\lambda_0^{(S)}$ takes big steps upwards, dwarfing any sampling bias in $\lambda_m^{(S)}$. The increase in correlation leads to a big increase in null baseline noncentrality which goes together with a decrease in the rejection rates of the CFI for the correctly specified model. The same results hold with increasing sample size n, whereas for increasing number of variables p it is less demarcated due to the opposing bias in $\lambda_m^{(S)}$.

Don't interpret CFI depending on RMSEA of null model?

As indicated before, additional specifications on the use of the general rule of thumb for CFI have been around. For example, one lesser known qualification advocated for on a popular web resources on SEM fit indices recommends that "CFI should not be computed if the RMSEA of the null model is less than .158 or otherwise one will obtain too small a value of the CFI" (Kenny, 2015). However, formal support for this recommendation was not given. Hence, we used the results from the main simulation study to follow up on the usefulness of this specific qualification in practice. We expected that if this rule of thumb works, cases where $\text{RMSEA}_0 < .158$ co-occur with a CFI value below the commonly adopted .95 threshold more often than not for models that fit.

As an initial rough effectiveness indicator of this rule of thumb we cross-classified all replications for each condition from the main simulation study based on whether the sample RMSEA_0 and CFI values were below or above their respective thresholds (see Table 3). On average the incidence of $\text{RMSEA}_0 < .158$ amounted to 31% of the cases. Given $\text{RMSEA}_0 < .158$, the probability for also obtaining a CFI value below .95 was on average 17.5% with a range across conditions between 0 and 84.2%. The reason for this wide range can be clearly illustrated by translating the $\text{RMSEA}_0 < .158$ into a corresponding required value for the null baseline noncentrality $\lambda_0^{158} = \text{RMSEA}_0^2 \times (n - 1) \times \text{df}_0$. This threshold null baseline noncentrality λ_0^{158} value indeed only depends on two design factors - the number of variables p ($\eta^2 = .483$), sample size n ($\eta^2 = .225$) -

, and their interaction $n \times p$ ($\eta^2 = .293$), but not on the third factor data correlation r (i.e., $\eta^2 = .000$ for r, $p \times r$, $r \times n$, & $p \times r \times n$). As one example, Table 4 clearly illustrates the ignorance of this $\text{RMSEA}_0 < .158$ threshold for the conditions where sample size $n = 200$ and $\text{df}_0 = 28$ (i.e., number of variables $p = 8$). Note that these results generalize across the other conditions. The $\text{RMSEA}_0 < .158$ specification wrongly assumes a null baseline noncentrality $\lambda_0^{.158}$ that remains constant regardless of the correlation r in the data, whereas CFI and its denominator the null baseline noncentrality λ_0 are highly sensitive to exactly this correlation.

Table 3

Cross-classification of all replications in the main simulation study based on their RMSEA_0 and CFI value relative to the corresponding thresholds.

	RMSEA_0	
CFI	$< .158$	$\geq .158$
$\geq .95$	24.91% [0-100%]	67.17% [0-100%]
$< .95$	6.29% [0-84.2%]	1.64% [0-21.1%]

Note. RMSEA_0 = RMSEA values for the null baseline model; CFI = CFI values for the estimated true model. For the the proposed rule of thumb to work, RMSEA_0 values below .158 ought to co-occur with CFI values below .95. Each cell in the cross-classification contains the overall average percentage and range of average percentages of replications across conditions in the main simulation study that is consistent with its thresholds-requirements.

In the end, the overall negative predictive value of the .158 rule of thumb appears to be not too reliable (i.e., $\Pr(\text{CFI} < .95 | \text{RMSEA}_0 < .158)$). Hence, it varies highly whether we can indeed expect too low CFI values given a correctly specified model when $\text{RMSEA}_0 < .158$. On the other hand, the correct decision of acceptable fit (i.e., $\text{CFI} \geq .95$) is taken in on average 95.8% (range across conditions = 52.9-100%) of the cases that $\text{RMSEA}_0 \geq .158$. Hence, the overall positive predictive value (i.e., $\Pr(\text{CFI} \geq .95 | \text{RMSEA}_0 \geq .158)$) of the .158 rule of thumb is more promising. The reason for this difference is that for specific

settings the null baseline noncentrality corresponding to the $\text{RMSEA}_0 = .158$ threshold is unreachable. This is illustrated in the latter columns of Table 4, where for this particular case of $n = 200$ and $p = 8$, RMSEA_0 values below .158 can only occur in conditions with correlations r below .3 (i.e., $\lambda_0^{(S)} < \lambda_0^{.158}$). Note that the specific breakdown point does vary depending on sample size n and number of variables p. In the end, this leads exactly to flagging down some of the conditions in which the CFI baseline for comparison is rather too small for effective model differentiation.

Table 4

Attainability of the threshold: Sensitivity of the null baseline noncentrality and CFI to data correlation r in relation to the constant RMSEA$_0$ rule of thumb and corresponding threshold in terms of the null baseline noncentrality $\lambda_0^{.158}$.

r	threshold		$\lambda_0^{(\Sigma)}$	$\lambda_0^{(S)}$			CFI		
	RMSEA_0	$\lambda_0^{.158}$		M	MIN	MAX	M	MIN	MAX
.1	.158	139.099	13	43	0	106	.95	.55	1.00
.2	.158	139.099	109	139	53	265	.98	.78	1.00
.3	.158	139.099	245	275	152	453	.99	.92	1.00
.5	.158	139.099	642	667	397	970	1.00	.96	1.00
.7	.158	139.099	1303	1324	1019	1655	1.00	.98	1.00
.9	.158	139.099	2798	2832	2409	3326	1.00	.99	1.00

Note. The results stem from the main simulation study and show an example for the conditions where the sample size $n = 200$ and the number of variables $p = 8$. RMSEA_0 = RMSEA threshold of the null baseline model; $\lambda_0^{.158} = .158$ threshold for RMSEA_0 translated in terms of null baseline noncentrality; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality; CFI = CFI value for the estimated true model.

In sum, despite its relatively good average positive predictive value, the proposed .158 rule of thumb does not fully meet its purpose. In its current form it is too general and ignores the role of one of the key components of CFI (cf. data correlation). In light of the

wide range of values and variation in performance, it does not seem advisable to utilize a fixed general RMSEA threshold as the conclusive answer for assessing whether or not to apply the CFI for fit assessment.

Discussion

If we would desire not mere mindless binary rule-following but more deliberate practice when assessing model fit, we need to better clarify what type of fit each of the different indices stand for and to provide a better insight in their inner workings to understand why fit indices behave like they do. In this study, we started with such endeavour for the Comparative Fit Index.

CFI is a relative model fit measure expressed as a ratio of the noncentrality of the model of interest to that of a baseline comparison model. In essence this implies that the CFI is in fact a standardized statistic where the standard of comparison is typically provided by the noncentrality of the null model that is by default chosen as comparison model. This does mean that one CFI is not the other because the baseline standard, the noncentrality of the null model, is determined by data dimensions (i.e., $n \times p$) and amount of multivariate dependence in the data (i.e., $|\mathbf{R}|$). This is important as the implications of absolute value judgement of good fit according to CFI might not correspond to the relative improvement CFI stands for. With a small CFI metric space, low relative improvement does not necessarily imply that a model is not good in terms of absolute fit, while a high relative fit given a large metric space can still be associated with a large amount of absolute misspecification. The broader the baseline, the less strict the CFI $\geq$ .95 rule of thumb becomes as more absolute misspecification is allowed for a model that is considered to adequately fit. This natural feature of a standardized/relative measure such as CFI, brings Moshagen and Auerswald (2018) to caution strongly against CFI's use for evaluating absolute fit of a single model.

However, such decontextualized assessment of fit of a single model is unfortunately quite common place in practice with the default application of the binary rule of thumb: CFI $\geq$ .95 means "good fit" whatever that might mean. If we formalize the latter as correctly identifying the true model as a good fitting model, with a binary decision rule

that works at least 95% of the time, our simulation results show that the rule of thumb needs to be adjusted based on data characteristics or only be applied under certain qualifications.

Qualifications for use of CFI's rule of thumb. Our results illustrate the theoretically derived principle that a wider basis for model differentiation is provided by increasing the three core components of the null baseline noncentrality – sample size n, number of variables p, and multivariate dependence as reflected by $|\mathbf{R}|$, the determinant of the data correlation matrix. This results in high rates of qualifying the correctly specified model as having good fit in high signal to noise conditions, that is high correlation with added high sample size regardless of the number of variables. In contrast, in low signal to noise conditions, that is low sample size and low correlation, the CFI $\geq .95$ rule was too strict and an increase of the number of variables made matters even worse. In the latter conditions, the null baseline model is already quite close in absolute fit to the correctly specified model, hence it is less likely to observe a huge relative change of 95% of that small distance even for a correctly specified model. Consequently, a word of caution for the current binary use of the CFI $\geq .95$ rule of thumb in such conditions is in order. Sample sizes below 200 are unfortunately not uncommon (Jackson et al., 2009; MacCallum & Austin, 2000) and the prevailing pragmatic idea that standardized factor loadings of .3 ($r = .09$) and .4 ($r = .16$) are sufficient for meaningful interpretation (Brown, 2015) seems too optimistic.

The CFI $\geq .95$ rule of thumb would approximately work in this 95% correct sense as a function of sample size and correlation: for $n = 1000$, a correlation of at least $r = .1$, for $n = 500$, a correlation of at least $r = .2$ is required, for $n = 200$ a correlation of at least $r = .3$, and for $n = 100$ a correlation of at least $r = .5$. Based on our simulation results, a conjecture could be put forward that a baseline noncentrality of $\lambda_0^{(S)} \geq 1400$ provides a sufficient broad metric space for fine-grained model differentiation using the CFI (e.g., conditions in line with this requirement had very narrow CFI range for the true model and far above the .95 rule of thumb). This is a conservative guideline as things do not necessarily look bad in all smaller baseline conditions. Although the general CFI

metric-space principle holds, the specific values suggested here are of course based on the limited set of levels of factors considered in the small simulation study, and would be somewhat adjusted with availability of results for more factor levels (e.g., extra sample size conditions) or even other design factors such as the data-generating model. Yet, the general identified patterns related to the CFI baseline are mostly data driven and core points and non-value specific recommendations can in that sense be trusted to generalize quite well.

We already mentioned that these type of additional qualifications, on when the CFI rule of thumb can be used, are not something new. Specifically, we looked into the recommendation not to use CFI if the RMSEA of the null model is less than .158 (Kenny, 2015). Even though this qualification does attempt to provide a more nuanced reporting of CFI, the simulation results showed that in light of its wide variation in performance across conditions, it is not advisable to use this specific qualification without careful deliberation. Yet, the underlying idea does contain merit as it essentially intends to filter out cases where there is a lack of covariance and high levels of noise in the data. Perhaps, we should not even consider SEM in such cases in the first place (e.g., Barrett, 2007) or at the minimum realize that it's not reasonable to expect a large relative fit difference from a null baseline model that itself is already very closely fitting to the data in an absolute parsimony fit sense.

Adjusting CFI's rule of thumb. Alternatively, instead of including additional qualifications on when to use CFI's rule of thumb, we could also adjust the rule of thumb depending on data characteristics. The general pattern of results shows that the CFI threshold should even become stricter in the more optimal situations (high correlation r, high sample size n: CFI 5% quantiles as high as .99), while it needs to be reduced considerably in the less optimal situations (low correlation r, low sample size n). The latter could even result in setting a threshold value as low as CFI $\geq .57$ for a specific condition ($n = 100, p = 24, r = .1$). When realistic CFI values for a true model cover such a broad range, CFI looses its informativeness for absolute model fit assessment.

Effect size. Another more drastic, but likely preferable alternative to including

additional qualifications on when to use CFI's rule of thumb or adjusting its threshold value as a function of data characteristics, would be to actually interpret CFI's value. In this respect, it is useful to see CFI as an extension of the linear regression model's R-square effect size measure to the broader SEM field. Both measures have indeed a similar setup:

$$\begin{aligned} r^2_{Y|\mathbf{X}} &= 1 - \tfrac{SS_{error}}{SS_{total}} \\ \text{CFI}_{(m,0)} &= 1 - \tfrac{\lambda_m}{\lambda_0} \\ \text{effect size} &= 1 - \tfrac{\text{misspecification target model vs saturated model}}{\text{misspecification null model vs saturated model}} \end{aligned}$$

This further clarifies that in essence, CFI is, like the R-square, a standardized effect size measure and hence all reservations with respect to interpretations of standardized effect size measures (e.g., Baguley, 2009) transfer to the interpretation of CFI. Such a realization has two major implications.

Firstly, CFI can be a useful benchmark metric for interpreting the relative magnitude of the effects within the same application dataset. Having a set of competing models, CFI can be used to quantify the effect size of the paths in which the models differ. In other words, we are using CFI as intended as an incremental comparative fit index among a set of models for the same dataset and interpreting its value in terms of relative magnitude.

Secondly, comparing CFI's across different datasets is not straightforward as given their standardized nature, a value of .95 is indeed similar in relative magnitude, but not necessarily in absolute magnitude. The latter would require that the denominator in CFI's formula remains constant across datasets. Where R-square is a relative reduction in variance not accounted for, and the denominator is a proxy for total variance in the outcome variable, CFI is a relative reduction in model noncentrality, and – when the baseline model is the null model – the denominator can be seen as a proxy for the amount of generalized variance in the manifest variables of the model, the determinant of the observed correlation matrix $|\mathbf{R}|$. An interpretation of CFI in terms of absolute magnitude would require an interpretation of the amount of generalized variance, that is the value of this determinant. The determinant of a correlation matrix can be seen

geometrically as the volume of the swarm of standardized data points, with $|\mathbf{R}| = 1$ in case of all zero-correlations (corresponding to a 'ball' in a multidimensional plane) and with $|\mathbf{R}| = 0$ for a matrix with perfect linear dependence (a ball flattened along at least one dimension). Whereas people in practice often already find it hard to interpret the absolute magnitude of a variance, it is fair to say that even fewer people have a good intuition about what a large or small generalized variance or determinant is for their dataset. The current lack of straightforward interpretability of CFI in terms of absolute magnitude essentially disqualifies it in practice for assessing the absolute fit of a single model or for comparing model fit between different datasets.

Nevertheless, the central role of this determinant should revive some interest in understanding classic measures of multivariate statistics (e.g., Anderson, 1958) to further our understanding of more modern SEM practices. In the meantime, we recommend implementing a reporting standard where next to the CFI also its denominator, the baseline model's noncentrality λ_0 is reported to provide some context for interpretation. These quantities are generally available or easy to request in common SEM software such as Mplus or R:lavaan. If the default null model is chosen as baseline, explicit reporting of its three key components – sample size n, number of manifest variables p, determinant of the observed correlation matrix $|\mathbf{R}|$ – would help in gaining some intuition on common reference values for these data characteristics[12] in your field of application and eventually allow for a better interpretation of relative and absolute magnitude of CFI even across datasets.

Other Considerations. One limitation of the current study is that we only considered the default null model in which all observed variables are uncorrelated while looking at the performance of CFI. However, it was already discussed by Bentler and Bonett (1980, p. 604) that "the incremental fit indices depend critically on the availability of a suitable framed null model". Widaman and Thompson (2003) argue that there are numerous situations in which the default null model would be an improper choice. Different alternatives for specification of a proper baseline model can be found in the liter-

[12] In a linear model, it is similarly good practice to report next to the R-square also the total variance of the outcome variable (or alternatively the residual standard deviation) to contextualize the percentage.

ature (e.g., Little, 2013; Widaman & Thompson, 2003). While Widaman and Thompson (2003) already touched upon it, going forward it is important to systematically evaluate the potential influence of the chosen null model on performance evaluation of the different comparative fit indices under different circumstances, as well as the substantive consequences of comparing a model of interest to a more meaningful baseline model.

In this study, we focused on the typical maximum likelihood estimator used in structural equation modelling, yet it would be of interest to expand the study to other estimators in particular for the categorical data case, both including limited-information estimators based on the polychoric correlation matrix or bivariate contingency tables as well as full-information estimators based on the item response patterns (cf. item response theory tradition). A move to the categorical case might also essentially call for a different baseline model; for categorical data, correlations are strongly constrained by their marginal distributions as mean and variance are intertwined.

Another avenue for further research would be to explore the impact of transitioning from classic estimates for the two noncentrality parameters in the CFI to bias-corrected estimates as for instance suggested by Raykov (2005). Raykov did add caution as for instance a bias-correction bootstrap estimate of noncentrality is feasible, but the properties of the approach for this particular case have not been fully studied. Yet deflating differential sampling bias in both numerator and denominator of CFI could potentially ensure that its sampling behavior is even more systematic and in line with the driving components of the baseline.

Conclusion

To conclude, the CFI does what it is supposed to do, but we haven't been using it in a smart fashion. The CFI is a relative fit measure where the standard for comparison is provided by the noncentrality of the (null) baseline model. The common CFI $\geq .95$ rule of thumb implies that regardless of context we are happy with a reduction of 95% of the misspecification by the null model. Current practices make us prone to hunting down this magic CFI $\geq .95$ value as a pseudo absolute fit measure disregarding the existence of the baseline. CFI as an absolute but meaningless criterion that needs to be fulfilled to achieve

an adequate model that can serve as starting point for further analysis. To help remedy this, we recommend that at a minimum a dual reporting standard is followed where both model of interest and the (null) baseline model are evaluated to provide proper context for interpretation of the CFI value. By making the presence of the baseline (and its core components) explicit in the reporting, the need to take it into account when interpreting fit indices also becomes explicit and non-ignorable. Even more optimal would be if CFI is not simply used as a mere number in a search for model adequacy but used as a real relative fit index intended to evaluate the relevance of cumulative theoretically motivated model restrictions in terms of % reduction in misspecification as measured by the baseline model (Bentler & Bonett, 1980).

References

Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd). Guilford Press.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research*, *37*(1), 1–36.

Graybill, F. A. (1983). *Matrices with Applications in Statistics* (2nd). Wadsworth International Group.

Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis* (6th). Pearson.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Jackson, D. L., Gillapsy, J. A., & Purch-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23.

Kenny, D. A. (2015). Measuring model fit. http://davidakenny.net/cm/fit.htm

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*(2-3), 220–239.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201–226.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 86–98.

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in strucutral equation modeling. *Psychological Methods*, *23*(2), 318–336.

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

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*, 1148–1172.

R Core Team. (2020). R: A language and environment for statistical computing.

Raykov, T. (2005). Bias-corrected estimation of noncentrality parameters of covariance structure models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 120–129.

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, Article 1715.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310–334.

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 21–40.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika*, *50*(3), 253–263.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37.

Appendix A: Noncentrality λ_0 of the null model

$$\lambda_m = \chi^2_m - \mathrm{df}_m \tag{1}$$

$$= F_m(n-1) - \mathrm{df}_m \tag{2}$$

$$= (\log|\hat{\boldsymbol{\Sigma}}_m| - \log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\hat{\boldsymbol{\Sigma}}_m^{-1}) - p)(n-1) - \mathrm{df}_m \tag{3}$$

Equations 1-3 outline how the noncentrality parameter of any model would be estimated as the difference between the model's chisquare against the saturated model and the model's degrees of freedom. The model's chisquare value is based on the product of the sample size n and the minimum value F_m of the used fit function. Under maximum likelihood estimation, F_m is a function of the discrepancy between the model-implied variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_m$ and the observed variance-covariance matrix $\mathbf{S}$ (e.g., Bollen, 1989), where p represents the number of observed variables and $\mathrm{tr}(\mathbf{X})$ and $|\mathbf{X}|$ are respectively the trace and determinant of a matrix $\mathbf{X}$.

Key in getting to the expression for the noncentrality λ_0 for the null model (Equation 2 in the main text) is that the minimal fit value F_0 for the null model can be further simplified using the fact that the model-implied covariance matrix under the null model comes down to a diagonal matrix $\mathbf{diag}(\mathbf{S})$ with the observed variances on the diagonal (cf. Equation 5). This results in $\mathbf{S}\hat{\boldsymbol{\Sigma}}_0^{-1}$ leading to a matrix with all ones on the diagonal such that the trace equals the number of observed variables p and cancels out the subsequent $-p$ term in the expression for F_0 (cf. Equation 6).

$$F_0 = \log|\hat{\boldsymbol{\Sigma}}_0| - \log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\hat{\boldsymbol{\Sigma}}_0^{-1}) - p \tag{4}$$

$$= \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\mathbf{diag}(\mathbf{S})^{-1}) - p \tag{5}$$

$$= \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| + p - p \tag{6}$$

Using the fact that the determinant of a matrix product can be split into products of determinants, each of the remaining two log determinants can be written out given that

a variance-covariance matrix $\mathbf{S}$ is a multiplicative function of a corresponding correlation matrix $\mathbf{R}$ and an inverse diagonal matrix with standard deviations on the diagonal. Thus we have

$$\log|\mathbf{S}| = \log|\sqrt{\mathbf{diag}(\mathbf{S})}\ \mathbf{R}\ \sqrt{\mathbf{diag}(\mathbf{S})}| \tag{7}$$

$$= \log|\sqrt{\mathbf{diag}(\mathbf{S})}| + \log|\mathbf{R}| + \log|\sqrt{\mathbf{diag}(\mathbf{S})}| \tag{8}$$

$$= \log\prod_{j=1}^{p}\sqrt{S_{jj}} + \log|\mathbf{R}| + \log\prod_{j=1}^{p}\sqrt{S_{jj}} \tag{9}$$

$$= \log\prod_{j=1}^{p} S_{jj} + \log|\mathbf{R}| \tag{10}$$

and

$$\log|\mathbf{diag}(\mathbf{S})| = \log|\sqrt{\mathbf{diag}(\mathbf{S})}\ \mathbf{I}\ \sqrt{\mathbf{diag}(\mathbf{S})}| \tag{11}$$

$$= \log\prod_{j=1}^{p} S_{jj} + 0 \tag{12}$$

where Equation 12 makes use of the fact that the correlation matrix of a diagonal variance-covariance matrix is an identity matrix $\mathbf{I}$ which determinant is exactly equal to 1.

The re-expressions of the log determinant terms in Equations 10 and 12 allow to simplify the expression for F_0 further by elimination

$$F_0 = \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| \tag{13}$$

$$= \log\prod_{j=1}^{p} S_{jj} - \log\prod_{j=1}^{p} S_{jj} - \log|\mathbf{R}| \tag{14}$$

$$= -\log|\mathbf{R}| \tag{15}$$

such that the estimated noncentrality of the null model comes down to

$$\lambda_0 = F_0(n-1) - \mathrm{df}_0 = -\log|\mathbf{R}|(n-1) - p(p-1)/2$$

where $p(p-1)/2$ is the degrees of freedom of the null model.

Appendix B: Results of main study

Table B1

CFI and its underlying noncentrality measures in numerator and denominator as a function of data correlation $\mathbf{R}$*, sample size* n*, and number of variables* p*.*

n	r	$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	$<.95$	Q.05			$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	$<.95$	Q.05
		$p = 4$							$p = 8$						
100	0.1	0.948	0.888	0.0	6.3	0.3	18.8%	0.74	0.813	0.608	0.0	22.7	2.7	41.6%	0.62
	0.2	0.819	0.777	13.9	19.7	0.6	20.1%	0.81	0.503	0.388	40.7	69.0	3.1	34.3%	0.83
	0.3	0.652	0.615	36.8	43.7	0.7	11.3%	0.91	0.255	0.203	108.5	137.0	3.4	19.2%	0.90
	0.5	0.312	0.304	110.3	116.2	0.8	1.6%	0.96	0.035	0.031	306.8	334.6	3.0	1.8%	0.96
	0.7	0.084	0.083	242.1	249.1	0.8	0.0%	0.98	0.001	0.001	637.3	667.0	3.1	0.0%	0.98
	0.9	0.004	0.004	553.9	560.8	0.9	0.0%	0.99	0.000	0.000	1385.0	1409.1	3.1	0.0%	0.99
		$p = 12$							$p = 24$						
	0.1	0.659	0.331	0.0	46.7	6.0	51.7%	0.61	0.292	0.015	0.0	152.6	32.2	84.2%	0.57
	0.2	0.275	0.145	63.1	133.0	5.7	34.4%	0.85	0.033	0.002	65.0	371.9	31.4	68.2%	0.81
	0.3	0.085	0.048	180.5	250.6	6.8	21.6%	0.91	0.002	0.000	337.7	653.9	32.3	45.8%	0.89
	0.5	0.003	0.002	509.3	583.2	6.6	2.6%	0.96	0.000	0.000	1065.7	1372.4	32.3	7.6%	0.95
	0.7	0.000	0.000	1042.0	1105.5	6.8	0.0%	0.98	0.000	0.000	2209.2	2517.1	32.3	0.0%	0.97
	0.9	0.000	0.000	2228.0	2294.2	6.4	0.0%	0.99	0.000	0.000	4712.2	5005.5	32.1	0.0%	0.98

Note. $\lvert\mathbf{R}_\Sigma\rvert$ = determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathrm{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\overline{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\overline{\lambda}_m^{(S)}$ = average sample noncentrality for the estimated true model; $< .95$ = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.

– Table B1 continued –

n	r	$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05			$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05
		$p = 4$							$p = 8$						
200	0.1	0.948	0.919	4.7	11.2	0.5	20.6%	0.76	0.813	0.704	13.4	43.0	2.7	31.8%	0.76
	0.2	0.819	0.795	33.9	40.5	0.7	12.9%	0.91	0.503	0.439	109.3	139.4	2.7	14.8%	0.91
	0.3	0.652	0.637	79.6	85.1	0.8	4.4%	0.95	0.255	0.225	245.1	275.3	2.7	3.0%	0.96
	0.5	0.312	0.308	226.6	232.3	0.7	0.1%	0.98	0.035	0.033	641.6	667.1	2.9	0.0%	0.98
	0.7	0.084	0.083	490.1	498.0	0.8	0.0%	0.99	0.001	0.001	1302.6	1323.8	2.8	0.0%	0.99
	0.9	0.004	0.004	1113.9	1119.3	0.7	0.0%	1.00	0.000	0.000	2798.0	2831.5	2.8	0.0%	1.00
		$p = 12$							$p = 24$						
	0.1	0.659	0.472	17.4	85.8	5.2	37.6%	0.78	0.292	0.071	0.0	260.1	17.6	47.4%	0.80
	0.2	0.275	0.202	192.3	258.9	4.9	12.4%	0.93	0.033	0.009	405.9	696.4	18.1	17.6%	0.92
	0.3	0.085	0.064	427.0	494.8	5.6	3.3%	0.96	0.002	0.001	951.3	1251.3	19.3	2.7%	0.96
	0.5	0.003	0.003	1084.6	1149.8	5.4	0.1%	0.98	0.000	0.000	2407.3	2692.8	19.0	0.0%	0.98
	0.7	0.000	0.000	2150.1	2222.3	5.4	0.0%	0.99	0.000	0.000	4694.5	4991.7	18.4	0.0%	0.99
	0.9	0.000	0.000	4521.9	4586.6	4.8	0.0%	1.00	0.000	0.000	9700.4	9991.0	17.9	0.0%	0.99

Note. $\lvert\mathbf{R}_\Sigma\rvert$ = determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathrm{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\overline{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\overline{\lambda}_m^{(S)}$ = average sample noncentrality for the estimated true model; $< .95$ = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.

– Table B1 continued –

n	r	$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05			$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05
		$p = 4$							$p = 8$						
	0.1	0.948	0.937	20.9	26.8	0.7	17.5%	0.87	0.813	0.770	75.4	103.4	2.7	20.1%	0.89
	0.2	0.819	0.811	93.7	99.3	0.8	3.1%	0.96	0.503	0.477	315.3	344.2	2.6	0.7%	0.97
	0.3	0.652	0.647	208.1	212.8	0.8	0.2%	0.98	0.255	0.243	654.7	684.9	2.7	0.0%	0.98
	0.5	0.312	0.311	575.6	581.9	0.7	0.0%	0.99	0.035	0.034	1646.0	1679.0	2.7	0.0%	0.99
	0.7	0.084	0.084	1234.3	1241.0	0.7	0.0%	1.00	0.001	0.001	3298.4	3323.2	2.7	0.0%	1.00
	0.9	0.004	0.004	2793.7	2803.4	0.7	0.0%	1.00	0.000	0.000	7037.1	7060.3	2.8	0.0%	1.00
500		$p = 12$							$p = 24$						
	0.1	0.659	0.577	142.5	210.9	4.7	16.7%	0.91	0.292	0.168	338.7	620.6	11.2	10.8%	0.94
	0.2	0.275	0.243	579.7	646.1	4.3	0.2%	0.97	0.033	0.020	1428.8	1708.3	12.1	0.0%	0.97
	0.3	0.085	0.076	1166.4	1233.0	4.3	0.0%	0.98	0.002	0.001	2792.3	3073.9	11.7	0.0%	0.99
	0.5	0.003	0.003	2810.4	2879.2	4.0	0.0%	0.99	0.000	0.000	6432.3	6714.4	13.2	0.0%	0.99
	0.7	0.000	0.000	5474.2	5549.3	4.3	0.0%	1.00	0.000	0.000	12150.1	12454.3	11.5	0.0%	1.00
	0.9	0.000	0.000	11403.8	11487.5	4.5	0.0%	1.00	0.000	0.000	24665.1	24938.3	12.3	0.0%	1.00

Note. $\lvert\mathbf{R}_\Sigma\rvert$ = determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathrm{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\overline{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\overline{\lambda}_m^{(S)}$ = average sample noncentrality for the estimated true model; $< .95$ = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.

– Table B1 continued –

n	r	$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05			$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05
		$p = 4$							$p = 8$						
1000	0.1	0.948	0.941	47.7	55.0	0.7	8.9%	0.93	0.813	0.790	178.9	208.4	2.5	5.3%	0.95
	0.2	0.819	0.815	193.4	198.6	0.7	0.1%	0.98	0.503	0.491	658.5	686.5	2.7	0.0%	0.98
	0.3	0.652	0.648	422.2	428.4	0.7	0.0%	0.99	0.255	0.249	1337.3	1365.9	2.8	0.0%	0.99
	0.5	0.312	0.311	1157.2	1164.5	0.7	0.0%	1.00	0.035	0.035	3320.0	3351.0	2.7	0.0%	1.00
	0.7	0.084	0.084	2474.5	2475.0	0.7	0.0%	1.00	0.001	0.001	6624.9	6634.5	2.6	0.0%	1.00
	0.9	0.004	0.004	5593.4	5596.0	0.7	0.0%	1.00	0.000	0.000	14102.2	14120.4	2.7	0.0%	1.00
		$p = 12$							$p = 24$						
	0.1	0.659	0.617	351.0	419.0	4.5	2.8%	0.95	0.292	0.222	953.4	1236.3	10.0	0.6%	0.97
	0.2	0.275	0.258	1225.4	1292.3	4.0	0.0%	0.99	0.033	0.025	3133.5	3416.4	10.7	0.0%	0.99
	0.3	0.085	0.081	2398.8	2458.1	4.0	0.0%	0.99	0.002	0.002	5860.7	6151.5	11.1	0.0%	0.99
	0.5	0.003	0.003	5686.8	5755.9	4.4	0.0%	1.00	0.000	0.000	13140.7	13412.6	9.7	0.0%	1.00
	0.7	0.000	0.000	11014.4	11058.4	4.2	0.0%	1.00	0.000	0.000	24576.3	24861.0	11.3	0.0%	1.00
	0.9	0.000	0.000	22873.7	22938.7	4.3	0.0%	1.00	0.000	0.000	49606.1	49921.5	10.4	0.0%	1.00

Note. $\lvert\mathbf{R}_\Sigma\rvert$ = determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathrm{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\overline{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\overline{\lambda}_m^{(S)}$ = average sample noncentrality for the estimated true model; $< .95$ = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.

5 Article 2: Multivariate Dependence

van Laar, S., & Braeken, J. (2022a). Caught of base: A note on the interpretation of incremental fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(6), 935–943. https://doi.org/10.1080/10705511.2022.2050730.

Caught Off Base: A Note on the Interpretation of Incremental Fit Indices

This note serves as a reminder that incremental fit indices are a form of standardized effect sizes and hence, all reservations with respect to interpretations of standardized effect sizes also transfer to their interpretation. Such a realization has major implications for the interpretation and use of incremental fit indices, for the theoretical (im)possibility of default universal rules of thumb in their application, and for simulation studies mapping incremental fit indices as if their value is comparable in an absolute sense across any and all conditions. A small but illustrative working example centered around the alleged impact of model type will drive these points home.

Model fit assessment and model comparison remain universally important but also confusing topics in structural equation modeling (SEM). Tons of model fit tests and diagnostic fit indices have been introduced for purpose of model fit assessment – with Marsh et al. for instance already looking at 29 fit indices early on in 1988 – and new developments are abundant and extend fit indices beyond their initial boundaries (e.g., non-normal data, bias-reduction; see for example Raykov, 2005; Yuan & Bentler, 2000). Recent practice has arguably converged to reporting multiple fit indices and following rules of thumb based on the work by Hu and Bentler (1999), with the chisquare statistic (χ^2), Root Mean Square Error of Approximation (RMSEA), and Comparative Fit Index (CFI) among the popular indices to use and report (Jackson et al., 2009). For model assessment guidelines and rules of thumb for fit indices to work, they should be proven to function rather universally across a broad scope of data and model characteristics. Yet, the extensive simulation literature on this matter has already put forward many factors that are influencing the general applicability of the rules of thumb (for a review, see e.g., Niemand & Mai, 2018) leading to a general caution on their universality.

This general caution is also readily ignored in practice where a binary search for adherence with the rules of thumb for a range of fit indices is the factual norm. The latter might come across as a surprise, but is in line with McDonald and Ho (2002) who

state that "it is sometimes suggested that we should report a large number of these indices, apparently because we do not know how to use any of them" (p. 72), resulting in a lack of deliberate decision making. In order to make more informed decisions with respect to the use of fit indices it is important to know how these fit indices work. However, as Lai and Green (2016) point out "the meaning of 'good' fit and how it relates to fit indices are not well understood in the current literature" (p. 234).

This manuscript sets out to remind/clarify what the meaning of good fit is for incremental fit indices and what implications this should have for their use in practice. The alleged impact of model type on incremental fit indices is used as a working example to elucidate the actual impact of the baseline as opposed to the type of target model.

Incremental Fit Indices

Incremental fit indices such as the Normed Fit Index (NFI: Bentler & Bonett, 1980), Comparative Fit Index (CFI: Bentler, 1990), or Tucker-Lewis Index (TLI: Tucker & Lewis, 1973) are part of a family of relative fit measures for structural equation modeling that involves locating a model of interest within a continuum of models from the worst fitting baseline model to the perfect fitting or saturated model. Incremental fit indices are much like SEM counterparts of r-square indices in linear regression.

$$
\begin{aligned}
r^2_{Y|\mathbf{X}} &= 1 - \frac{SS_{error}}{SS_{total}} \\
\text{NFI}_{(m,b)} &= 1 - \frac{\chi^2_m}{\chi^2_b} \\
\text{CFI}_{(m,b)} &= 1 - \frac{\lambda_m}{\lambda_b} = 1 - \frac{\chi^2_m - \text{df}_m}{\chi^2_b - \text{df}_b} \\
\text{TLI}_{(m,b)} &= \frac{\chi^2_b/\text{df}_b - \chi^2_m/\text{df}_m}{\chi^2_b/\text{df}_b - 1} \\
\text{effect size} &= 1 - \frac{\text{misspecification of target model 'm' vs saturated model}}{\text{misspecification baseline model 'b' vs saturated model}}
\end{aligned}
\tag{1}
$$

Note. $r^2_{Y|\mathbf{X}}$ = r-squared, relative reduction in prediction error of Y given predictors $\mathbf{X}$; SS_{error} = error sum of squares, sum of squared differences between each data point y_i and their estimated value $\hat{y}_i$; SS_{total} = total sum of squares, sum of squared differences between each data point y_i and the average $\overline{y}$; NFI = Normed Fit Index; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; with λ_m = noncentrality parameter of a model of interest; λ_b = noncentrality parameter of a baseline model; χ^2_m = chisquare of a model of interest; χ^2_b = chisquare of a baseline model; df_m = degrees of freedom of a model of interest; df_b = degrees of freedom of a baseline model.

Equation 1 shows that each of the measures renorms the misspecification of the target model[13] in terms of the magnitude of the corresponding misspecification of a baseline model. In other words, the baseline model functions as the standard of comparison.

Null model as Baseline

When incremental fit indices are seen in practice, the default baseline model is the null model where all manifest variables are assumed to be uncorrelated. Hence, the core component in the denominator of the incremental fit indices then becomes χ^2_0 , the chisquare of the null model (with degrees of freedom $\text{df}_0 = I(I-1)/2$) and I the number of manifest variables). Under the default maximum likelihood estimator, the latter chisquare reduces to minus the log determinant of the observed correlation matrix $-\log|\mathbf{R}|$ (up to a sample size factor) (for the derivation, see Appendix A). Thus, the standardized metric of the incremental fit indices with null baseline is set by this determinant, a single number representing a generalized measure of variance across your entire dataset. By definition, the determinant of a correlation matrix can be seen geometrically as the volume of the swarm of standardized data points, with $|\mathbf{R}| = 1$ in case of all zero-correlations (corresponding to a 'ball' in a multidimensional plane) and with $|\mathbf{R}| = 0$ for a matrix with perfect linear dependence (a 'ball' flattened along at least one dimension). As Lai and Green (2016) correctly mention, how the determinant changes as a function of a single particular correlation in the correlation matrix is generally opaque. What is clear however, is that the determinant is a real multivariate measure and not simply represents the magnitude of the average correlation, but more something like the magnitude of the dominant correlation (the determinant is equal to the product of eigenvalues of the correlation matrix). Although perhaps not coming across as the most intuitive metric, this determinant does form the core of the standardized metric underlying the popular incremental fit indices in structural equation modeling. Thus, in essence, incremental fit indices are in fact a form of standardized effect size measure and hence, all reservations with respect to interpretations of standardized effect size measures (e.g., Baguley, 2009)

[13] NFI uses absolute misspecification as given by the model's chisquare to the saturated model, CFI uses the model's noncentrality parameter ($\lambda = \chi^2 - \text{df}$), and TLI the ratio of chisquare to degrees of freedom of the model.

also transfer to their interpretation. Such a realization has major implications for the interpretation and use of incremental fit indices, for the theoretical (im)possibility of default universal rules of thumb in their application, and for simulation studies mapping incremental fit indices as if their value is comparable in an absolute sense across any and all conditions. We will drive these points home using a small but illustrative working example centered around the alleged impact of model type and end with a brief discussion elaborating on these implications.

Impact of Model Type?

Reviewing the literature for the differential impact of model type on the behavior of fit indices leads to calls for caution when intending to apply general cutoff criteria across different model types. Considering a range of SEM models, Fan and Sivo (2007) concluded for instance that CFI sampling distributions are sensitive to differences in model type and that this becomes more apparent with increased model misspecification. Similarly, in their famous benchmark study, Hu and Bentler's (1999) simulation results showed differences between simple and more complex structured confirmatory factor analysis models. When comparing simple and approximate simple structure factor models Beauducel and Wittmann (2005) further observed differences among fit indices and what magnitude of secondary loading misspecification they tolerate depending on the rule of thumb applied.

Changing the model type implies changing where the correlation can be found in the model's implied correlation matrix. A one-factor model with equal loadings for 6 observed variables implies a homogeneous correlation all across the 6-by-6 correlation matrix $\mathbf{R}_1$. In contrast, an orthogonal two-factor model with independent cluster structure and equal loadings for each of the three variables per factor implies a block-structured correlation matrix $\mathbf{R}_2$, with 0 correlation on the between-block cells and homogeneous correlation for within-block cells (see Equation 2).

$$\mathbf{R}_1 = \begin{bmatrix} 1 & r & r & r & r & r \\ & 1 & r & r & r & r \\ & & 1 & r & r & r \\ & & & 1 & r & r \\ & & & & 1 & r \\ & & & & & 1 \end{bmatrix} \quad \mathbf{R}_2 = \begin{bmatrix} 1 & r & r & 0 & 0 & 0 \\ & 1 & r & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & 1 & r & r \\ & & & & 1 & r \\ & & & & & 1 \end{bmatrix} \tag{2}$$

So does the behavior of the incremental fit indices really depend on which type of model is being considered?

Three Data-Generating Models

We will consider three data-generating population models M1, M2, and M3. M1 is the aforementioned one-factor model with equal factor loadings, and both M2 and M3 take the form of the aforementioned orthogonal multi-factor model with independent cluster structure and equal factor loadings (see also Equation 2). The difference between models M2 and M3 is that in the former the degree of multivariate dependence as given by the determinant of the model-implied correlation matrix $|\mathbf{R}|$ is equal to that in model M1, whereas in the latter the size of the within-block correlation r_b (or similarly, the square root of the homogeneous factor loading) is equal to that of model M1.

Study Design

Two Simulation Scenarios. To materialize this, consider the following two scenarios where sample size $n = 200$, number of variables $I = 12$, and degrees of freedom df $= 54$. Model M1 was set to have a within-block correlation of $r_b = .40$ resulting in determinant $|\mathbf{R}_1| = .02$ in scenario 1 or a within-block correlation of $r_b = .2$ resulting in determinant $|\mathbf{R}_1| = .27$ in scenario 2. Building from there, Model M2 and M3 were set to contain $B = 3$ independent cluster blocks with $I_b = 4$ indicators per block (i.e., $I = 3 \times 4 = 12$), where for model M2 the within-block correlation r_b was set such that

the determinant of its implied correlation matrix would equal[14] that of model M1 and for model M3 the within-block correlation would simply be set equal to that of model M1. Table 1 summarizes the relevant features of the three data-generating models under both scenarios. Notice that models M2 and M3 also have close to equal average implied correlation ($\bar{r}$). The two scenarios only differ in the amount of correlation present in the data.

Crossfitting: 3×2 ***conditions.*** For each data-generating model – M1, M2, and M3 –, 5000 replicates were generated by simulating sample covariance matrices $\mathbf{S}_m$ drawn from a Wishart distribution with population covariance matrix composed from the $I \times I$ model-implied population correlation matrix $\mathbf{R}_{\text{M}}$ and I population variances sampled from a uniform distribution on the interval $[.75, 2]$. To each replicate, both a one-factor model and an orthogonal three-factor model with independent cluster structure were fitted using maximum likelihood estimation. This cross-fitting procedure results in having a correctly specified and one misspecified model for each data-generating condition. Data simulation and analyses were conducted in R (R Core Team, 2020) through custom scripts in combination with the lavaan package (Rosseel, 2012).

Study Objective. This study design will aid in gaining insight into how different fit indices operationalize "model fit" and in particular how incremental fit indices should be interpreted as a function of their baseline when dealing with both correctly as well as misspecified target models. Note that the sample size and the number of variables are purposively kept constant to exclude potential confounding due to the model size effect on bias in the sample chisquare (e.g., Moshagen, 2012).

[14] Given the homogeneity within and across blocks, the required within-block correlation can be obtained from the fact that the determinant for M2 reduces to the product of the within-block determinants and the relation $|\mathbf{R}_{\text{M1}}| = [1 + (I-1)r_b][1 - r_b]^{I-1}$ (e.g., Graybill, 1983).

Table 1

Study Design: Two Estimated Models Cross-Fitted across Three Data-Generating Models.

		data characteristics							**estimated model**	
data-generating model	df	n	I	B	I_b	r_b	$\lvert\mathbf{R}\rvert$	$\bar{r}$	correctly specified	misspecified
Scenario 1										
M1: one-factor	54	200	12	1	12	.40	.02	.40	one-factor	multi-factor
M2: multi-factor equal $\lvert\mathbf{R}\rvert$	54	200	12	3	4	.53	.02	.14	multi-factor	one-factor
M3: multi-factor equal r_b	54	200	12	3	4	.40	.11	.11	multi-factor	one-factor
Scenario 2										
M1: one-factor	54	200	12	1	12	.20	.27	.20	one-factor	multi-factor
M2: multi-factor equal $\lvert\mathbf{R}\rvert$	54	200	12	3	4	.30	.27	.08	multi-factor	one-factor
M3: multi-factor equal r_b	54	200	12	3	4	.20	.55	.05	multi-factor	one-factor

Note. df degrees of freedom of the data-generating model (i.e., $I(I+1)/2$ sufficient statistics - 24 estimated parameters); sample size n; I number of manifest indicator variables; B number of independent cluster blocks; I_b number of indicators per block; r_b within-block correlation; $\lvert\mathbf{R}\rvert$ determinant of the model-implied population correlation matrix as expression of the degree of multivariate dependence; $\bar{r}$ average model-implied correlation. Non-zero factor loadings in data-generating models are constrained to $\sqrt{r_b}$; estimated models have no such equality constraints. Multi-factor models are orthogonal with an independent cluster structure (cf. blocks).

Results

Correctly-Specified Models

Absolute Fit. Estimating a correctly-specified model results in a sample chisquare statistic χ^2_m of the target model m to the saturated model that has a near-zero value plus some upwards bias that is a function of sample size and the number of variables (Moshagen, 2012). The latter two data characteristics are constant across the three data-generating model conditions, which should result in similar bias magnitude. Hence, if we fit correctly-specified models to data of each of the three data-generating models, we would theoretically expect to see the exact same central chisquare distribution to pop up for the chisquare model fit statistic. Figure 1 illustrates and confirms these theoretical predictions based upon the 5000 replicates. For the chisquare statistic χ^2_m the distribution is indeed equivalent up to minor Monte Carlo variation under each of the three data-generating models when a correctly specified model is fitted, with about 92.5% of the 5000 replications per data-generating model resulting in a non-statistically-significant chisquare statistic (i.e., $\chi^2_m \leq 72.15$, the 5% critical value for df = 54). Withstanding the difference in the amount of data correlation between the scenarios, these results do apply to both scenario 1 and scenario 2.

As a corollary, given that the RMSEA is a function[15] of only the target model's chisquare, degrees of freedom, and sample size, the same equivalence of distributions across the three data generating model conditions also holds for this member of the family of parsimony fit indices. For the RMSEA, equivalent distributions were indeed observed ($M = .015$, and $SD = .016$, across all models) with values for 95% of the replicates falling in the interval $[.00, .05]$.

Incremental Fit. The same equivalence of distribution across all of the data-generating models does not apply for the incremental fit indices, neither across scenarios nor within a scenario. For instance, although the CFI is on average as high as .99 in scenario 1, only the distribution under M1 and M2 is similar, but characterized by heavier

[15] Root Mean Square Error of Approximation: $\text{RMSEA} = \frac{\sqrt{\chi^2_m - \text{df}}}{\sqrt{\text{df}(n-1)}}$

tails in the case of M3 with a lower adjacent[16] CFI value of .95 and a minimum of .89 compared to a lower adjacent CFI value of .97 and a minimum of .94 for both M1 and M2 (see Figure 2). When applying the commonly adopted .95 rule of thumb, this would result in assessing 4% of the correctly specified M3 models as showing non-acceptable fit to the data, compared to close to 0% for M1 and M2. With the lower amount of data correlation in scenario 2, this pattern of findings reproduces but with larger sampling variation in CFI values under all models, resulting in assessing 14 and 15% of replicates under M1 and M2 as non-acceptable according to the CFI $\geq$ .95 rule of thumb with lower adjacent CFI values of .92 and .92 and minima of .81 and .84 compared to 29% of non-acceptably fitting replicates under M3 with lower adjacent CFI value of .85 and minimum of .70.

The equivalence of CFI distributions under M1 and M2 is due to both having a similar CFI numerator (i.e., based on the χ^2_m of a correctly specified model with the same degrees of freedom and equal sample size) and denominators with similar baseline value based on $\chi^2_0 = -\log(|\mathbf{R}|)(n-1)$, reflecting the degree of multivariate dependence in the data (see Table 1). In contrast, M3 also has a similar numerator but has a smaller baseline which makes it harder to differentiate between the model of interest and the baseline model, resulting in the heavier CFI tails under M3. In scenario 2, with the amount of data correlation being lower compared to scenario 1, the smaller baseline for all three data-generating conditions amplifies the variation in CFI including numerous observed values that are not even in line with the common rule of thumb guidelines for correctly specified models.

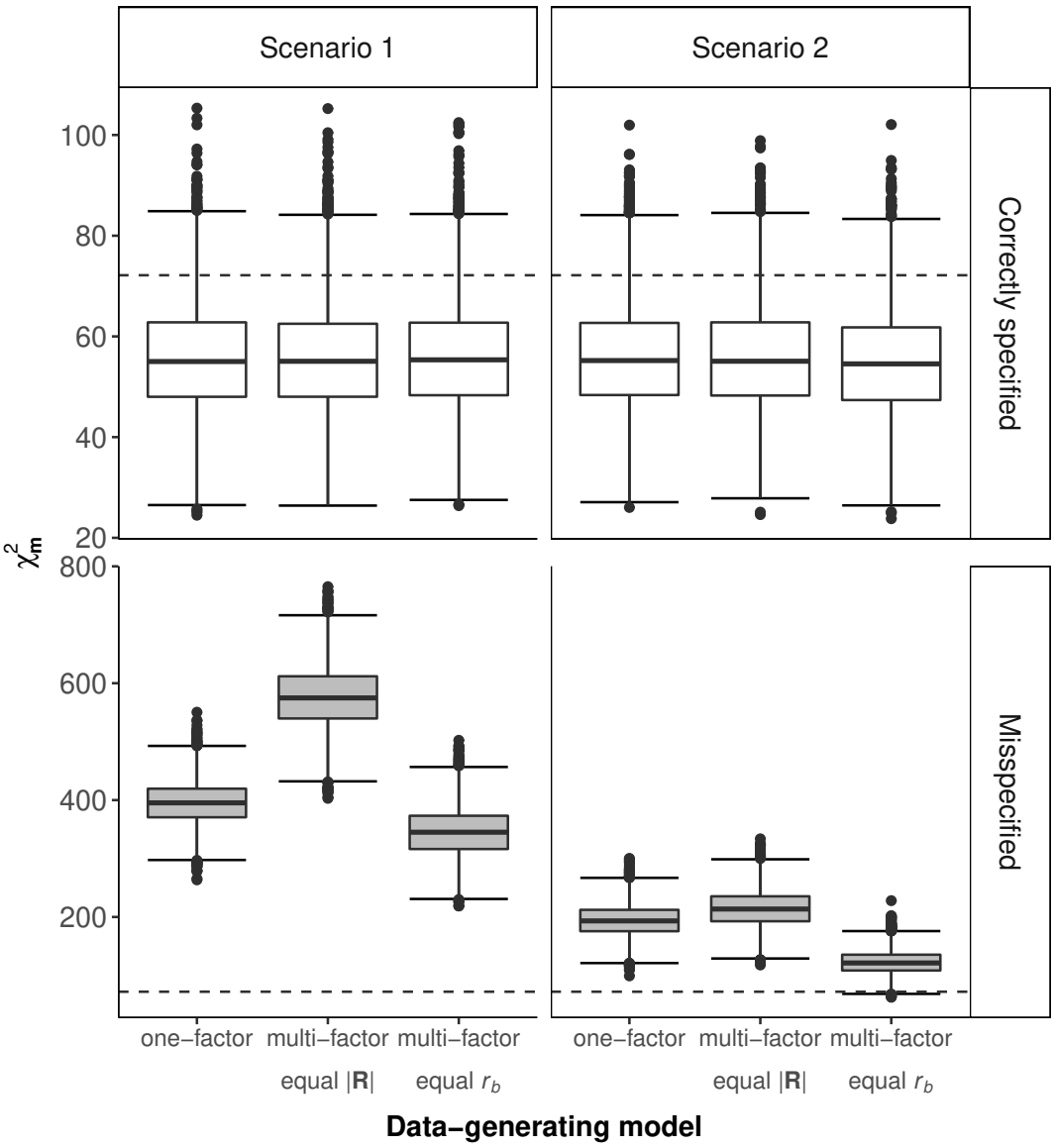
Trends similar to CFI's apply to other incremental fit indices, but the increased sampling variance in scenario 2 and the heavier tail under M3 now apply to both the lower and upper tail of the distribution as both TLI and NFI are, in contrast to CFI, not restricted to an upper bound of 1. In sum, these trends show that the degree of multivariate dependence plays an integral part in the observed differences in CFI distribution, the performance of common rules of thumb, and variation in the sampling distribution

[16]Lower Adjacent Value: the smallest observation above or equal to the lower inner fence (i.e., first quartile minus the interquartile range) in a boxplot.

of the incremental fit indices by changes in the baseline for comparison, regardless of changes in model type.

Figure 1

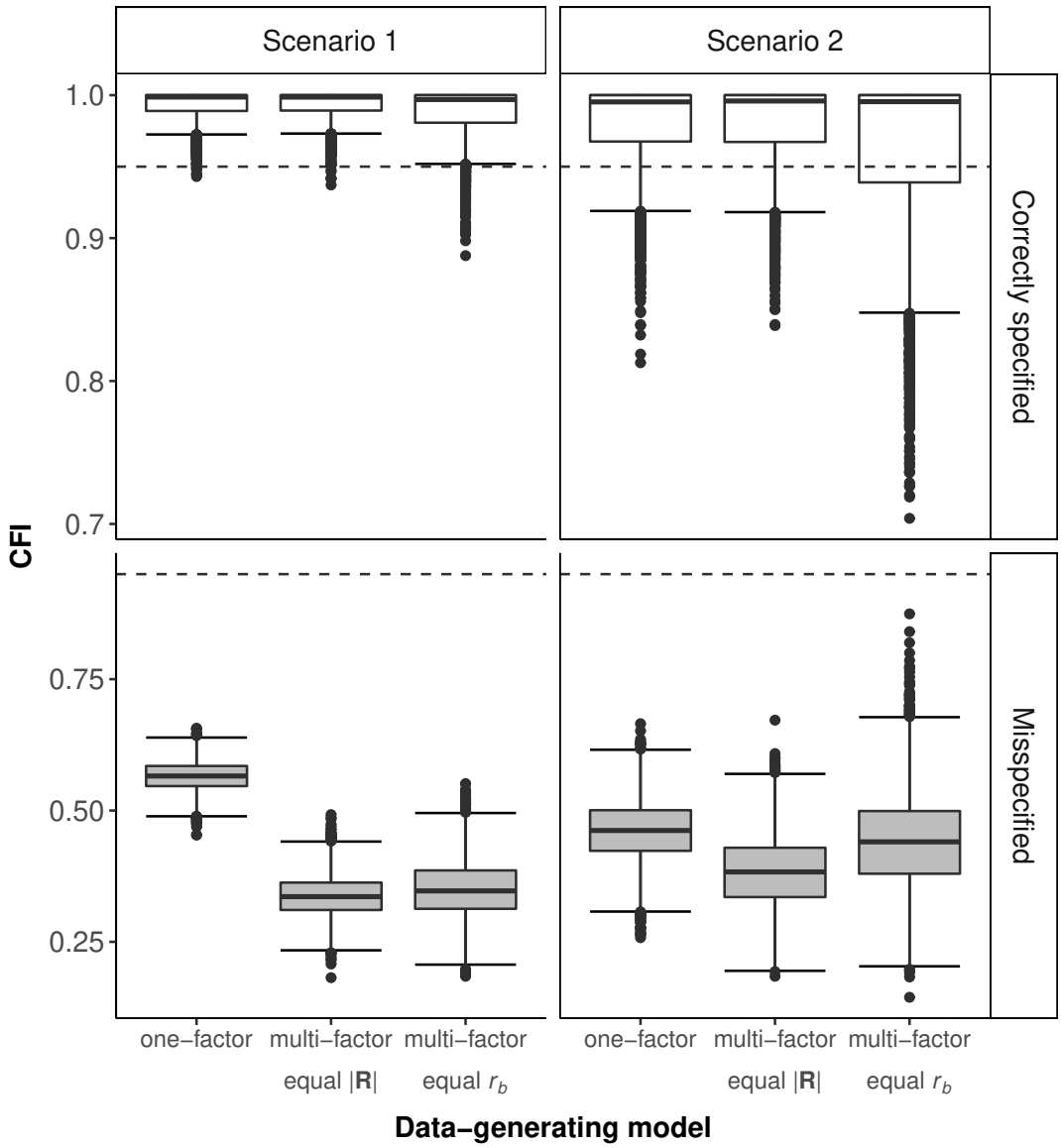
χ^2-distribution under Correctly- and Misspecified Models.



Note. The dotted line corresponds to the 5% critical value $\chi^2_{\text{df}=54} = 72.15$. With χ^2_m chisquare of the model of interest; $|\mathbf{R}|$ determinant of the model-implied population correlation matrix as expression of the degree of multivariate dependence; r_b within-block correlation. In both scenarios, sample size $n = 200$. The misspecified model is a multi-factor model for the one-factor model, and vice versa (see also Table 1).

Figure 2

CFI-distribution under Correctly- and Misspecified Models.



Note. The dotted line corresponds to the commonly adopted .95 CFI rule of thumb. With $|\mathbf{R}|$ determinant of the model-implied population correlation matrix as expression of the degree of multivariate dependence; r_b within-block correlation. In both scenarios, sample size $n = 200$. The misspecified model is a multi-factor model for the one-factor model, and vice versa (see also Table 1).

Misspecified Models

Absolute Fit. To consider misspecified models we fitted an orthogonal multi-factor model with independent cluster structure to data generated under M1(one-factor model with equal loadings), and a one-factor model to data generated under models M2 and M3 (orthogonal multi-factor models with independent cluster structure and equal loadings) (see Table 1). The resulting misspecified models, denoted by M1′, M2′, and M3′, have absolute misspecification as measured by $\chi^2_{\text{M}'}$ of a similar magnitude for M1′ ($\chi^2_{\text{M1}'} : M = 395$) and M3′ ($\chi^2_{\text{M3}'} : M = 345$), but about one and a half times larger misspecification for M2′ ($\chi^2_{\text{M2}'} : M = 576$) in scenario 1 (see left panel of Figure 1). For scenario 2, the chisquare values reduced as they were bounded by the lower amount of data correlation. Differences in chisquare values between models were more compressed with M2′ still the lowest ($\chi^2_{\text{M2}'} : M = 215$), now more closely followed by M1′ ($\chi^2_{\text{M1}'} : M = 194$), but still a good distance to M3′ ($\chi^2_{\text{M3}'} : M = 123$). Under each data-generating condition, the misspecified model resulted in rejecting the chisquare test of equal fit to the saturated model for almost exactly 100% of the replicates.

In terms of parsimony-adjusted absolute fit as measured by the RMSEA the chisquare values translated to an average RMSEA of .18, .22, and .16 for M1′ to M3′, respectively under scenario 1 and reduced to about half those values in scenario 2 (with the lower amounts of data correlation) to an average RMSEA of .11, .12, and .08 for M′1 to $M3'$, respectively. As a consequence, applying the popular rule of thumb of RMSEA below .08 in scenario 2, would wrongly assess M3′ as an acceptable fitting model for 52% of the replicates.

Incremental Fit. When looking at incremental fit indices, the magnitude pattern of misspecification shifts compared to the absolute fit indices. For scenario 1, M1′ results in higher incremental fit values than both M2′ and M3′, and the latter two being equal in size (e.g., see left panel of Figure 2; CFI: $M = .56, .34, \& .35$, respectively). The magnitude of CFI values seem to imply that M1′ is the least misspecified, and M2′ and M3′ the most misspecified among the three models (i.e., M1′ < (M2′, M3′)). In contrast, the magnitude order of χ^2 indicated M3′ and M1′ to be the least misspecified and M2′ the

most misspecified (i.e., $(\text{M3}', \text{M1}') < \text{M2}'$).

How can these irreconcilable differences in assessment of the magnitude of model misspecification or model fit be explained? Well, M2′ and M3′ are both one-factor models wrongly fitted to data from a multi-factor, whereas M1′ is a one-factor model wrongly fitted to a multi-factor model, and hence the obvious culprit for these CFI differences must be the difference in model type? Yet, by making such an inference, we would be caught off base by not accounting for the nature of incremental fit indices and applicable baseline differences. Whereas chisquare and RMSEA are more absolute measures of misspecification (raw or parsimony-adjusted), the incremental fit indices are relative measures with the amount of absolute misspecification under the baseline model as standardized metric.

Although M1′ and M3′ have similar χ^2_m values (i.e., basis of the numerator in incremental fit indices) for the target model, the baseline model in case of data generated under M1 has a larger χ^2_0 value than under M3, leading to $\text{M1}' > \text{M3}'$ in CFI value. Hence, relatively speaking in CFI terms, the model M1′ is less badly misspecified compared to the baseline model for data from M1 than is the model M3′ compared to the baseline model for data from M3. Furthermore, a large χ^2_m is divided by a large baseline chisquare in M2′'s case and that happens to result in a CFI value similar to dividing a smaller target model chisquare by a smaller baseline chisquare in M3′'s case.

In other words, by trying to compare CFI values across models fitted on different datasets, we are looking at values on different standardized metrics as if they were comparable in an absolute sense and are now essentially ignoring the fact that we are comparing different units, literally, percentages of different baseline totals. Note that the same reasoning applies to scenario 2, although the pattern of incomparable values across models differs.

Implications

What all of this hopefully clarifies, is that we should resist the temptation to interpret values of incremental fit indices as if they were comparable in an absolute sense because they are only comparable in the case that their baselines are comparable at the data level (e.g., for CFI the noncentrality parameter of the baseline model λ_b) and not at the mere

conceptual level (i.e., it is not sufficient that both baseline models are the null model). Such a realization has major implications for the interpretation and use of incremental fit indices, for the theoretical (im)possibility of default universal rules of thumb in their application, and for simulation studies mapping incremental fit indices as if their value is comparable in an absolute sense across any and all conditions.

Theoretical (im)possibility of default universal rules of thumb. The fact that, in contrast to absolute fit indices, the distribution of incremental fit indices even varies across correctly specified models of equal degrees of freedom and with equal sample size (cf. compare top panels of Figure 1 and 2) implies that adopting a universally applicable general cutoff rule of thumb might not be the most fruitful idea for incremental fit indices. This is not illogical. When placing a target model of interest along a relatively small baseline-to-saturated continuum as in scenario 2 (i.e., in case of a null baseline reflected by a small value of $|\mathbf{R}|$), it will always be closely fitting in absolute sense to both the baseline and the saturated model, as all models are relatively alike. This implies that model differentiation is unreliable in case of a small baseline, incremental fit indices become less informative, and placing a fixed threshold for a universal rule of thumb becomes nigh impossible (see also, van Laar & Braeken, 2021). The opposite holds in case of a large baseline.

Baseline differences as confounder in simulation studies. Realizing the non-ignorability of the baseline not only applies to SEM practitioners in the field, but also to past and future simulation studies where values of CFI, TLI, and family are simply tracked regardless of baseline comparability, leading to an obvious confound in their design, comparative statements, and recommendations for relative fit measures. In general, we argue that to further advance our joint understanding of goodness-of-fit measures and their behavior in practice within the SEM field, we need more theoretically driven and less exploratory simulation studies. The latter are too much at risk of making conclusions based on artifacts in the chosen design factors. One element in an exploratory study design potentially impacts many other easily overlooked confounding factors under the hood.

Determinant not average pairwise correlation. In SEM, the relative model discrepancy to the null baseline, in incremental fit indices stemming from the chisquare, does not take into account the location of the correlation in the data that your model fails to capture nor does it encode how much of the average correlation your model has captured, but instead it encodes how much of the dominant correlation (i.e., the determinant is the product of eigenvalues of $\mathbf{R}$) in the data the model captures. The central role of this determinant should revive some interest in understanding classic measures of multivariate statistics (e.g., Anderson, 1958) to further our understanding of more modern SEM practices. Whereas people in practice often already find it hard to interpret the absolute magnitude of a variance, it is fair to say that even fewer people have a good intuition about what a large or small determinant (i.e., generalized variance) is for their dataset.

Explicit reporting of this determinant[17] $|\mathbf{R}|$ would help in gaining some intuition on common reference values for this data characteristic in your field of application and eventually allow for a better interpretation of the relative and absolute magnitude of incremental fit indices with the null model as baseline, even across datasets. By making the presence of the core components of the null baseline explicit in the reporting, the need to take it into account when interpreting incremental fit indices also becomes explicit and non-ignorable (for a small reporting example and corresponding R syntax, see Appendix B).

Note that this rationale with respect to interpretation is not necessarily limited to incremental fit indices. There are other fit measures it could be extended to, even though their baseline for meaningful interpretation might be different. For example, the Standardized Root Mean Square Residual (SRMR) fit index is also a standardized measure, and hence similar interpretation and practice recommendations should apply here. The core difference to the incremental fit indices considered here is that SRMR is residual-based and not chisquare-based. As a consequence, SRMR's metric is not a function of the determinant but of the average observed pairwise correlation $\bar{r}$. In our small working ex-

[17] The determinant of the observed correlation matrix $|\mathbf{R}|$ can be easily extracted from default software. For R::lavaan, this can be extracted from the fitted model, in the example syntax stored in an object labeled "fit": `exp(-(fitmeasures(fit)[["baseline.chisq"]]/(inspect(fit, "nobs")-1)))`.

ample, the SRMR distributions for correctly specified models would indeed be equivalent under M2 and M3, but not under M1, as the former two have equal average correlation values but differ from M1's average correlation value. In other words, SRMR evaluates fit in an average pairwise dependence sense, in contrast to incremental fit indices who evaluate fit in terms of multivariate dependence (i.e., $|\mathbf{R}|$). Realizing this difference helps in understanding what type of model fit each fit index codes for. Yet for all standardized fit indices the base for interpretation needs to be taken into consideration and absolute value judgements across any and all conditions are not recommended.

Transferability. Although our working example is rather small, the underlying principles should apply across different scenarios. Even when extending the scope to models involving mean-structure, other baselines than the null model (e.g., Rigdon, 1998; Widaman & Thompson, 2003), non-normality corrections, or different estimation methods, the formulas for numerator and denominator and the character and metric of the baseline might slightly change, but the practical implication that incremental fit indices are only large or small in comparison to a data-specific baseline, and not a universal threshold reference value, will never disappear.

Practical Recommendation

CFI, TLI, and the entire incremental fit family are improperly treated in the current all too common one-off model assessment approach where they are seen as an absolute value in a mere search for a model adequacy threshold number. Instead, in a reasoned model comparison strategy, incremental fit indices are a useful benchmark metric for interpreting the relative magnitude (i.e., effect size) of the paths in which the set of competing models differ. Thus, we should strive to use incremental fit indices (Bentler & Bonett, 1980) as intended, to evaluate the relevance of cumulative theoretically motivated model restrictions in terms of % reduction in absolute misspecification as measured by the adopted baseline model.

References

Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in cfa based on data with slightly distorted simple structure. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 41–75.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509–529.

Graybill, F. A. (1983). *Matrices with Applications in Statistics* (2nd). Wadsworth International Group.

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution*. University of Chicago Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Jackson, D. L., Gillapsy, J. A., & Purch-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23.

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*(2-3), 220–239.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391–410.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 86–98.

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*, 1148–1172.

R Core Team. (2020). R: A language and environment for statistical computing.

Raykov, T. (2005). Bias-corrected estimation of noncentrality parameters of covariance structure models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 120–129.

Rigdon, E. E. (1998). The equal correlation baseline model for comparative fit assessment in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 63–77.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

van Laar, S., & Braeken, J. (2021). Understanding the comparative fit index: It's all about the base! *Practical Assessment, Research, and Evaluation*, *26*, Article 26. https://doi.org/10.7275/23663996

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1).

Appendix A
The Chisquare of the Null Model is Proportional to Minus the log Determinant of the Observed Correlation Matrix ($\chi_0^2 \propto -\log|\mathbf{R}|$)

For the default CFI with a null model as baseline, the value of CFI is based on the ratio of misspecification between the model of interest and the null model:

$$\text{CFI}_{(m,0)} = 1 - \frac{\lambda_m}{\lambda_0} = 1 - \frac{\chi_m^2 - \text{df}_\text{m}}{\chi_0^2 - \text{df}_0} \tag{1}$$

Equation 1 shows how the misspecification of both models would be estimated by their noncentrality parameter, being the difference between the model's chisquare of exact fit against the saturated model and the model's degrees of freedom. Focusing on the denominator of CFI, the standard of comparison, and hence the core component of CFI, is then $\chi_0^2 - \text{df}_0$, the chisquare of the null model with degrees of freedom $\text{df}_0 = I(I-1)/2$) and I the number of manifest variables. The chisquare value of the null model can be rewritten as the product of the sample size n and the minimum value F_0 of the used fit function to estimate the models (i.e., $\chi_0^2 = F_0(n-1)$).

Under maximum likelihood estimation, F_0 is a function of the discrepancy between the model-implied variance-covariance matrix $\hat{\mathbf{\Sigma}}_0$ under the null model and the observed variance-covariance matrix $\mathbf{S}$ (e.g., Bollen, 1989), where $\text{tr}(\mathbf{X})$ and $|\mathbf{X}|$ are respectively the trace and determinant of a matrix $\mathbf{X}$ (cf. Equation 2).

$$\begin{aligned} F_0 &= \log|\hat{\mathbf{\Sigma}}_0| - \log|\mathbf{S}| + \text{tr}(\mathbf{S}\hat{\mathbf{\Sigma}}_0^{-1}) - I && (2)\\ &= \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| + \text{tr}(\mathbf{S}\mathbf{diag}(\mathbf{S})^{-1}) - I && (3)\\ &= \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| + I - I && (4) \end{aligned}$$

Key in getting to the expression for the chisquare χ_0^2 for the null model as mentioned in the main text (i.e., $-\log|\mathbf{R}|$ up to a sample size factor), is that the minimal fit value F_0 for the null model can be further simplified using the fact that the model-implied covariance matrix under the null model comes down to a diagonal matrix $\mathbf{diag}(\mathbf{S})$ with

the observed variances on the diagonal (cf. Equation 3). This results in $\mathbf{S}\hat{\mathbf{\Sigma}}_0^{-1}$ leading to a matrix with all ones on the diagonal such that the trace equals the number of observed variables I and cancels out the subsequent $-I$ term in the expression for F_0 (cf. Equation 4).

Using the fact that the determinant of a matrix product can be split into products of determinants, each of the remaining two log determinants can be written out given that a variance-covariance matrix $\mathbf{S}$ is a multiplicative function of a corresponding correlation matrix $\mathbf{R}$ and an inverse diagonal matrix with standard deviations on the diagonal. Thus we have

$$\log|\mathbf{S}| = \log|\sqrt{\mathbf{diag}(\mathbf{S})}\ \mathbf{R}\ \sqrt{\mathbf{diag}(\mathbf{S})}| \tag{5}$$

$$= \log|\sqrt{\mathbf{diag}(\mathbf{S})}| + \log|\mathbf{R}| + \log|\sqrt{\mathbf{diag}(\mathbf{S})}| \tag{6}$$

$$= \log\prod_{j=1}^{I}\sqrt{S_{jj}} + \log|\mathbf{R}| + \log\prod_{j=1}^{I}\sqrt{S_{jj}} \tag{7}$$

$$= \log\prod_{j=1}^{I} S_{jj} + \log|\mathbf{R}| \tag{8}$$

and

$$\log|\mathbf{diag}(\mathbf{S})| = \log|\sqrt{\mathbf{diag}(\mathbf{S})}\ \mathbf{I}\ \sqrt{\mathbf{diag}(\mathbf{S})}| \tag{9}$$

$$= \log\prod_{j=1}^{I} S_{jj} + 0 \tag{10}$$

where Equation 10 makes use of the fact that the correlation matrix of a diagonal variance-covariance matrix is an identity matrix $\mathbf{I}$ which determinant is exactly equal to 1.

The re-expressions of the log determinant terms in Equations 8 and 10 allow to simplify

the expression for F_0 further by elimination

$$F_0 = \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| \tag{11}$$

$$= \log\prod_{j=1}^{I} S_{jj} - \log\prod_{j=1}^{I} S_{jj} - \log|\mathbf{R}| \tag{12}$$

$$= -\log|\mathbf{R}| \tag{13}$$

such that the denominator of CFI under the null model comes down to

$$\lambda_0 = \chi_0^2 - \mathrm{df}_0 = F_0(n-1) - \mathrm{df}_0 = -\log|\mathbf{R}|(n-1) - I(I-1)/2$$

Appendix B
Mini Example to Report Incremental Fit Indices with Corresponding R::lavaan Code

The SEM-package lavaan (Rosseel, 2012) in the free statistical software environment R (R Core Team, 2020) contains a built-in dataset variant of a well-known study by Holzinger and Swineford (1939). Situated in the study of human intelligence, the dataset contains scores on $I = 9$ cognitive ability tests (named variables x1 to x9 in the dataset) for $n = 301$ children. In practice, we advocate the use of incremental fit indices as intended, that is in the context of a reasoned model comparison strategy. Without being able to elaborate too much on specifics of the field nor dataset, we can still posit a fairly realistic set of competing models for the current context as an example in case, but with a somewhat simplified underlying theoretical motivation.

Set of competing models

A historical finding in the intelligence field is that cognitive tests, no matter their specifics, tend to positively correlate within a general population. This would correspond to a so-called positive manifold as reflected by the appropriateness of a one-factor model M_1 covering all 9 tests. Yet the 9 cognitive tests are said to have some common structural elements, with the first three tests being more the visuo-spatial type, the second three tests being more verbal-text related, and the last three more speed-based. It would be natural to expect these clusters to also be reflected in the strength of the intercorrelations between the test scores. Yet how this exactly surfaces, one can disagree about. Model M_{2a} considers three orthogonal factors, one for each of the three independent item clusters. This models also implies that intercorrelations among cognitive tests of different type would be negligible. Model M_{2b} with three oblique factors, one for each of the three independent item clusters, offers a less strict perspective by implying that the dominant correlation is within the clusters, but allowing some correlation between clusters. A final model M_3 covers all bases by considering a one factor model but with residual correlations among cognitive tests within the same cluster.

The model comparison strategy further involves locating the set of competing models

within a within a continuum of models from the worst fitting baseline null model M_0 to the perfect fitting saturated model M_S (Bentler & Bonett, 1980). The results are summarized and reported in Table B1. Corresponding R-code for the models and results can be found at https:// osf. io/ f6jnm/? view_ only=e367c654fbcd47248667e170442592c3.

Results

We can see that accounting for the implied positive manifold or the expectation that performance on cognitive tests correlate by default as in M_1, reduces the specification error in terms of the multivariate degree of dependence present in the data by 68% ($CFI_{(M_1,M_0)} = .68$). Note how in a linear regression, one would generally already be quite happy with such a relative reduction in predictor error variance as implied by an r-square of .68. Although the model does not fit close to perfect in an absolute sense, there is sufficient to disregard the implied uncorrelatedness of test performances by model M_0. At the same time we see that ignoring the positive manifold idea and only accounting for the cluster structure as in M_{2a} leads to a reduction of 86%, an additional 18% reduction in misspecification error of the multivariate dependence compared to M_1. This finding implies that the dominant correlation structure in the dataset is indeed between cognitive tests of the same type. Allowing for some structural intercorrelation between the clusters does reduce misspecification somewhat more with an additional 7%, amount to a total reduction of 93% under M_{2b}. Further covering both perspectives with a structural positive manifold and variable residual interdependence within a cluster, as in M_3, leads to additional reduction of 5% in misspecification error, bringing us, relatively speaking within 2% ($CFI_{(M_3,M_0)}$=.98), in the immediate neighbourhood of the 'perfect' yet unstructured saturated model M_S.

Simplified conclusion

In a reasoned model comparison strategy, incremental fit indices are a useful benchmark metric for interpreting the relative magnitude (i.e., effect size) of the paths in which the set of competing models differ. Together these results imply that the paths corresponding to the cluster structure in terms of cognitive test type are clearly pronounced, but that not all cognitive tests adhere to a strict clustering and still intercorrelate across

type as well. When inspecting the correlation matrix among the cognitive tests, you can also clearly see the cluster structure, but also the first visuo-spatial test correlating with the majority of other tests regardless of type.

Notice that in our assessments there was no explicit need of rules of thumbs nor a focus on absolute fit, as in the end interest would be more about strengths of the different perspectives as put forward by the competing models. This appears to us as more healthy approach than a one-off model assessment approach using binary conclusions based on indefensible universal rules of thumb (e.g. CFI$\geq$ .95). For one specific study, the value of reporting determinant, sample size, and number of variables might not be directly apparent. Yet, these summary statistics would become relevant once you intend to compare incremental fit indices across different studies to assess whether one study's 93% is comparable to another study's 95%, and for general meta-analysis purposes. Hence, we recommend including these by default, and doing so is luckily extremely simple in practice.

Table B1

Model Comparison Results for the Set of Competing Models.

	M_0	M_1	M_{2a}	M_{2b}	M_3	M_S
χ^2	919	312	154	85	35	0
df	36	27	27	24	18	0
p	<.001	<.001	<.001	<.001	0.010	1.000
λ	883	285	127	61	17	0
$\text{CFI}_{(m,0)}$	0.00	0.68	0.86	0.93	0.98	1.00
$\lvert\mathbf{R}\rvert = 0.047$, $n = 301$, $I = 9$.						

Note. λ = noncentrality for the estimated model (i.e., $\lambda = \chi^2 - \text{df}$); CFI = CFI value for estimated model (i.e., $\text{CFI}_{(m,0)} = 1 - \frac{\lambda_m}{\lambda_0}$); $|\mathbf{R}|$ = determinant of the observed correlation matrix (i.e., the degree of multivariate dependence); n = sample size; I = number of items. Here one would typically further clarify the model specifications and highlight the differences among the models. Yet, to keep the appendix compact see text above.

6 Article 3: Prevalence & Impact

van Laar, S., & Braeken, J. (2022b). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501. https://doi.org/10.1111/jedm.12317

Random responders in the TIMSS 2015 student questionnaire: A threat to validity?

The low-stakes character of international large-scale educational assessments implies that a participating student might at times provide unrelated answers as if s/he was not even reading the items and choosing a response option randomly throughout. Depending on the severity of this invalid response behavior, interpretations of the assessment results are at risk of being invalidated. Not much is known about the prevalence nor impact of such *random responders* in the context of international large-scale educational assessments. Following a mixture item response theory (IRT) approach, an initial investigation of both issues is conducted for the Confidence in and Value of Mathematics/Science scales in the Trends in International Mathematics and Science Study (TIMSS) 2015 student questionnaire. We end with a call to facilitate further mapping of invalid response behavior in this context by the inclusion of instructed response items and survey completion speed indicators in the assessments and a habit of sensitivity checks in all secondary data studies.

International large-scale educational assessments are used to describe, compare, and monitor student achievement in different educational domains and across different countries. In general, by providing information on contextual factors as provided by the student questionnaire and staff survey with respect to the learning processes that can be related to the student outcomes on the achievement tests, these assessments aim to inform curriculum and education policy to improve learning (e.g., International Association for the Evaluation of Educational Achievement [IEA]: Trends in International Mathematics and Science Study [TIMSS] and Organisation for Economic Co-operation and Development [OECD]: Programme for International Student Assessment [PISA]). The impact of international large-scale educational assessments on policymaking has been widely documented (e.g., Hopfenbeck et al., 2018). The amount of participants together with the inclusion of different countries and repeated assessments over time makes these large-scale educational assessments an extensive source of potentially valuable information in

national and international contexts. This treasure trove of information is also recognized and exploited in an ever-increasing number of studies that use data collected in these assessments to answer a wide variety of research questions (for a review, see Hopfenbeck et al., 2018).

Although the availability of such extensive amounts of information in publicly available databases sounds very promising in terms of research opportunities, the validity of the conclusions we draw is of course dependent on the quality of the assessment and the corresponding data. To ensure the highest quality, both the central as well as national institutes behind these international large-scale educational assessments invest a lot of time, effort, and resources in the design, data collection, analysis, and preparation of the data and its documentation for the databases of these assessments. Yet, one factor that organizing parties logically lack control over is the actual response behavior of the students participating in these international large-scale educational assessments.

This lack of control with respect to actual response behavior is of course not unique to the survey part specifically or international large-scale educational assessments in general, yet this type of assessment can be expected to be extra susceptible to invalid response behavior due to their low-stakes character and generally young target population. In absence of direct consequences or other incentives, the students that are required to fill in these international large-scale educational assessments, might not always respond accurately or thoughtfully (e.g., Curran, 2016; Eklöf, 2010), but instead these young adolescents might respond without meaningful reference to the test items or survey questions (Berry et al., 1992). Regardless of whether it is due to insufficient effort, thoughtlessness, or lack of seriousness, such behavior would make that responses no longer accurately reflect knowledge, abilities, or opinions related to the assessment content, but are being distorted by contextual factors (e.g., Cronbach, 1950; Messick, 1984). Depending on how prevalent this invalid response behavior is across the educational assessments, any type of inference based on these assessments, either research conclusions or policy changes, is at risk of being invalidated in spirit of the old adagium “garbage in, garbage out”.

Although there are a lot of conjectures about invalid response behavior and its conse-

quences for the validity of results in the international large-scale educational assessments, the corresponding evidence base on its prevalence and its impact on inferences is rather scarce. This is rather striking and hugely unfortunate as invalid response behavior is directly linked to data quality, and data quality is at the very essence of all secondary analyses that are run on the datasets of these assessments. We see three core challenges contributing to the non-ideal current status that are all three connected to the inherently exploratory character of the act of data quality monitoring for invalid response behavior.

First, invalid response behavior is a broad concept giving rise to a whole range of at times inconsistently used definitions and terminology such as insufficient effort, disengaged, careless, unmotivated, random, inconsistent, non-contingent, variable, or content-independent responding (cf., Huang et al., 2012) and in addition, it invigorates debate regarding plausible underlying mechanisms. Within the context of international large-scale educational assessments, these plausible underlying mechanisms have attracted research attention, whereas the basic identification and *prevalence* of invalid response behavior has not gained as much traction.

For example, students' self-reported test motivation (Eklöf, 2007) or effort (Butler & Adams, 2007) has been measured using custom survey scales (often added as a national option) and correlated to the achievement scores on the cognitive test component of the assessments for a specific country or set of countries ($r \approx .25$, see e.g., Eklöf et al., 2014; Hopfenbeck & Kjærnsli, 2016). Yet, how does self-reported motivation/effort translate into actual invalid responses given on the overall assessment or help in assessing their prevalence? In a study by Eklöf et al. (2014) 37% of students reported they did not do their very best on the test, whereas 81% of the students agreed they could have worked harder on the test. Students' self-reported perceived intentions are a complex indirect measure and do not necessarily have a simple one-to-one translation into the actual prevalence of invalid response behavior demonstrated in the assessment. Given the scope of the international large-scale educational assessments, one can also wonder whether self-reported generic motivation/effort is a constant across the assessment and equally applicable to all parts of the cognitive achievement test component and survey

component of the assessment. Furthermore, the self-report approach has a circular character to it considering that it comes down to asking students how motivated they are in a questionnaire for which they might not be motivated to begin with. Thus, although underlying mechanisms are of definite interest to potentially intervene and remedy invalid response behavior, it is also a challenging subject and not directly fruitful for answering the more primordial question of how *prevalent* this invalid response behavior is and what *impact* it has on inferences based on the international large-scale educational assessments.

Second, a rather loose and unsystematic link exists between the definition of invalid response behavior adopted in a study, mostly in terms of theorized unobservable intentions of participants to the assessment, and the subsequently chosen operationalized measure of invalid response behavior. Different types of post-hoc diagnostic methods are available for the detection of invalid response patterns, including response time analyses, outlier analyses, individual consistency measures, and person-fit statistics (for an overview, see e.g., Curran, 2016; Meade & Craig, 2012). Yet most of the detection tools are either generic and hence not clearly linked to a specific definition of a type of invalid response behavior or when they are more specifically aimed at the detection of a set of invalid response patterns, they lack specificity in the sense that they also pick up other response patterns (e.g., Hong et al., 2020; Karabatsos, 2003). The difficulty with most methods lies with the decision of, to some extent, arbitrary cutoff values to distinguish between individuals showing in/valid response behavior (Curran, 2016; Hong et al., 2020). Curran, Kotbra, and Denison (as cited in Meade & Craig, 2012) showed that the prevalence of invalid response behavior varied from 5% to 50% depending on how invalid response behavior was operationalized (for similar results Beck et al., 2019; Huang et al., 2012). In the context of international large-scale educational assessments, the validity study by Hopfenbeck and Maul (2011) on the PISA 2006 learning strategy scale provides for instance a good discussion of the potential value, but also the interpretation difficulties when employing invalid-response detection techniques to further our understanding of the interaction between students and these assessments. One thing is clear; When trying to assess and identify invalid response behavior, it is important to pay attention to the

specific methodology used as reported results highly vary as a function of these operationalizations, making the definition of invalid response behavior more operationally than conceptually grounded.

Third, the sheer size and diversity of the assessments make that data quality monitoring in international large-scale assessments in education has much more ground to cover than a similar endeavor in research using a personality inventory for instance. Preventive measures taken by TIMSS that will contribute to data quality, are the implemented booklet design and pauses in-between assessment parts (Mullis & Martin, 2013) such that test burden and testing time are kept within boundaries for the students. Standard data management procedures are also in place to ensure the highest quality and consistency for the TIMSS data files (e.g., out-of-range values; multiple responses; recoding logically invalid or missing responses). Yet, to the best of our knowledge TIMSS does not focus on underlying response processes (e.g., motivation, effort, or willingness to cooperate) nor do their manuals (e.g., Martin et al., 2016; Mullis & Martin, 2013) mention the use of different detection methods for flagging invalid response patterns or behavior for the student questionnaire. One can argue whether this data quality monitoring and assessment is the responsibility of the organizing party, of the research community, or up to individual researchers as part of a sensitivity check for their specific study. The Standards for Educational and Psychological Testing (AERA, 2014) do call on test developers, administrators, and researchers alike to document sources of construct-irrelevant variance to provide further context for test-based analyses and inferences.

This Study

In what follows, we will investigate the prevalence of so-called *random responders* and their impact on inferences related to the TIMSS student questionnaire. This survey part of the international large-scale educational assessments typically receives both less investment and attention when compared to achievement tests, although both are equally important to put results in context (e.g., Rutkowski & Rutkowski, 2010). After clarifying why we chose to focus on this particular type of invalid response behavior, we outline the adopted mixture item response theory (IRT) approach to detect random responders and

how it operationalizes "random" and hence defines the type of invalid response behavior it can study. Using this mixture IRT approach, we will identify the students which are likely engaging in random response behavior, resulting in a prevalence estimate of random responders. An impact assessment is run by means of a sensitivity study comparing the results of analyses with and without the identified random responders. The approach provides educational researchers with a model-based procedure to chart the issue of random responding in line with the Standards for Educational and Psychological Testing (AERA, 2014). We will illustrate the method for two student survey scales in each TIMSS domain – "Value of" and "Confidence in" Mathematics/Science – for a subset of five countries. This study design allows to make a tentative exploration of where the biggest source of variation in prevalence and impact of random responders lies: At the country side or the scale side.

Random Responders

Here, we explicitly focus on so-called "*random responders*". Random responding is defined as a response set (Cronbach, 1950) in which a person provides mostly unrelated responses to a survey scale of interrelated items as if s/he was not even reading the particular items and choosing a response option randomly throughout.

Risk factors. Meade and Craig (2012) highlight four risk factors that contribute to the occurrence of such invalid response behavior: limited respondent interest, survey duration/length, lack of personalization/large social distance, and environmental distractions. Unfortunately, the international large-scale educational assessments can be considered to tick all these boxes except for the latter one (i.e., assuming relatively standardized test administration conditions in participating classes across countries). First, the assessments are low-stakes, not directly of interest to the student, students don't receive any personal feedback afterwards, and yet participation is implicitly compulsory. In the specific case of TIMSS, this means that once a school agrees to participate in the assessment, all eligible students in the sampled classes are expected to participate in the assessment (Martin et al., 2016). Second, the assessments are quite comprehensive and hence lengthy, making them prone to fatigue, inattentiveness, or boredom effects. The actual testing times for

the achievement part of different assessments are in general set between 80–120 minutes (e.g., technical guides of TIMSS: Mullis & Martin, 2013; PIRLS: Mullis & Martin, 2015; PISA: OECD, 2017). In addition, the survey component of the assessment is typically administered after the achievement test component, further exacerbating the issue for the former component by adding an additional 15–35 minutes of testing time. Finally, there is no social connection with the big organizations behind these assessments and, given that participants are part of a random sample, there is no room for personalization. In the end, it does seem reasonable to expect that some of these factors are somewhat moderated depending on how the student is introduced to the survey by their classroom teacher or by the national attitude towards these assessments (e.g., Sjöberg, 2007). Yet, overall it seems natural to expect some degree of random response behavior to surface in the international large-scale educational assessments.

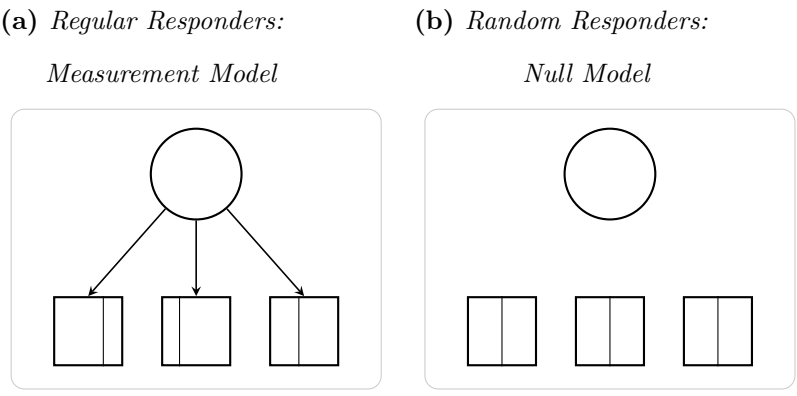
Prevalence. As indicated earlier, estimates of the prevalence of random responders in the international large-scale educational assessments are currently lacking in the literature. Prevalence estimates from other domains vary widely depending on the population and assessment being investigated, but also based on the specific method used to identify the random responders (Credé, 2010; Meade & Craig, 2012). The former variation across populations and assessments is logical, but also implies that it is hard to make predictions for the prevalence in the international large-scale educational assessments where many distinct populations are involved and also a wide variety of content domains are considered in the survey. Credé (2010, p.602) summarizes the current state of knowledge: "the rate of random responding is nonzero for most populations and is likely to fall somewhere between 1% and 10%, although higher rates are certainly possible under certain circumstances.".

Definition: Operationalization. Repeating an earlier general point, the variation in prevalence estimates across methods stresses that it is eventually the operationalization by the method that sets the definition of what/who a random responder is, and hence that this is a crucial choice in the study of this phenomenon. Here, our adopted operational framework is depicted in Figure 1. In this path diagram, the squares represent the

observed responses on the items of a survey scale and the circle represents the underlying latent trait (e.g., ability, knowledge, attitude, ...) of the person responding to this scale. Regular responders are expected to answer consistently according to their own latent trait such that it can be considered the common cause underlying the given item responses by a person as indicated by the arrows going from the circle to the squares in Figure 1a. Formally, this implies that a person's item responses are conditionally independent given the person's latent trait. In contrast, for random responders, their item responses do not necessarily reflect their latent trait and are expected to be mutually independent (cf. the absence of arrows between circle and squares in Panel 1b). More specifically, the random responders are expected to respond uniformly at random, such that each response option has an equal probability of occurrence. This is symbolized in Panel 1b by the vertical line bisecting the square area in equal halves to represent the corresponding response distribution within an item.

Figure 1

Framework to Operationalize and Define Random Responders in terms of (conditional) Independence and Uniformity of Item Responses.



Note. Symbols follow standard path diagram conventions, with squares representing observed variables (i.e., item responses); circles, latent variables (i.e., trait to be measured by the scale of items); arrows indicating dependence relations; vertical lines, categorical thresholds.

Mixture IRT approach. We will adopt a mixture IRT approach (for a review, see Sen & Cohen, 2019) to explicitly model the possibility of two underlying yet unobserved groups in the population, students engaging in regular response behavior versus students

engaging in invalid random response behavior. In our approach, we extend an instance of the HYBRID model by Yamamoto (1989) from binary in/correct responses on achievement tests to the polytomous case for survey responses (for other example extensions, see e.g., Jin et al., 2018). The resulting model is a mixture IRT model in which the component model for the regular responders is a graded response model (Samejima, 1969) consistent with panel A in Figure 1 and the component model for the random responders is a null model consistent with panel B.

More formally, the likelihood of a person's p item response vector $\mathbf{Y}_p$ under the mixture model is formed by the weighted sum of the mixture component model likelihoods

$$\begin{aligned} l(\mathbf{Y}_p = \mathbf{y}_p) = &\Pr(C = \ RR) \Pr(\mathbf{Y}_p = \mathbf{y}_p | C = \ RR) + \\ &\Pr(C = \backslash RR) \Pr(\mathbf{Y}_p = \mathbf{y}_p | C = \backslash RR), \end{aligned}$$

with usual restrictions that component weights $\Pr(C = RR)$ and $\Pr(C = \backslash RR)$ sum up to 1 and are each larger than 0. The component weight $\Pr(C = RR)$ can be interpreted as a prevalence estimate for random responders, a model-based estimate for the percentage of random responders on the scale.

The mixture component model for the *regular responders* (i.e., $C = \backslash RR$) follows the graded response model where item responses of a person are conditionally independent given the person's latent trait θ_p

$$\Pr(\mathbf{Y}_p = \mathbf{y}_p | C = \backslash RR) = \int_\theta \prod_i \Pr(Y_{pi} = y_{pi} | \theta_p) h(\theta) d\theta. \tag{1}$$

In the graded response model, the conditional cumulative distribution function (cdf) of answering in a category k ($k = 1, 2, \ldots, K$) or lower on item i given the person's latent trait θ_p is written as

$$F(Y_{pi} = k | \theta_p) = \frac{1}{1 + \exp\left(-\alpha_i \left[\theta_p - \beta_{ik}^{(\backslash RR)}\right]\right)},$$

in which α_i is recognized as item discrimination parameter and $\beta_{ik}^{(\backslash RR)}$ as item category threshold parameter. The item response probabilities are formed as differences of adjacent

category cdf's: $\Pr(Y_{pi} = k|\theta_p) = F(Y_{pi} = k|\theta_p) - F(Y_{pi} = k-1|\theta_p)$.

The mixture component model for the *random responders* follows a reduced formulation omitting the common latent trait that previously linked responses within a person, such that item responses are mutually independent as in a null model

$$\Pr(\mathbf{Y}_p = \mathbf{y}_p|C = RR) = \prod_i \Pr(Y_{pi} = y_{pi}), \tag{2}$$

with cdf formulated as

$$F(Y_{pi} = k) = \frac{1}{1 + \exp\left(-\beta_k^{(RR)}\right)}.$$

The item category threshold parameters are different from those in the regular responder component model and are fixed to $\beta_k^{(RR)} = -\log(K/k - 1)$ such that each response category has an equal chance of occurrence.

Detection of Random Responders. The mixture model approach allows to classify persons using their posterior most likely component membership as regular or as random responder based on their observed item response pattern (i.e., $\Pr(C = RR|\mathbf{Y}_p)$ vs $\Pr(C = \backslash RR|\mathbf{Y}_p)$). Hence in contrast to more untargeted methods, the identification of random responders is now the result of an explicit link between the conceptual definition of the invalid response behavior and its observable expression in terms of expected response patterns as formalized by the mixture model. There is no need to set arbitrary thresholds for classification as classification is internal to the model approach and the crispness of the classification can be evaluated using accepted criteria such as entropy.

Impact. The presence of random responders to your survey scale will essentially add noise to your sample data and therefore has the potential to confound measurement and related inferences. The general intuition is that the random responders themselves can be considered a source of measurement error for your survey data and therefore that you would expect a general attenuation effect (Spearman, 1904) to occur for any correlation with a measure from a scale affected by random responders. So, we would be at risk of underestimating correlations, factor loadings, reliability, and other related statistics between variables and/or constructs.

In practice, this attenuation-intuition is a slight overgeneralization as the impact of random responders will depend on several factors (i.e., the percentage of random responders in your sample, the consistency of the random responders across scales, and the distribution of scale scores for the regular responders) and can essentially lead to either of three options: no change, attenuation, but also the reverse, inflation (for an overview, see Credé, 2010). Credé (2010) concludes that even with percentages of random responders as low as 5%, observed correlations can be significantly distorted "in a manner that is comparable to the effects of other important study artifacts such as range restriction, dichotomization of continuous variables, and score unreliability" (p.609). Thus, we should not underestimate the potential threat that random responders could form to the validity of our inferences based on survey data from the international large-scale educational assessments.

Method

The data that will be used stems from the Trends in International Mathematics and Science Study (TIMSS) 2015 cycle. TIMSS is an international large-scale educational assessment used to monitor mathematics and science achievement among representative samples of fourth- and eighth-grade students across different countries and is conducted every four years since 1995. Next to achievement measures, TIMSS also collects information about the context for learning through among others a student questionnaire focusing on students' engagement and attitudes towards learning mathematics and science.

Sample

From the 40 unique countries (i.e., technically all education systems including the regional benchmarking participants) who administered TIMSS to the eighth grade in 2015, a sample of five countries was selected for analysis based on their average country achievement scores in mathematics and science, and their type of science program. With the highest performing country – Singapore –, two countries with mathematics and science achievement above the TIMSS scale average of 500 – England and Norway –, and two countries with achievement scores below average – Malaysia and Jordan –, the

selection covers the whole achievement scale. All selected countries have an educational curriculum containing an integrated science program. This implies that these countries all administered the same student survey with the science-related student questionnaire scales referring to science as a general subject (i.e., no distinction between biology; earth science; chemistry; and physics). This helps between-country comparability of at least some of the many contextual factors.

Note that Norway will be represented twice. For the 2015 assessment, Norway administered TIMSS to both the eighth grade, as well as the ninth grade[18]. Given that we don't expect any substantial differences in random response behavior between both adjacent cohorts, we believe we can consider both Norwegian grades as an opportunity for direct internal replication of results in at least one country.

Measures

Value/Confidence in Achievement Domain. Within each of the TIMSS achievement domains (i.e., Mathematics and Science), we focused on two scales – the Value of Mathematics/Science (VoM/VoS) and the Confidence in Mathematics/Science scales (CiM/CiS) – from the student questionnaire. Hence, we have two constructs in two domains. The relation between student Value/Confidence and academic achievement has gained much attention in the literature and is of direct interest to educational stakeholders, and is therefore a relevant subject area to investigate the potential impact of random responders on the validity of inferences.

The Confidence in Science (CiS) scale contains eight items, and the Value of Science scale (VoS), Confidence in Mathematics scale (CiM), and Value of Mathematics scale (VoM) contain nine items each. For each item, a student indicated to what extent s/he agrees with the given statement on a 4-point Likert scale, ranging from 1 (*agree a lot*) to 4 (*disagree a lot*) (for the statements, see Martin et al., 2016).

Academic Achievement. Per domain, Mathematics and Science, a set of five so-called plausible values is provided as an estimate for the student's latent underlying

[18]The government argued that due to the nature of the first grade in the Norwegian school system, Norwegian grade 9 is more comparable to grade 8 internationally. After 2015, only the fifth grade and the ninth grade will participate in TIMSS.

proficiency on the achievement test part of TIMSS. The plausible values are a consequence of the booklet design underlying the achievement test, where the number of items is so large that it is strategically distributed across the sample of students in a country, such that a simple sum or average score is no longer directly comparable, and more advanced estimates are needed (e.g., Von Davier et al., 2009). These plausible values will be used and analyzed accordingly as a measure of student achievement when investigating the impact of random responders on the relation between Confidence/Value and academic achievement across the selected countries.

Design. The main consideration to reduce the response burden of students was the implementation of a booklet design. Overall, the achievement test for the eighth-grade students consists of about 450 multiple-choice and constructed-response items, yet students only answer a limited range of items given the applied design. Students are assigned to one booklet and for the eighth grade each booklet consists of four blocks with 12-18 items. The blocks are administered in two parts (i.e., focus on mathematics or science) and the testing time for each part was set at 45 minutes, with a 30-minute break in between (e.g., Mullis & Martin, 2013). After a second break, the student questionnaire was administered to every student that took part in the main assessment. For the selected sample of countries, the complete student questionnaire consisted of 10 scales and additional items on student background information (Martin et al., 2016). The testing time for the student questionnaire was set at 30 minutes. Students were not allowed to leave the room or start with a new section even if they had already completed the task within the set time frame (Martin et al., 2016). Hence, there is no reward for rushing through the assessment as students had to remain seated in class and received the same break time. The actual testing time for an eighth-grade student in the TIMSS 2015 assessment was set at 120 minutes in total plus the time for the two breaks (Mullis & Martin, 2013).

Statistical Analysis

The mixture models were estimated using Mplus Version 8.1 (Muthén & Muthén, 1998–2017) through the MplusAutomation package for R version 0.7-3 (Hallquist & Wiley, 2018). We used full-information maximum likelihood estimation with robust stan-

dard errors and the accelerated expectation-maximization algorithm with 400 random starts, 100 final stage optimizations, and 10 initial stage iterations. Model estimates and prevalence/impact statistics accounted for the TIMSS sampling design through the total student weights. To not confound results with any measurement non-invariance issues, the mixture model was estimated for all country-scale combinations separately. Analysis scripts were run under R version 4.0.0 (R Core Team, 2020).

Modelling approach. For each scale, a series of three models was estimated: the null model; the graded response model; and the mixture model. Model comparison through information criteria (i.e., AIC & BIC, Wagenmakers & Farrell, 2004) allows to assess (i) the initial starting ground for each of the component models (null & graded response model) in the mixture, (ii) reasonableness of considering people consistently responding to the scale (graded vs null), and (iii) reasonableness of considering the two types of responders instead of one homogeneous population (graded vs mixture). With $K = 4$ ordinal response categories for an item, category thresholds were fixed at $\beta_k^{(RR)} = \{-1.099, 0, 1.099\}$ for the null models.

Classification Validity Checks. To ascertain whether the mixture model provides a solid basis for further classification, and hence the detection of random responders, we implemented two classification validity checks. First, we required a classification entropy of at least .70 to ensure that the mixture is able to provide a crisp classification separation of the sample in a group of random responders and a group of regular responders. Second, the component model for regular responders in the mixture was inspected to ensure that it indeed reflected persons consistently responding on a unidimensional scale (i.e., compatible with a common underlying latent trait); when two or more standardized item discrimination parameters (i.e., factor loadings) dropped below .40, this validity criterion was not met. Cases for which the inclusion criteria do not turn out positively will be disregarded for further analyses and reporting (i.e., some blanco cells might appear in the results).

Random Responders: Group Validation. For the group of students classified as random responders, we will evaluate whether their response patterns are indeed displaying

the characteristics as prescribed under the adopted random responder definition: (i) responses between items ought to be unrelated as evidenced by close-to-zero inter-item correlations, (ii) marginal response distributions per item ought to be close to uniform, and (iii) with individuals tending to make use of the full range of the likert scale as evidenced by the average number of response alternatives used across the survey scale. Empirical evidence against would question the operational success of the approach, dismissing the random responder class of the mixture model as a spurious class accommodating some undefined model assumption violations of the regular IRT measurement model(e.g., Bauer & Curran, 2014). In contrast, empirical evidence in favor would support a strong interpretation in terms of a random responder group and population heterogeneity.

Random Responders: Prevalence and Overlap. We assessed prevalence by the number of students in our sample classified as a random responder by the model. Overlap in classification as random responder was assessed pairwise, across scales, using a simple percentage of those jointly classified as random responder with the corresponding odds ratio as an effect size measure for the interdependence between both classifications.

Random Responders: Impact. The impact of students being flagged as random responders on results and conclusions was evaluated by comparing results with and without random responders (i.e., "without" means here that corresponding observations on the scale on which respondents were flagged, were set as missing). We inspected the scale score's distribution and reliability, but also correlations between scales across/within domains and with achievement in the corresponding domain. In evaluating impact focus is on effect size measures and graphical representations to avoid a too narrow perspective focusing on mere statistical significance.

Results

Mixture Model Results

Model Comparison

Model fit of the series of three item response models for each of the four scales per country are presented in Appendix A (see Table A1). In Table 1, we zoom in on the

results for the Confidence in Mathematics (CiM) scale among Norwegian ninth-grade students as similar results applied to the other scales and countries.

Table 1

Model Fit of the Series of three Models for the Confidence in Mathematics Scale for Norwegian Ninth-Grade Students.

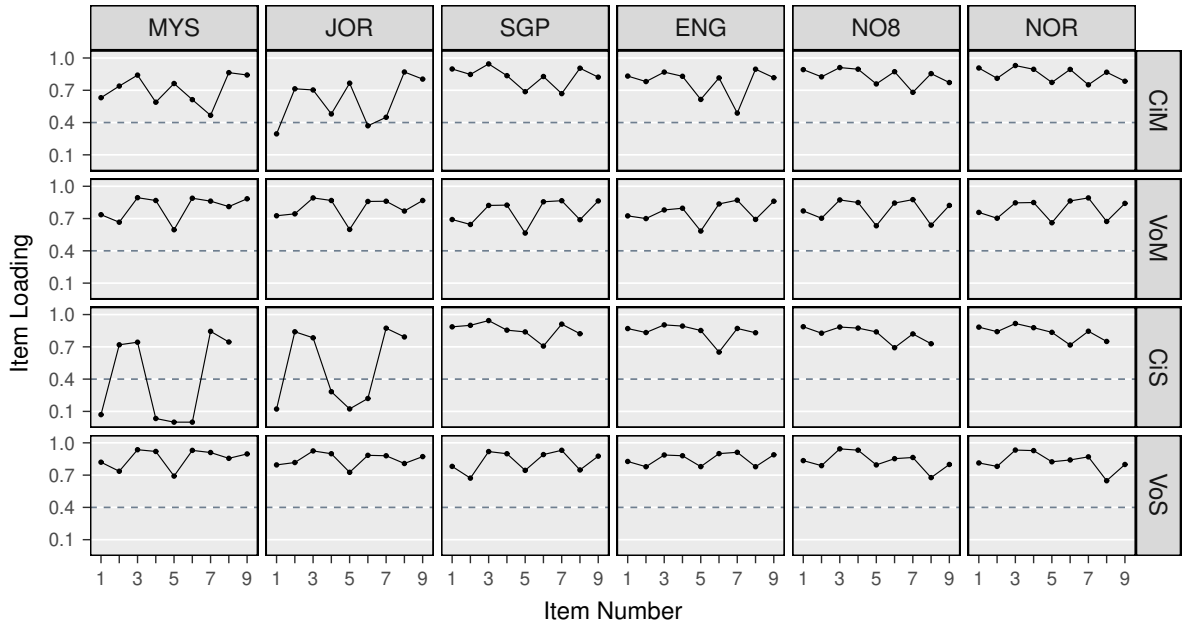
model	#par	-LL	AIC	BIC	wAIC	wBIC
M1: null model	0	56372	112743	112743	0	0
M2: graded response model	36	41082	82236	82468	0	0
M3: mixture model	37	40499	81071	81309	1	1

Note. #par = number of parameters; -LL = -log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; wAIC = weighted AIC; wBIC = weighted BIC.

The huge improvement in fit for the graded response model (M2) over the null model (M1) supported the notion that the scale is unidimensional and that at least a substantial portion of people respond accordingly. Yet, assuming a completely homogeneous population was untenable as the random responder mixture model (M3) on its turn outperformed the graded response model (M2). Although differences in AIC and BIC were smaller in relative magnitude within this last comparison, the differences were consistently in favor of the mixture model, with model weights unanimously distributed to this specific model among the set of 3 alternative models for the data at hand. This supported the notion of population heterogeneity in the manner of responding to the scale and the mixture modelling approach. Note that the pattern of results was also internally replicated for Norwegian eighth-grade students.

Figure 2

Loading Patterns of the Non-random Responders Component Model in the Mixture Model.



Note. In line with responding consistently across items on a unidimensional scale, strong item loadings are expected for the non-random or "regular" responders. Loadings below .40, as indicated by the dashed grey line, are considered weak. The figure row facets are formed by the scales with acronyms being construct (i.e., Confidence in/Value of) domain (Mathematics/Science) combinations. Items {2, 3, 5, 8, 9} on the CiM scale and items {2, 3, 7, 8} on the CiS scale are reverse-coded due to negative item wording. The figure column facets are formed by the countries: Malaysia, Jordan, Singapore, England, and Norway eighth and ninth grade, respectively.

Within-Class Model Characterization

We first characterized the estimated loadings of the non-random responders component model in the mixture model to verify that the non-random responder class is indeed responding consistently across items in line with a unidimensional scale. Under the assumption that the scales are unidimensional, strong loadings were expected across all items in the regular responder class. Otherwise the interpretation of those ending up classified as not being a member of the random responder but of the "regular" (i.e., non-random) responder class would no longer be comparable across scales and countries.

Generally, a clear unidimensional structure was found with average loadings ranging from .67 to .88 across countries and scales. However, deviations from this general trend

occurred for the Confidence in Science scale in Jordan and Malaysia. In these cases, only a subset of items – specifically the reversed worded items – had strong loadings, whereas the loadings reduced to zero for the remaining items (see Figure 2). In Jordan, the Confidence in Mathematics scale showed a similar but less demarcated pattern, where strong loadings for reverse-coded items were combined with more moderate to weak loadings for the remaining items. In all three cases the validity criterion was not met and the non-random responder class could not be simply interpreted as regular responders to a unidimensional scale, and consequently these cases were therefore omitted from further analyses.

Validation of the Random Responders Class

To validate that the students classified as random responders have a response pattern that can be considered "random", we took a closer look at the observed response patterns, as well as the correlation between item pairs. From a theoretical viewpoint, and as implied by the underlying model, the responses in the population were expected to be uncorrelated and every response option was expected to have an equal chance of occurrence. Yet in practice, when looking at the students classified according to the model, results will be prone to sampling variation and classification errors. Consequently, the observed distribution of responses and the theoretically expected distribution cannot be expected to be one-to-one comparable. Similarly, correlations will not be exactly zero and might show some deviations in either direction. Moreover, this sampling variation in results might be further enlarged by the potentially low sample size in the random responder class. Yet, to have confidence that the random responder class was actually "random", the responders in the random class should still follow the expected patterns approximately.

Table 2

Use of Response Scale: Average Number of Response Options Selected per Scale by Regular and Random Responders.

ISO	group	Mathematics				Science			
		Confidence		Value		Confidence		Value	
		n	count	n	count	n	count	n	count
NOR	\RR	4360	2.47	4547	2.47	4392	2.33	4400	2.31
	RR	298	3.34	99	3.47	253	3.22	237	3.46
NO8	\RR	4471	2.49	4639	2.38	4453	2.30	4525	2.29
	RR	275	3.44	88	3.29	290	3.23	209	3.43
ENG	\RR	4438	2.63	4641	2.50	4279	2.30	4543	2.29
	RR	289	3.49	83	3.37	421	3.34	133	3.44
SGP	\RR	5634	2.46	5840	2.42	5563	2.18	5722	2.13
	RR	454	3.38	246	3.46	520	3.13	355	3.39
JOR	\RR	x	x	7469	2.03	x	x	7224	1.74
	RR	x	x	272	3.35	x	x	477	3.44
MYS	\RR	8282	2.42	9041	2.15	x	x	8184	1.80
	RR	1394	3.48	604	3.40	x	x	1282	2.90

Note. The results show to what extent the complete response scale is used. The ISO codes refer to the countries: Norway, ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia. The group variable refers to the different groups of responders, with \RR = regular responders or the whole sample without random responders; RR = random responders; n = weighted sample size; and count = the average number of response options selected.

Use of Response Scales. Figure 3a and Figure 3b show the distribution of the observed responses given by the average proportion of how often a specific response category was selected among all students in a specific class for each country-scale combination. The common trend within the regular responder classes was that on average these students tended to select the "agree" response options (i.e., reflected by the top two categories in the figure) and they did this to a larger degree than their counterparts in the random re-

sponder class. In contrast, the random responder classes showed a strong general pattern where all response options were equally well represented. Malaysia might be singled out by a lower proportion of "disagree a lot" responses on the Value of Science scale. Yet, in this case the distribution of responses was restricted by the baseline probability for this specific category. Across all items this response option was only selected by 2.6% of the total number of participating students. Furthermore, students in the random responders class used on average one additional category when responding to the questionnaire (i.e., $\overline{\text{RR}} = 3.4$ and $\overline{\backslash\text{RR}} = 2.3$; see Table 2) which was in line with a more random and less consistent use of the response scale.

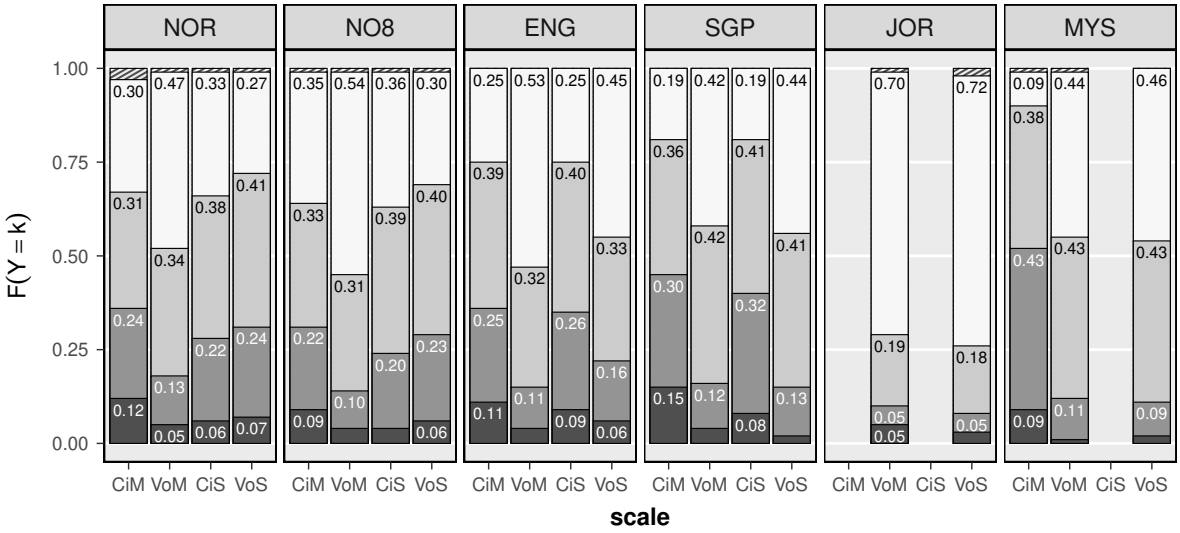
Dependence between item pairs. Whereas the regular or non-random responders showed a degree of consistency in responding with average correlations across item pairs ranging from .43 to .64 for all country-scale combinations, the overall relation between item responses was lacking for all the random responder groups. On average the random responder classes showed near-zero correlations across item pairs (see Table 3). Even though individual item pairs showed some sampling variation in either direction for the random responders, the strength of the relation between the item responses was of different orders of magnitude than for the corresponding regular or non-random responder classes. Within each scale-country combination the majority of individual item pairs showed weak positive or negative relations at best.

Overall, these clear differences in observed response patterns between the classified groups were in line with the theoretically motivated specifications of the mixture model and provided further support for regarding the random responder classes as "random responders" (i.e., response patterns for students in these classes are conform expectation for random responders) and not as mere spurious classes accommodating undefined residual misfit.

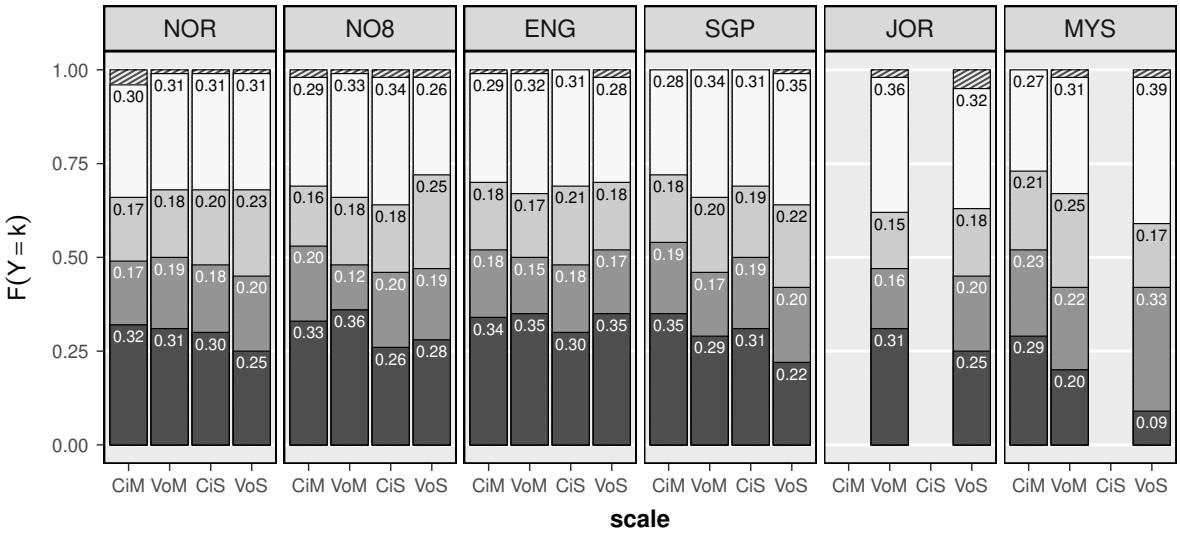
Figure 3

Use of Response Scale by Regular and Random responders.

(a) *Regular Responders: Distribution of Selected Response Options Across Scales and Countries*



(b) *Random Responders: Distribution of Selected Response Options Across Scales and Countries*



Note. The vertical axis in Panel A and Panel B provides the average cumulative proportion for selecting a specific category or lower across scale items, respectively for the regular responders and random responder group. The number in each bar element reflects the average response proportion for that specific category in answering to a specific scale. From bottom to top the reflected categories range from "disagree a lot" to "agree a lot" and end with the missing responses. Under the component model for the random responders, the probability for each category was expected to be 0.25. The scale acronyms are construct (i.e., Confidence in/Value of) domain (Mathematics/Science) combinations. The figure column facets and ISO codes are formed by the countries: Norway ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia.

Table 3

Average correlation across all item pairs for the 'Confidence in' and 'Value of' Mathematics and Science scales.

		Regular Responders						Random Responders					
	ISO	Mathematics			Science			Mathematics			Science		
		M	Q1	Q3	M	Q1	Q3	M	Q1	Q3	M	Q1	Q3
Confidence	NOR	0.64	0.59	0.69	0.59	0.53	0.64	0.03	-0.21	0.30	0.06	-0.21	0.43
	NO8	0.59	0.55	0.65	0.56	0.52	0.63	0.00	-0.29	0.28	0.04	-0.22	0.38
	ENG	0.51	0.40	0.61	0.61	0.58	0.67	0.02	-0.21	0.26	0.04	-0.23	0.38
	SGP	0.60	0.53	0.68	0.64	0.59	0.68	0.06	-0.17	0.27	0.13	-0.14	0.32
	JOR	x	x	x	x	x	x	x	x	x	x	x	x
	MYS	0.43	0.34	0.52	x	x	x	0.02	-0.21	0.32	x	x	x
Value	NOR	0.51	0.45	0.58	0.59	0.54	0.63	0.02	-0.11	0.12	0.07	-0.10	0.22
	NO8	0.49	0.42	0.54	0.60	0.54	0.66	0.05	-0.09	0.23	0.04	-0.12	0.18
	ENG	0.46	0.39	0.53	0.61	0.56	0.67	0.00	-0.16	0.18	0.03	-0.15	0.09
	SGP	0.46	0.40	0.55	0.56	0.48	0.64	0.02	-0.07	0.14	0.08	-0.04	0.16
	JOR	0.51	0.44	0.60	0.61	0.57	0.66	-0.05	-0.15	0.03	0.00	-0.09	0.04
	MYS	0.52	0.43	0.61	0.62	0.54	0.71	0.06	0.00	0.12	0.09	0.04	0.14

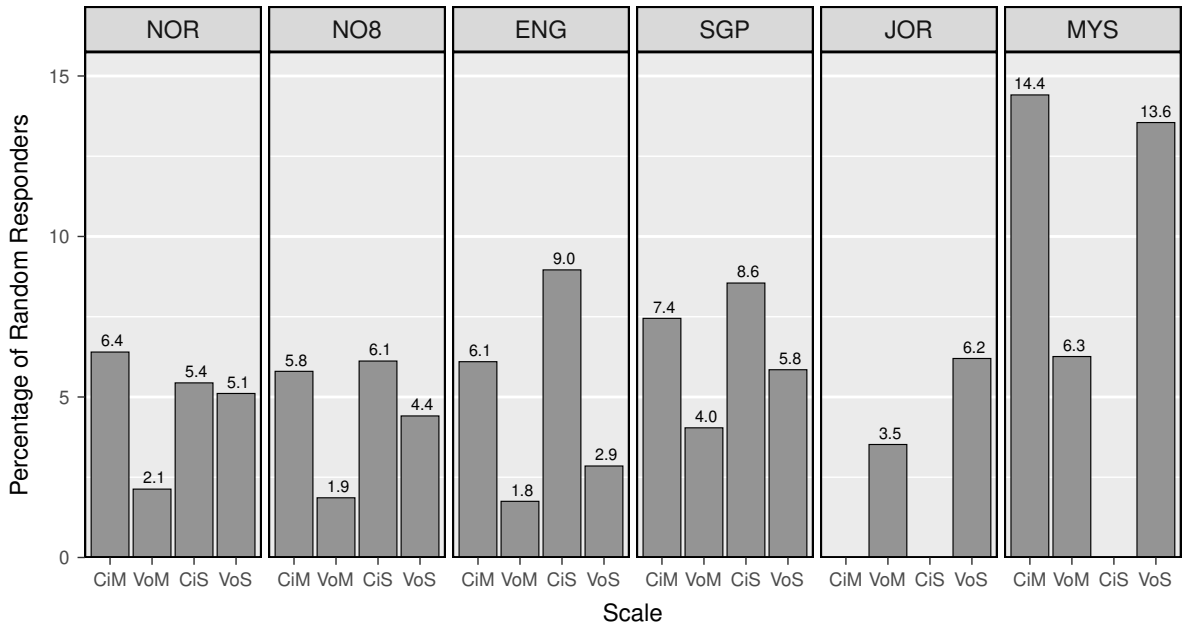
Note. The results show the dependence in responses across all item pairs. For the random responder group, correlations are expected to be centered around zero. The ISO codes refer to the countries: Norway, ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia. With M = mean; Q1 = first quartile; Q3 = third quartile.

Prevalence of Random Responders

The classification results using the mixture approach indicated that the prevalence of random responders was estimated on average at 6% across countries and scales. The regular non-random responder class and random responder class were well-separated as indicated by entropy values well above .70 (see Appendix B, Table B1), leading to a crisp classification (i.e., a participant's posterior membership probability for one class tended to clearly outweigh the other class' membership probability).

Figure 4

Percentage of Random Responders under the Mixture Approach.



Note. The figure column facets are formed by the countries: Norway ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia. The scale acronyms are construct (i.e., Confidence in/Value of) domain (Mathematics/Science) combinations.

Variation across countries and scales. Most of the variation in prevalence occurred between countries, with the average prevalence of random responders across scales ranging from 4.5% for Norway's eighth grade (i.e., 4.8% for the ninth grade) up to 11.4% for Malaysia. With respect to the different scales, the across-country averaged prevalence rates were somewhat closer together with 3.3% for the VoM scale, 6.3% for the VoS, 7.3% for the CiS scale, and 8.0% for the CiM scale.

Zooming in further on the results some tentative prevalence patterns were observed (see Figure 4). The most affected cases were displayed by Malaysia on the Confidence in Science scale with 14.4% of random responders and the 'Value of Science' scale with 13.6% of random responders. Notice that beyond Malaysia no other country showed prevalence rates above 10%. In addition, the overall random responder prevalence rates on the Confidence scales tended to be higher than those on the Value scales. Especially within the mathematics domain this pattern was unequivocal, with prevalence rates that were up to three times higher.

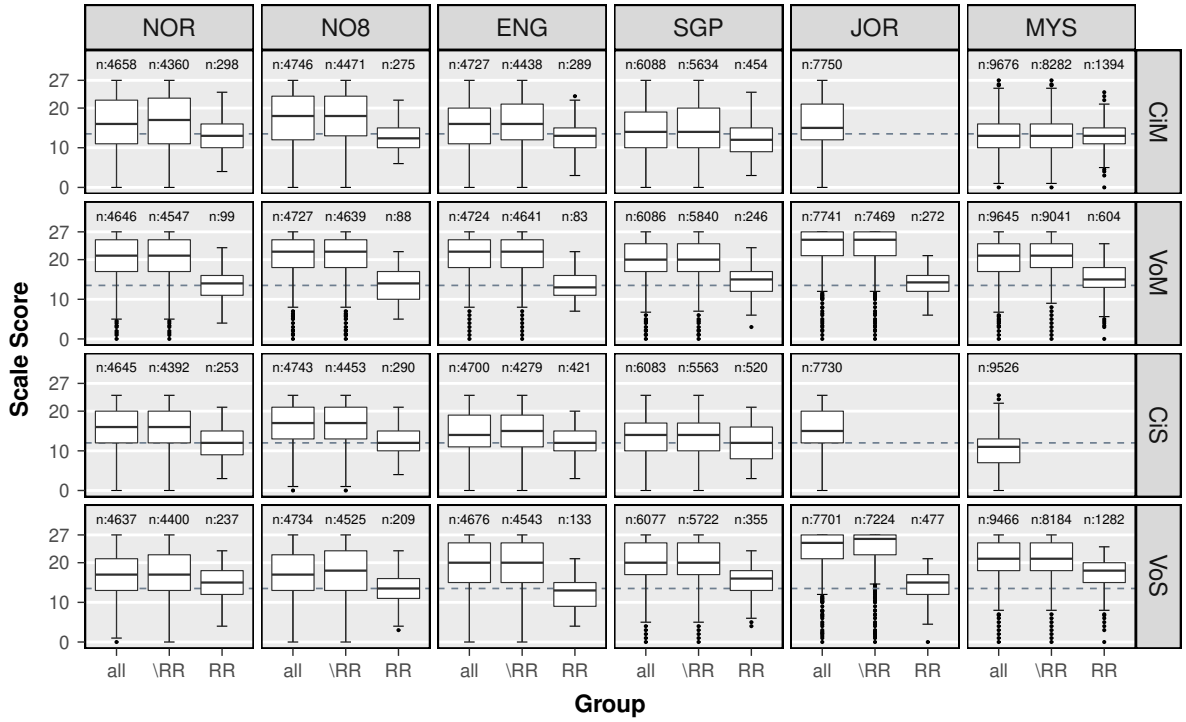
Overlap across scales. Overall, the generally low overlap percentage and low to moderate classification log odds ratios across pairs of scales (see Appendix B, Table B1) provided no indication to conclude that it were always the same students that were classified as random responder across all four scales. Only about 1% of students ended up being classified as a random responder across individual pairs of scales. Yet note that the minimum prevalence rate across the pair of scales serves as an upper bound to this overlap. In general, there was however stronger dependence in classification between scales assessing the same construct (i.e., Confidence or Value) compared to between scales assessing the same domain (i.e., Mathematics or Science).

Impact of Random Responders on Scale-related Inferences

To assess the potential impact of the student classified as random responders on inferences related to the Confidence and Value scales under investigation, we conducted a small sensitivity study. With this objective in mind we computed relevant statistics of interest with and without the random responders present in the data (for the latter, answers were recoded as missing for the random responders). In addition, we will also report on the relevant statistics for the random responder groups to characterize how these compared to the regular or non-random responder groups. Thus, three sets of results are reported: results for the whole sample, for the whole sample without random responders, and for the random responders only. Note that due to relatively low prevalence rates of random responders, the latter third set of results was often based on correspondingly quite low sample sizes.

Figure 5

Distribution of Mean Scale Scores With and Without Random Responders.



Note. For the scale scores, item responses were recoded such that higher mean values were indicative of higher Confidence/Value levels (0 = 'disagree a lot'; 1 = "disagree a little"; 2 = "agree a little"; and 3 = 'agree a lot'). The figure row facets are formed by the scales with acronyms being construct (i.e., Confidence in/Value of) domain (i.e., Mathematics/Science) combinations. The maximum scale score is 27 for the CiM, VoM, and VoS scales and 24 for the CiS scale. The grey lines indicate the scale midpoints. The figure column facets are formed by the countries: Norway ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia. The horizontal axis refers to different groups of responders, with all = the whole sample; \RR = regular responders or the whole sample without random responders; RR = random responders.

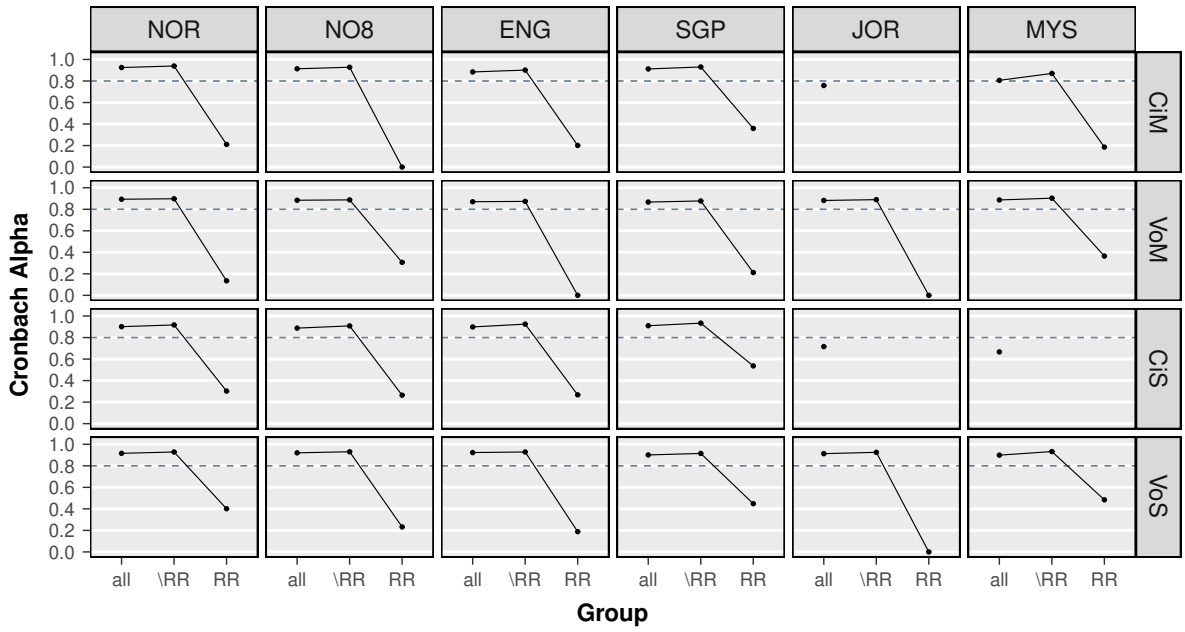
Scales Scores. In general, the random responders tended to score as expected closer around the midpoint of the scale (i.e., 12 for CiS and 13.5 for the other scales), visibly lower than the non-random responder groups (i.e., an average difference of 4.7 points) and more homogeneous as a group (i.e., less variation in score distribution) (see Figure 5). Even though there were large differences in score distribution between those two responder groups, the distribution of scale scores stayed fairly stable when removing the random responders (i.e., whole sample vs. non-random responders only). Hence, at the level of the scale score distribution, the impact of the generally small proportion of random responders was rather limited. In general, the average scale score went up by just .2

points across countries and scales when removing the random responders.

Cronbach's Alpha. Looking at reliability, similar patterns of results were found. In line with expectations, the reliability of the scale for the random responder groups was too low to be meaningful (see Figure 6), while reliability for all of the available non-random responder groups was above .80. Thus again there were large differences in results between the two responder groups, yet excluding the random responders resulted in a minimal increase in Cronbach's alpha. Yet, do keep in mind that the starting reliability for the whole sample was already close to or above .80, and that obviously it was hard to effectuate a large increase in those cases (i.e., the maximum increase was .06 points for Malaysia on the CiM scale).

Figure 6

Scale Reliability With and Without Random Responders.



Note. The figure row facets are formed by the scales with acronyms being construct (i.e., Confidence in/Value of) domain (i.e., Mathematics/Science) combinations. The figure column facets are formed by the countries: Norway, ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia. The horizontal axis refers to different groups of responders, with all = the whole sample; \RR = regular responders or the whole sample without random responders; RR = random responders.

Between-scale Correlations. To examine the relations between Confidence and Value on the same domain (i.e., CiM:VoM and CiS:VoS) or relations within constructs

across the Mathematics and Science domain (i.e., CiM:CiS and VoM:VoS), we looked at pairwise correlations with and without random responders. As mentioned before, on the scales where students were classified as a random responder, their respective scale score was set as missing. In this case, results for the random responder group are not discussed separately, as the number of students being classified as random responders simultaneously across different scale pairs is too limited for meaningful interpretation. In general, the exclusion of the random responders from the whole sample gave rise to correlation differences ranging only from -.04 to .02 across countries and scales (see Table C1). The largest difference was observed for England where a correlation of .50 between Confidence and Value of Science increased to .54.

Correlations with Achievement. In addition, we examined how "Confidence in" and "Value of" Mathematics/Science as reported by the students in the questionnaire were related to the students' achievement scores on the corresponding domain in the assessment. For the random responder groups, the average across countries correlation between Confidence/Value and Achievement was $r = .25/.11$ for Mathematics and $r = .21/.14$ for Science (see Table C1b). In comparison, for the regular responder groups, the average across countries correlation between Confidence/Value and Achievement was $r = .49/.17$ for Mathematics and $r = .36/.21$ for Science. The difference in the relation between Confidence and Achievement was stronger than between Value and Achievement. The higher correlation between Confidence and Achievement was in line with theory, as the Confidence scale in TIMSS can also be regarded as a proxy for self-efficacy, a known correlate of achievement. Yet again, although the differences between the random and regular responder groups were at times quite sizeable, the impact of the random responders on the point estimates remained rather limited. The differences in the correlations with achievement ranged from -.03 to .03 points across countries and scales when excluding the random responders from the total sample. In this case, the largest difference was observed for Malaysia for both correlations with Mathematics Achievement.

Discussion

The aim of the present study was to investigate, using a mixture IRT approach, the prevalence of random responders and their impact on scale-related inferences related to the TIMSS 2015 student questionnaire. The prevalence of random responders was assessed as non-zero across all country-scale combinations of the subset under study with prevalence rates ranging from 1.8% to 14.4%. These estimates are in line with prevalence rates found in for instance personality research (for an overview see e.g., Credé, 2010) and by rapid-guessing approaches in low-stakes achievement tests in an educational context (e.g., Wise et al., 2020).

Individual-level. From a validity perspective, empirical support for two different responder groups present in the data, portraying different underlying response patterns, might in itself already have some implications. Compared to the regular responder groups, the random responder groups tended to use more response options in a more uniform fashion and their response patterns were less consistent (i.e., showing near-zero correlations across item pairs). Given that the observed responses for the random responder groups are so different from what would be expected if students respond according to their own opinions and beliefs related to the questionnaire content, the interpretation of scores is no longer informative for students classified as random responder. In this sense, random responders are by their very definition a threat to validity at the individual level.

Group-level. In some cases the differences in response behavior also lead to large group differences in the scale-related impact statistics (i.e., internal scale statistics, correlations across scales, and correlations with achievement). It has especially been these differences between the actual responder groups (e.g., (un)motivated; (in)sufficient effort groups) that have been of particular interest in the context of international large-scale educational assessments (for examples in terms of average achievement scores see e.g. Eklöf et al., 2014; Hopfenbeck & Kjærnsli, 2016; or average accuracy scores see e.g., Michaelides et al., 2020). Because the random responder group by definition tends to score near the survey scale average, larger group differences on the relevant impact statistics coincide with more homogeneous outspoken responses (i.e., far from the theoretical scale average)

by the regular group. In our study, the results showed that in the most extreme cases, the regular or non-random responder groups had scale scores up to nine points higher given a maximum scale score of 27; scale reliabilities up to .93 points higher; or correlations up to .34 points higher, compared to the random responder groups. Such differences are quite disconcerting and hence random responders can be a real threat to validity at the group level when studying differences between groups that contain disproportionate numbers of random responders.

Aggregated-level. Moving one step further, we compared analysis results for the whole sample with and without random responders included. Overall, the impact of random responders on the inferences at the aggregate level was rather limited. This does not mean that there were no quantitative differences in analysis results with and without random responders included, yet these differences were not representing any relevant qualitative changes in the results and conclusions. This result at the aggregate level is in contrast to the validity consequences at the individual level and the group level.

Group differences in themselves are not a sufficient precondition for finding inferential impact at the aggregate level. For example, the impact of random responders on the correlations between substantive measures is not only influenced by the difference in correlation between the two responder groups but also by the prevalence rate of the random responders and of how strongly their total scale scores are separated from the regular responders. Keeping everything else equal, impact will increase as the proportion of random responders or score separation increases (Credé, 2010).

In this study, the overall prevalence rates for the random responder groups stayed within limits and the average group differences were small. All combined this results in the random responders forming a non-influential and small outlier group that is out crowded by the typically large samples (i.e., $n = 4000$ as the target sample size in TIMSS) of regular responders in the international large-scale educational assessments. Compare this to other non-educational contexts such as personality tests or psychological assessments where the ruling impression is that random responders would have a large impact, but where sample sizes tend to be smaller and prevalence rates and group differences tend

to be larger. Yet in an educational context, the limited impact on the aggregated level results is not a one-off finding in the literature. Wise et al. (2020), for example, found in a large-scale educational achievement testing context that aggregated school-level scores remained rather stable after filtering out random responders based on their response times (i.e., so-called rapid guessing response behavior). Using similar methodology, Wise (2006) found a small positive effect on the correlation between information literacy and SAT achievement scores; the reported increase of .01-.03 is similar in magnitude to what was found in the present study.

Limitations and Other Considerations. The non-extreme prevalence of random responders and robustness of the reported results might give the community reason to be moderately optimistic about data quality of the survey and ignorability of the issue of random responders. Yet we do have to keep in mind the necessary caveats with respect to generalization of the findings.

First, the study considered one set of rudimentary inferential results. More complex analyses could concern more scales, variables, and interrelations. In these situations, small effects of the presence of random responders could potentially accumulate to noticeable differences in inferential conclusions at some point. Especially when the focus would be on the outcome of the statistical significance filter (cf., the tentative simulation study by Rios (2021) pointing at an increased type-I error rate in measurement invariance testing). Alternatively, there could also be situations in which the presence of random responders in itself is sufficient to raise some concerns, regardless of their influence on scale-related inferences. For example, representativity of the sample could be impaired if it would be non-random groups of students engaging in random response behavior. If these students are removed, any conclusions or policy recommendations based on these results might be somewhat restricted.

Second, the study investigated only one particular model-defined way of invalid response behavior. Giving responses that are more in line with a random responder than with a regular responder is only one manifestation of invalid response behavior. In practice, many other (more systematic) types of invalid response behavior could have been

considered (e.g., straightlining, inconsistent responding, and speeding). By focusing on one specific pattern of invalid response behavior, the prevalence estimates will be a conservative estimate of overall invalid response behavior. Different patterns of invalid response behavior also have different impact and it could be interesting to see if their combined effect would lead to more pronounced impact results. Yet, different patterns of invalid response behavior will require different methods for detection (e.g., Hong et al., 2020; Huang et al., 2012; Meade & Craig, 2012).

Third, the study considered a sample of five countries and two scales per domain, the Value of Mathematics/Science and the Confidence in Mathematics/Science, from the TIMSS 2015 eighth-grade student questionnaire. There is no guarantee that the results will transfer perfectly to other survey scales. Furthermore, scales with few items and/or items on which the sample of respondents do not use the full range of response options will make it hard to distinguish between any type of responder, random or non-random. Generalization to other countries, other TIMSS cycles, or other international large-scale educational assessments is not guaranteed. Eklöf (2010) for instance suggests that the motivation to participate in TIMSS might be influenced by grade level, and there might be other covariates or contextual factors that can potentially give rise to more/less random response behavior. In this respect, the transition of TIMSS to a computer-based assessment administration system and observed mode effect (Fishbein et al., 2018), might or might not be related to the typical increase in random response behavior when comparing paper-and-pencil to computer-based administration (Beach, 1989). Therefore, regardless of our initial set of positive results it would be good to remain cautious and further investigate, monitor, and keep track of random response behavior and the impact this might have on results and conclusions from international large-scale educational assessments.

Conclusion

When further exploring invalid response behavior in the context of international large-scale educational assessments, we do want to call for establishing a clear link between the operational detection method and the definition of invalid responders, without the

need to put forward intangible theories on underlying reasons or intentions. In contrast, adopting generic definitions such as "insufficient effort responding" would "underscore the cause of the response behavior without presupposing specific response patterns or outcomes" (Huang et al., 2012, p.100). Yet, the responses students provide on a questionnaire are the only visible piece of information we have in order to say something about the behavior of those students on the questionnaire. Regardless of the used detection method, we are only able to flag those students likely engaging in invalid response behavior. Their response pattern does not provide any explanation of why they responded as they did. Different manifestations of response behavior can have similar underlying mechanisms, whereas similar responses can be caused by different intentions. The definition of "random responder" following the proposed mixture approach does not imply that the person is consciously or intentionally randomly responding in an *absolute* sense, but merely that they have a response pattern that is more consistent in a *relative* sense with people whose responses to the scale would be unrelated across items than with people whose item responses are related in a common-cause latent variable sense[19]. Admittedly, this is a very operational definition, but it does not confound interpretation with implications and/or connotations about (un)conscious intentions that are hard to capture. The response pattern is more random than regular, the individual's underlying response processes are unknown. Instead of directly diving into the deep with the more ambitious higher-order goal of understanding students' intentions and response processes, it is more realistic and fruitful at this point to first tackle the already difficult enough task of implementing a thorough data quality monitoring system for international large-scale educational assessments and make a start at more systematically charting the land of invalid response behavior in these assessments. The mixture IRT approach used in this study is one possible tool in such monitoring system.

Practical Recommendations. The low-stakes character of international large-scale educational assessments and individual and cross-cultural/national differences cannot be made to disappear, but it is important to remain vigilant on data quality to not destroy

[19]Yamamoto (1989) would refer to this as the "unscalable" class.

for policymakers and researchers alike, the treasure trove of information collected in these large endeavours. However, currently it is mostly left to the users of the large-scale assessments to assure the validity of the results (Braeken, 2016) and take care of the broader data quality. From this perspective, it would be ideal if the community of researchers using the data from the international large-scale educational assessments forms a habit to by default conduct and report sensitivity checks to study the robustness of their findings. Flagging suspect survey responders and inspecting whether results change substantially when these flagged responders are removed. The current study can be seen as one example illustration of this practice.

However, it seems natural that a more structured data quality monitoring process would be centrally organized or at least facilitated. To the organizing parties and stakeholders of the international large-scale educational assessments, we would therefore make a plea for the default inclusion of proven standard survey measures to facilitate detection of invalid response behavior (e.g., Breitsohl & Steidelmüller, 2018; Leiner, 2019): (i) the inclusion of an instructed response item (e.g., "Please mark slightly agree") or bogus item (e.g., "I have never used a computer") at a few random moments throughout the survey for an individual pupil in combination with a warning at the start of the survey that such items can be included, and (ii) the provision of individual survey completion speed indicators to help track rushed responding (cf. analogue to rapid guessing in the achievement context, e.g., Michaelides et al., 2020; Wise et al., 2020) with the ethical requirement that participants are informed about such data being collected. The recent move to computer-based assessment, would make it straightforward to implement both measures. The extra structure in design and data will make the regular response patterns more predictable and facilitates the detection of irregular invalid responses.

In an ideal world, all stakeholders would have access to a chart of invalid response behavior indicators and trends across countries and scales to further guide any inferences and policy recommendations based on the international large-scale educational assessments and provide a possibility for an informed response to those sceptical about the value of the survey data and robustness of related inferences. In other words, a lot of

work is still on the table.

References

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

Bauer, D. J., & Curran, P. J. (2014). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*(1), 3–29.

Beach, D. A. (1989). Identifying the random responder. *The Journal of Psychology*, *123*(1), 101–103.

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-fit as an index of inattentive responding: A comparison of methods using polytomous survey data. *Applied Psychological Measurement*, *43*(5), 374–387.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340.

Braeken, J. (2016). International larges-scale assessments: Elephants at the gate? In S. Ludvigsen (Ed.), *Northern Lights on PISA and TALIS* (pp. 195–216). Nordic Council of Ministers.

Breitsohl, H., & Steidelmüller, C. (2018). The impact of insufficient effort responding detection methods on substantive responses: Results from an experiment testing parameter invariance. *Applied Psychology*, *67*(2), 284–308.

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*(3), 279–304.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*(3), 311–326.

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

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.

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education*, *6*, Article 11.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2), 312–345.

Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal*, *27*(3), 406–422.

Hopfenbeck, T. N., Lenkeit, J., Masri, Y. E., 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.

Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing*, *11*(2), 95–121.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods*, *21*(1), 197–225.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298.

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229–248.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*(3), 215–237.

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.

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., & Martin, M. O. (2015). *PIRLS Assessment Framework: 2nd Edition*. TIMSS & PIRLS International Study Center, Boston College.

Muthén, L. K., & Muthén, B. O. (1998–2017). Mplus User's Guide. Eighth Edition.

OECD. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving, revised edition*. OECD Publishing.

R Core Team. (2020). R: A language and environment for statistical computing.

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

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, *42*(3), 411–430.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191.

Sjöberg, S. (2007). PISA and "real life challenges": Mission impossible? In S. Hopman, G. Brinck, & M. Retzl (Eds.), *PISA According to PISA* (pp. 203–224). LIT Verlag.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI monograph series. Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114.

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

Yamamoto, K. (1989). Hybrid model of IRT and latent class models. *ETS Research Report Series*, *RR-89-41*.

Appendix A: Model comparison results

Table A1

Model Fit of the Series of three Models for the "Confidence in" and "Value of" Mathematics and Science scales.

(a) *Norway ninth grade*

	model	Mathematics						Science					
		#par	-LL	AIC	BIC	*w*AIC	*w*BIC	#par	-LL	AIC	BIC	*w*AIC	*w*BIC
Confidence	M1: null model	0	56372	112743	112743	0	0	0	50674	101349	101349	0	0
	M2: graded response model	36	41082	82236	82468	0	0	32	35007	70079	70285	0	0
	M3: mixture model	37	40499	81071	81310	1	1	33	34424	68915	69127	1	1
Value	M1: null model	0	57530	115060	115060	0	0	0	57248	114496	114496	0	0
	M2: graded response model	36	35647	71365	71597	0	0	36	39452	78977	79208	0	0
	M3: mixture model	37	35513	71099	71338	1	1	37	39043	78159	78398	1	1

(b) *Norway eighth grade*

	model	Mathematics						Science					
		#par	-LL	AIC	BIC	*w*AIC	*w*BIC	#par	-LL	AIC	BIC	*w*AIC	*w*BIC
Confidence	M1: null model	0	58458	116915	116915	0	0	0	51770	103539	103539	0	0
	M2: graded response model	36	41934	83939	84172	0	0	32	35637	71337	71544	0	0
	M3: mixture model	37	41253	82579	82819	1	1	33	34936	69938	70151	1	1
Value	M1: null model	0	58319	116638	116638	0	0	0	58333	116666	116666	0	0
	M2: graded response model	36	33466	67005	67237	0	0	36	39209	78491	78724	0	0
	M3: mixture model	37	33309	66692	66931	1	1	37	38829	77733	77972	1	1

Note. #par = number of parameters; -LL = -log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; *w*AIC = weighted AIC; *w*BIC = weighted BIC.

- Table A1 continued -

(c) *England*

	model	Mathematics						Science					
		#par	-LL	AIC	BIC	*w*AIC	*w*BIC	#par	-LL	AIC	BIC	*w*AIC	*w*BIC
Confidence	M1: null model	0	58612	117223	117223	0	0	0	51849	103698	103698	0	0
	M2: graded response model	36	45122	90317	90549	0	0	32	38500	77063	77270	0	0
	M3: mixture model	37	44525	89124	89364	1	1	33	37524	75114	75327	1	1
Value	M1: null model	0	58707	117414	117414	0	0	0	58089	116177	116177	0	0
	M2: graded response model	36	34736	69545	69777	0	0	36	36299	72670	72902	0	0
	M3: mixture model	37	34643	69360	69600	1	1	37	36118	72311	72550	1	1

(d) *Singapore*

	model	Mathematics						Science					
		#par	-LL	AIC	BIC	*w*AIC	*w*BIC	#par	-LL	AIC	BIC	*w*AIC	*w*BIC
Confidence	M1: null model	0	75857	151714	151714	0	0	0	67364	134728	134728	0	0
	M2: graded response model	36	57023	114117	114359	0	0	32	47680	95423	95638	0	0
	M3: mixture model	37	55935	111943	112192	1	1	33	45978	92022	92244	1	1
Value	M1: null model	0	75895	151789	151789	0	0	0	75719	151437	151437	0	0
	M2: graded response model	36	47650	95373	95614	0	0	36	45127	90325	90567	0	0
	M3: mixture model	37	47321	94716	94965	1	1	37	44511	89096	89344	1	1

Note. #par = number of parameters; -LL = -log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; *w*AIC = weighted AIC; *w*BIC = weighted BIC.

- Table A1 continued -

(e) *Jordan*

	model	Mathematics						Science					
		#par	-LL	AIC	BIC	wAIC	wBIC	#par	-LL	AIC	BIC	wAIC	wBIC
Confidence	M1: null model	0	94436	188873	188873	0	0	0	83868	167736	167736	0	0
	M2: graded response model	36	79716	159505	159755	0	0	32	66057	132177	132400	0	0
	M3: mixture model	37	78189	156452	156709	1	1	33	64395	128856	129086	1	1
Value	M1: null model	0	94947	189895	189895	0	0	0	94388	188777	188777	0	0
	M2: graded response model	36	48073	96218	96469	0	0	36	43446	86963	87213	0	0
	M3: mixture model	37	47956	95987	96244	1	1	37	43222	86518	86775	1	1

(f) *Malaysia*

	model	Mathematics						Science					
		#par	-LL	AIC	BIC	wAIC	wBIC	#par	-LL	AIC	BIC	wAIC	wBIC
Confidence	M1: null model	0	120108	240216	240216	0	0	0	104802	209605	209605	0	0
	M2: graded response model	36	94482	189036	189294	0	0	32	97084	194233	194462	0	0
	M3: mixture model	37	91196	182465	182731	1	1	33	93651	187369	187605	1	1
Value	M1: null model	0	119834	239669	239669	0	0	0	117425	234851	234851	0	0
	M2: graded response model	36	71853	143778	144036	0	0	36	70607	141285	141543	0	0
	M3: mixture model	37	70441	140955	141221	1	1	37	68086	136246	136511	1	1

Note. #par = number of parameters; -LL = -log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; wAIC = weighted AIC; wBIC = weighted BIC.

Appendix B: Prevalence and Overlap of Random Responders

Table B1

Prevalence and Overlap of Random Responders across Scales.

ISO	Mathematics						Science						Overlap							
	Confidence			Value			Confidence			Value			CiM:VoM		CiS:VoS		CiM:CiS		VoM:VoS	
	n	E	RR	n	E	RR	n	E	RR	n	E	RR	%	LOR	%	LOR	%	LOR	%	LOR
NOR	4658	0.89	6.4%	4646	0.94	2.1%	4645	0.89	5.4%	4637	0.91	5.1%	0.4	1.2	0.5	0.8	1.5	2.0	0.5	1.9
NO8	4746	0.90	5.8%	4727	0.95	1.9%	4743	0.88	6.1%	4734	0.92	4.4%	0.3	1.3	0.6	1.0	1.6	2.0	0.6	2.6
ENG	4727	0.86	6.1%	4724	0.96	1.8%	4700	0.86	9.0%	4676	0.93	2.9%	0.3	1.2	0.5	0.9	2.1	1.8	0.4	2.5
SGP	6088	0.87	7.4%	6086	0.92	4.0%	6083	0.90	8.6%	6077	0.93	5.8%	0.6	0.9	1.0	0.9	1.9	1.5	1.1	2.0
JOR	7750	x	x	7741	0.9	3.5%	7730	x	x	7701	0.90	6.2%	x	x	x	x	x	x	1.1	2.2
MYS	9676	0.84	14.4%	9645	0.93	6.3%	9526	x	x	9466	0.91	13.6%	3.1	2.0	x	x	x	x	2.8	1.9

Note. The ISO codes refer to the countries: Norway, ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia; n = sample size for the whole sample; E = entropy; RR = percentage of random responders; Overlap % = percentage of students being classified as random responder under the pair of scales; LOR: log odds ratio of being classified as random/non-random responder under the pair of scales. The overlap is determined for all pairwise combinations of scales. Scale acronyms are formed as construct (i.e., Confidence in/Value of) domain (Mathematics/Science) contractions.

Appendix C: Scale Related Impact of Random Responders

Table C1

Correlations With and Without Random Responders.

(a) *Between-Scales Correlations.*

ISO	CiM:VoM			CiS:VoS			CiM:CiS			VoM:VoS		
	all	\RR	RR	all	\RR	RR	all	\RR	RR	all	\RR	RR
NOR	0.38	0.40	*	0.42	0.45	*	0.44	0.46	*	0.54	0.55	*
NO8	0.34	0.36	*	0.43	0.45	*	0.40	0.40	*	0.52	0.52	*
ENG	0.34	0.35	*	0.50	0.54	*	0.28	0.29	*	0.48	0.48	*
SGP	0.35	0.35	*	0.47	0.47	*	0.16	0.16	*	0.44	0.44	*
JOR	0.35	x	x	0.38	x	x	0.36	x	x	0.51	0.51	*
MYS	0.24	0.27	*	-0.05	x	x	0.03	x	x	0.41	0.39	*

(b) *Correlations with Achievement.*

ISO	CiM:Mathematics			VoM:Mathematics			CiS:Science			VoS:Science		
	all	\RR	RR	all	\RR	RR	all	\RR	RR	all	\RR	RR
NOR	0.61	0.62	0.28	0.25	0.25	0.23	0.45	0.45	0.28	0.20	0.20	0.10
NO8	0.58	0.58	0.27	0.19	0.18	0.29	0.37	0.36	0.18	0.14	0.13	0.15
ENG	0.47	0.47	0.25	0.13	0.12	0.01	0.40	0.40	0.28	0.26	0.25	0.13
SGP	0.42	0.42	0.27	0.13	0.13	0.01	0.25	0.25	0.10	0.27	0.27	0.19
JOR	0.38	x	x	0.16	0.16	-0.08	0.39	x	x	0.17	0.17	-0.02
MYS	0.34	0.37	0.17	0.23	0.20	0.18	-0.18	x	x	0.24	0.21	0.31

Note. Correlations are based on total scale scores, as well as achievement scores. The scale acronyms are construct (i.e., Confidence in/Value of) domain (i.e., Mathematics/Science) combinations. For the scale scores, item responses were recoded such that higher mean values were indicative of higher Confidence/Value levels (0 = "disagree a lot"; 1 = "disagree a little"; 2 = "agree a little"; 3 = "agree a lot"). The ISO codes refer to the countries: Norway, ninth and eighth grade respectively, England, Singapore, Jordan, and Malaysia. Results are given for different groups of responders, with all = the whole sample; \RR = regular responders or the whole sample without random responders; RR = random responders. With respect to the between-scale correlations in Panel A, results for the random responder group are not presented (*), due to too limited number of students being classified as random responder across different scale pairs.

7 Article 4: Where

van Laar, S., & Braeken, J. (2022c). *Prevalence of Random Responders as a function of Scale Position and Questionnaire Length in the TIMSS 2015 eighth-grade Student Questionnaire*. Manuscript under review.

Prevalence of Random Responders as a function of Scale Position and Questionnaire Length in the TIMSS 2015 eighth-grade Student Questionnaire

This study examined the impact of two questionnaire characteristics, scale position and questionnaire length, on the prevalence of random responders in the TIMSS 2015 eighth-grade student questionnaire. While there was no support for an absolute effect of questionnaire length, we did find a positive effect for scale position, with an increase of 5% in random responding over the course of the questionnaire (in both the shorter and the longer version). However, scale character turned out to be an unexpected but more important determinant. Scales about students' confidence in mathematics or science showed an increase of 9% in random responding, which is double the impact of scale position. Potential mechanisms underlying the confidence case and general implications of the results for questionnaire design are discussed.

Survey answers can be distorted by construct-irrelevant factors that influence response behavior. A potential measurement validity problem arises here as the corresponding scale scores might no longer accurately reflect knowledge, abilities, or opinions related to the survey content (e.g., Cronbach, 1950; Messick, 1984). A prominent contextual factor that tends to trigger such invalid response behavior is a low-stakes-low-effort situation, a context characterizing for instance most international large-scale educational assessments such as IEA's Trends in International Mathematics and Science Study (TIMSS) or OECD's Programme for International Student Assessment (PISA). For students participating in these types of assessments, there are no personal consequences linked to their responses on the assessment, and hence, students might not always respond accurately or thoughtfully, but instead shift to responding with the lowest effort (e.g., Curran, 2016; Eklöf, 2010).

In this study, we will specifically focus on *random responding*, which is one type of response behavior that is considered a typical expression of this low-stakes-low-effort

context where students provide "responses without meaningful reference to the test questions" (Berry et al., 1992, p.340). Specifically, using TIMSS 2015 as a case study, we will investigate the prevalence of random responders among the students across the different scales of the TIMSS eighth-grade student questionnaire and in the different participating countries. Profiting from the large-scale of the TIMSS study and the natural variation in questionnaire version among countries, the potential impact of two construct irrelevant external factors, scale position and questionnaire length, on random responding will be explored.

Questionnaire characteristics in Context: Scale Position × Questionnaire Length

Scale position. One of the most common risk factors that has been hypothesized to influence response quality is item position. In the context of low-stakes assessments in the personality and survey literature, invalid response behavior appears to become more frequent near the end of the questionnaire, regardless of the specific content of the items considered (Bowling et al., 2021; Galesic & Bosnjak, 2009). With respect to random responding, rapid-guessing research provides an example of the specific impact of item position on this type of behavior. The underlying idea is that when responses are given too fast, students have not been able to accurately reflect on the given questions and the "answers given during rapid-guessing are essentially random" (Wise & Kong, 2005, p.167) and no longer reflective of their true knowledge or abilities. For achievement tests, rapid guessing studies have shown that items located closer to the end of the assessment tend to receive more random responses overall (e.g., Wise et al., 2009).

Although most research has focused on item-level position effects, every extra scale added to a questionnaire can be seen as an additional group of items that need to be answered, and hence the position effect would naturally extend to the scale level. For example, in a small-scale study with university students, Merritt (2012) included one additional scale on affective committment to a questionnaire, either at the beginning or the end, with the latter position resulting in more invalid responding. Similarly, when looking at two blocks of items of differing contents (i.e., numeracy and literacy) in an

educational achievement test, Goldhammer et al. (2017) found for both blocks that when presented in the first versus second part of the assessment, more invalid response behavior was observed in the latter position. As students progress through the questionnaire, they can be prone to experience for example boredom, disinterest, inattentiveness, or fatigue, and as a consequence, provide responses that are no longer accurate or thoughtful.

Hypothesis 1. Scales at a later position in the questionnaire display a higher prevalence of random responders compared to scales at an earlier position.

Questionnaire length. Based on the notion of similar underlying mechanisms, a second potential risk factor that has been put forward is questionnaire length (e.g., Meade & Craig, 2012). However, the literature shows mixed results with respect to the relation between questionnaire length and response quality. Herzog and Bachman (1981) used two types of questionnaires in their study, a short 45-minute version and a long 2-hours version, and found higher levels of overly uniform responding in the longer questionnaire. In a similar fashion, longer internet surveys were characterized by lower completion rates (e.g., Deutskens et al., 2004; Galesic & Bosnjak, 2009). In contrast, Boe et al. (2002) found that the 'persistence to respond' to the TIMSS 1995 student questionnaire, as measured by the percentage of missing responses across the entire questionnaire, was not significantly related to the length of the administered questionnaire. Furthermore, in a set of small-scale studies with university students, Gibson and Bowling (2020) showed that the influence of questionnaire length for personality assessments is dependent on the context of questionnaire administration and on the operationalization or detection method for invalid response behavior. Even though the literature is not unanimously in agreement, we would still expect that a longer questionnaire length coincides with more random responding overall, even in the TIMSS student questionnaire, as it has been stated that "among the few documented problems detected by the national monitors were students complaining about the length of the Student Questionnaire" (Martin et al., 2016, p.6.19).

Hypothesis 2. Longer questionnaires display a higher prevalence of random responders compared to shorter questionnaires.

Position* × *Length. The final external factor that we will take into consideration is the interplay between scale position and questionnaire length. Wise et al. (2009) for example wondered whether adjustments to questionnaire length might be sufficient to counteract the observed position effects. Yet current literature shows that it is hard to pinpoint a generic criterion for the optimal length of a questionnaire as this would among other things depend on the amount of invalid response behavior that is considered acceptable, as well as on more pragmatic contextual factors (e.g., the context of administration) (Bowling et al., 2021). In addition, invalid response behavior appears related to questionnaire length or the number of questions overall. For example, Deutskens et al. (2004) not only found that fewer respondents are finishing an internet survey as it gets longer, but that the respondents would finish less of the longer questionnaire percentwise. Hence, respondents' subjective perception of questionnaire length and the pace at which they proceed through the questionnaire might actively moderate potential position effects. A longer questionnaire might drain a respondent's resources at a faster pace by sheer negative anticipation for what's still waiting ahead. Note that this would imply a synergistic interaction between scale position and questionnaire length.

Hypothesis 3. In longer questionnaires, scales at a later position in the questionnaire display an even higher prevalence of random responders compared to scales at an earlier position, than in a shorter questionnaire.

This study

When studying the impact of scale position and questionnaire length on random response behavior an ideal setup would be a large-scale experimental design where we, under controlled scale-content conditions manipulate these two external content-irrelevant factors and randomize thousands of participants across the experiment while administering the resulting questionnaire versions under low-stakes conditions to our target population of high school students. Yet, such an extensive experiment might not be a realistic endeavor. As illustrated in the previous subsection, studies in the literature are mostly based on personality questionnaires administered to relatively small convenient samples of university students in typical Western countries, on internet surveys with somewhat

larger but still non-random samples of participants, or on achievement tests in combination with response-time data (cf. rapid guessing). Simply extrapolating the evidence base from these types of study designs and contexts to random responding on low-stakes questionnaires for high school students in large-scale educational assessments in a more international context seems not warranted. Thus, more specific tailored research is needed to answer our research questions.

Here, we will be using the eighth-grade student questionnaire of *Trends in International Mathematics and Science Study* (TIMSS) 2015 as a specific case study. Profiting from the large-scale of TIMSS, the study has large representative random samples of eighth-grade students in each of the participating countries, bringing along extra generalization support and potential systematic country variation that can be of interest to educational stakeholders. Furthermore, the TIMSS 2015 student questionnaire provides natural variation in scale position and questionnaire length across countries as two versions of the questionnaire were administered. The specific version that was administered in a country depended on the structure of the science curriculum program taught by that country. The student questionnaire under the separated science program is much longer than under the integrated science program (i.e., respectively 19 and 10 scales beyond basic demographics/background information). The order of the scales in each version remains constant across administrations and the first 6 scales and the last scale of both versions were similar for all students. Furthermore, most scales had a similar setup with respect to question format and answer alternatives, with some being replicates if it were not for subject domain differences (e.g., confidence in biology or confidence in chemistry). All these features allow studying random response behavior as a function of the two content-irrelevant factors of interest: scale position and questionnaire length.

Note that there are no response times available for the student questionnaire (so far, these have typically only been available for the achievement tests part of the international large-scale assessments), and hence popular rapid guessing methodology to identify random responders is not an option. Self-report data or convincing psychological effort-related proxies are also lacking. Instead, we will rely on an operationalization

of random response behavior by van Laar and Braeken (2022) that is directly based on the questionnaire responses given at scale level and uses a mixture item response theory (IRT) approach (for a review, see Sen & Cohen, 2019) to explicitly model the possibility of two underlying yet unobserved groups in the population, students engaging in regular response behavior versus students engaging in invalid random response behavior across the items of a scale.

Method

The data that will be used comes from the *Trends in International Mathematics and Science Study* 2015 cycle. TIMSS is an international large-scale educational assessment used to monitor mathematics and science achievement among representative samples of fourth- and eighth-grade students across different countries. Besides the achievement measures, TIMSS also collects information about the home, school, and classroom context for learning. As mentioned before, in this study we focus on the non-achievement part of the assessment, with a specific focus on the student questionnaire. Besides some basic demographics and background information, the main focus of the student questionnaire lies with students' attitudes towards learning mathematics and science (Mullis & Martin, 2013).

Assessment Duration. For the eighth grade, the achievement test of TIMSS consisted of two sections (i.e., focus on mathematics or science). For each of these sections the testing time was set at 45 minutes with a 30-minute break in between (e.g., Mullis & Martin, 2013). Only after a second break, the student questionnaire was administered as a third section. The student questionnaire was administered to every student that took part in the TIMSS 2015 achievement test. The testing time for the student questionnaire was set at 30 minutes. The total testing time for an eighth-grade student in the TIMSS 2015 assessment (i.e., all 3 sections) is then 120 minutes in total plus the time for the two breaks (Mullis & Martin, 2013). Students were not allowed to leave the room or start with a new section even if they had already completed the task within the set time frame (Martin et al., 2016). Hence, there is no reward for rushing through the assessment as students had to remain seated in class and everyone also gets the same break time.

Student Questionnaire Length

For the eighth grade, there are two versions of the student questionnaire. The version that is administered depends on the science curriculum program within a country. One version is for countries teaching science as a single or general subject (i.e., integrated science program), while the other version is for countries teaching science as separate subjects (i.e., separated science program). This distinction between the science programs also comes with natural variation in questionnaire length as implied by the different number of survey scales within the specific versions of the questionnaire. The separated science program has the longer questionnaire (i.e., 19 scales) with an additional 9 scales compared to the student questionnaire for the integrated science program (i.e., 10 scales).

Student Questionnaire Scales

The student questionnaires contain survey scales related to the following domains: school climate for learning, school safety, and student engagement and attitudes towards mathematics or science (Martin et al., 2016) (for information on the specific scales see Table 1). The three scales affected by the structure of the science program are the "Students Like Learning Science", "Students' View on Engaging Teaching in Science Lessons" and "Students Confident in Science" scales. For countries with an integrated science program, each of these scales only appears once, while for countries with a separated science program each of these scales is available for each science domain separately (i.e., in order of appearance: Biology, Earth Science, Chemistry, and Physics). The science scales in both student questionnaires do have the same structure. For the items in the separated student questionnaire, it is just the word 'science' that is replaced by the name of the specific science domain (e.g., 'I enjoy learning science' vs 'I enjoy learning chemistry').

The set of scales contains between 7 and 10 items for each scale, for which a student needed to indicate to what extent s/he agrees with the given statement or indicate how often a specific situation has occurred to them on a 4-point Likert scale, ranging from 1 (*agree a lot* or *at least once a week*) to 4 (*disagree a lot* or *never*).

Table 1

Overview of Scales in the TIMSS 2015 Student Questionnaire.

Scale	Items	Response Options	Position	
			ISP	SSP
Domain: School climate				
Students' sense of school belonging	7	1 (agree a lot) – 4 (disagree a lot)	0	0
Domain: School safety				
Student bullying	9	1 (at least once a week) – 4 (never)	1	1
Domain: Student engagement and attitudes				
Students like learning mathematics	9	1 (agree a lot) – 4 (disagree a lot)	2	2
Students' views on engaging teaching in mathematics lessons	10	1 (agree a lot) – 4 (disagree a lot)	3	3
Students confident in mathematics	9	1 (agree a lot) – 4 (disagree a lot)	4	4
Students value mathematics	9	1 (agree a lot) – 4 (disagree a lot)	5	5
Students like learning science*	9	1 (agree a lot) – 4 (disagree a lot)	6	{6, 9, 12, 15}
Students' views on engaging teaching in science lessons*	10	1 (agree a lot) – 4 (disagree a lot)	7	{7, 10, 13, 16}
Students confident in science*	8	1 (agree a lot) – 4 (disagree a lot)	8	{8, 11, 14, 17}
Students value science	9	1 (agree a lot) – 4 (disagree a lot)	9	18

Note. *For this scale, there is a distinction between countries with an integrated or a separated science program, respectively referring to one general scale related to science as a single subject or to four separate scales related to each of the specific science subjects. In the questionnaire for countries with a separated science program, these scales are grouped per domain and appear in the following order: biology, earth science, chemistry, and physics. In the corresponding statements "science" is then replaced by the specific subject name. ISP = integrated science program; SSP = separated science program. Note that the first scale at position 0 represents the first substantive scale after 14 more general background questions.

Scale Position

Scale position is defined by the order in which the survey scales appear in the student questionnaire. Starting at position zero is the first substantive scale (i.e., students' sense of school belonging) that followed after 14 more general questions about students' background. After this first scale, the other survey scales followed in succession in the questionnaire. An overview of the survey scales and their position in each version of the student questionnaire can be found in Table 1.

TIMSS Sample: Countries

All 40 regular participating countries that administered the eighth-grade TIMSS assessment in 2015 or 2016 have been included in the analyses. Note that some countries used the opportunity to administer the TIMSS assessment to the ninth grade instead of the regular eighth grade for better comparability with curricula (i.e., Botswana and South Africa), for better comparability of results with other countries (i.e., Norway) or to better match the TIMSS age criteria (i.e., England and New Zealand) (e.g., Martin et al., 2016). Of the included countries, 29 teach an integrated science program, while the other 11 countries teach a separated science program[20]. In what follows, we will refer to the countries by the ISO country codes as used in the TIMSS data files (see also footnote 20).

Prevalence of Random Responders

A mixture item response theory model framework (Mislevy & Verhelst, 1990; Sen & Cohen, 2019; Yamamoto, 1989) was adopted to operationalize and define the target outcome variable of interest $PREV(RR)$, the prevalence of random responders on a particular survey scale. The approach by van Laar and Braeken (2022) assumes that there are two distinct, yet unobserved latent groups of responders in the population expressing

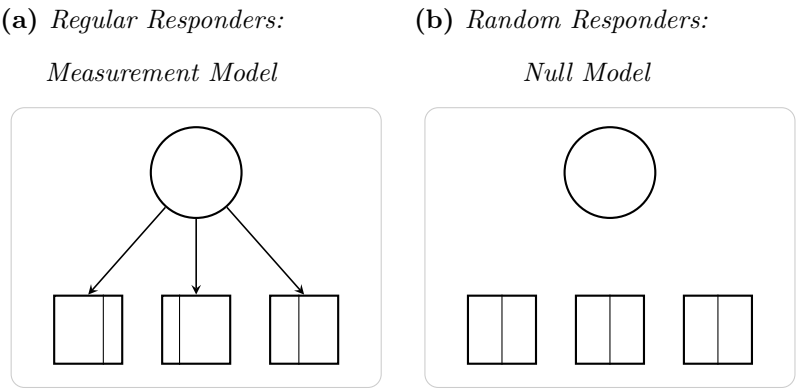
[20]Integrated Science Program: Australia, AUS; Bahrain, BHR; Botswana, BWA; Canada, CAN; Chile, CHL; Chinese Taipei, TWN; Egypt, EGY; England, ENG, Hong Kong SAR, HKG; Iran, Islamic Rep. of, IRN; Ireland, IRL; Israel, ISR; Italy, ITA; Japan, JPN; Jordan, JOR; Korea, Rep. of, KOR; Kuwait, KWT; Malaysia, MYS; New Zealand, NZL; Norway, NOR; Oman, OMN; Qatar, QAT; Saudi Arabia, SAU; Singapore, SGP; South Africa, ZAF; Thailand, THA; Turkey, TUR; United Arab Emirates, ARE; United States, USA.

Separated Science Program: Armenia, ARM; Georgia, GEO; Hungary, HUN; Kazakhstan, KAZ; Lebanon, LBN; Lithuania, LTU; Malta, MLT; Morocco, MAR; Russian Federation, RUS; Slovenia, SVN; Sweden, SWE.

different response behavior on a survey scale: regular or non-random responders and random responders (see Figure 1).

Figure 1

Mixture IRT model Framework to Define and Operationalize Random Responders in terms of Independence and Uniformity of Item Responses.



Note. Symbols follow standard path diagram conventions, with squares representing observed variables (i.e., item responses); circles, latent variables (i.e., trait to be measured by the scale of items); arrows indicating dependence relations; vertical lines, response category thresholds. Reprinted under the terms of CC-BY-NC from "Random responders in the TIMSS 2015 student questionnaire: A threat to validity?" by S. van Laar and J. Braeken, 2022, *Journal of Educational Measurement*.

The regular responders are expected to respond consistently according to their own opinions and beliefs related to the questionnaire content of the items on the scale, in line with a traditional latent variable measurement model (see Figure 1a, the 'circle' is the common cause of the 'squares') such as the graded response model (Samejima, 1969). In contrast, the random responders are expected to provide responses that do not reflect their opinions and beliefs, but are more haphazard, in line with a null model implying independent item responses that have an equal chance of falling in either of the possible response categories (see Figure 1b, the 'squares' are mutually disconnected, nor influenced by the 'circle'; all squares are divided into uniformly equal category parts).

Under the mixture IRT model, the likelihood of a person p's item response pattern $\mathbf{y}_p$ (see Equation 1) is written as a weighted sum of the two mentioned model expressions: the joint probability of the observed item response pattern given the person's latent trait

value under the graded response model multiplied by $\Pr(\backslash RR)$ the prior probability for a person to be a member of the regular responder group plus the joint probability of the observed item response pattern under the null model multiplied by $\Pr(RR)$ the prior probability for a person to be a member of the random responder group.

$$\begin{aligned} \mathcal{L}(\mathbf{Y}_p = \mathbf{y}_p) = & \\ & \Pr(\backslash RR) \quad \textstyle\prod_i \Pr(Y_{pi} = y_{pi} | \theta_p, \backslash RR) \\ & + \\ & \Pr(RR) \qquad \textstyle\prod_i \Pr(Y_{pi} = y_{pi} | RR) \end{aligned} \tag{1}$$

Although seemingly much more complex, this mixture model in fact has only one additional parameter[21] when compared to the regular measurement model. This parameter $\Pr(RR)$ can be interpreted as a model-based estimate of the prevalence of random responders on the survey scale for the item response data the mixture model is applied to.

Thus, this mixture IRT model was estimated for each of the scale-country combinations in the current study. The resulting estimates for the mixture weight $\widehat{\Pr}(C = RR)$ will be used as an estimate of the prevalence of random responders on the survey scale for that country, and hence is the actual outcome variable $PREV(RR)$ for further analyses targeting our core research questions. If the mixture model for a specific country-scale combination failed either of two quality checks, the corresponding outcome was set to missing. First, the measurement model for the regular responders in the mixture was inspected to ensure that it reflected a clean unidimensional model (i.e., compatible with the assumed common trait for the survey scale). This criterion was not met when two or more standardized item discrimination parameters (i.e., factor loadings) were below .40. Secondly, a classification entropy of at least .70 was required to ensure that the mixture model was able to provide a good enough distinction between the two latent groups of responders. To further assess model adequacy we gathered model comparison

[21] The part of the model accommodating the possibility of random responders in the population has no unknown parameters as item response probabilities are known and assumed to be uniformly equal across categories and items, such that only the mixture weights $\Pr(RR)$ and $\Pr(\backslash RR)$ remain as extra model parameters, which reduces to one given that $\Pr(RR) + \Pr(\backslash RR) = 1$.

evidence using BIC and BIC weights (Nylund et al., 2007; Wagenmakers & Farrell, 2004) contrasting the null model with the graded response model and the mixture IRT model.

Statistical Analysis

A cross-classified linear mixed model approach was adopted to investigate how the prevalence of random responders on a survey scale varied as a function of the scale's position in the questionnaire and the length of the student questionnaire it is part of. The study design has a cross-classified cluster structure as multiple prevalence estimates are observed within each country (i.e., across scales), but also for each survey scale multiple prevalence estimates are observed (i.e., across countries). As a consequence, the outcome variable $PREV(RR_{cs})$ in the model is the random responder prevalence for a given country c on a given scale s, reflecting the countries-by-scales cross-classification. A series of five models was fitted to investigate the main research questions. As a baseline model we used a varying-intercepts model ($\mathcal{M}_0$) capturing variation in the prevalence of random responders across countries and scales, accounting for the heterogeneity and dependence structure implied by the cross-classified study design:

$$\begin{aligned} PREV(RR_{cs}) &= \beta_0 + \beta_{0c} + \beta_{0s} + \varepsilon_{cs} \\ \beta_{0c} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{country}) \\ \beta_{0s} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{scale}) \\ \varepsilon_{cs} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{residual}) \end{aligned} \tag{2}$$

The general intercept β_0 reflects the average expected prevalence of random responders for an average country on an average scale. The country-varying (random) intercept β_{0c} and scale-varying (random) intercept β_{0s} allow for a systematic deviation in the prevalence for a specific country c or specific scale s, respectively. The residual ε_{cs} allows for unexpected deviations in prevalence for a specific country-scale combination not accommodated by both country and scale main effects in the model.

Our core research questions would imply that when adding scale position and questionnaire length as predictors to this model, variation in both features would be related to the systematic variation in prevalence of random responders across scales (i.e., σ^2_{scale}).

Hence, the four models building on the presented baseline model will incrementally add both predictors (and their interaction) to the equation. The percentage of systematic variation in prevalence among the survey scales that is accounted for by the predictors (i.e., R^2_{scale}) will be used as a general effect size measure for each model.

Statistical Software. The mixture IRT models were estimated using Mplus Version 8.2 (Muthén & Muthén, 1998–2017) through the MplusAutomation package for R version 0.7-3 (Hallquist & Wiley, 2018) (for an example of Mplus syntax see Appendix A). We used full-information maximum likelihood estimation with robust standard errors and the expectation-maximization acceleration algorithm with a standard of 400 random starts, 100 final stage optimizations, and 10 initial stage iterations. Mixture model estimates accounted for the TIMSS sampling design through the total student weights. The cross-classified mixed models were estimated using the lme4 package for R version 1.1-27 (Bates et al., 2015). As recommended by Snijders and Bosker (2012) we used residual maximum likelihood estimation for estimation of the model parameters, but maximum likelihood estimation for model comparison inference by means of likelihood ratio tests. All analysis scripts were run under R version 4.0.0 (R Core Team, 2020).

Results

Descriptives

Data. Given that 29 countries teach an integrated science program and 11 countries teach a separated science program the study started with 499 country-scale combinations. However, prevalence estimates are not available for all combinations. For 7 combinations this was related to data collection procedures (i.e., the scale was not administered or the data is not available for public use), while 35 combinations did not fulfill the mixture model quality checks (for an overview see Table 2). Together, this results in an effective sample size of 457 country-scale combinations for further analyses. Across all 457 combinations, the null model is never supported (BIC weight = 0 for all, average BIC = 145747) and the model comparison evidence is close to unanimously in favor of the mixture IRT model (average BIC = 90608; BIC weight = 1 for 435 combinations), with

the regular graded response model (average BIC = 91855) being favored in only 13 combinations (all representing the 'Student bullying' scale with prevalence estimates below 1%). On average 93% (range: 70–100%) and 89% (range: 47–100%) of the scales have an effective prevalence estimate for countries with an integrated science and a separated science program structure, respectively. For survey scales shared by both science programs, prevalence estimates are available for 95% (75–100%) of the countries; for scales unique to the integrated science and separated science programs, this comes down to 90% (69–100%) and 85% (73–100%) of the corresponding countries, respectively. In sum, we have a solid empirical basis for further analyses.

Baseline Model $\mathcal{M}_0$. The estimated prevalence of random responders on an average scale ranged from 6.3% to 15.4% ($M = 8.9\%$) across countries. The estimated prevalence for an average country ranged from 1.9% to 20.2% ($M = 8.9\%$) across scales. The variation in prevalence ($\hat{\sigma}^2_{total} = 37.8$) was primarily due to systematic differences between scales ($\hat{\sigma}^2_{scale} = 25.5$, 67% of the total variance) and only to a lesser extent to systematic differences between countries ($\hat{\sigma}^2_{country} = 3.8$, 10% of the total variance).

The systematic variation in the prevalence of random responders and how it relates to scale position and questionnaire length will be discussed in the next subsection, but first the systematic variation across countries will be briefly addressed. The expected prevalence of random responders for an average country on an average scale was estimated to be about $\hat{\beta}_0 = 8.9\%$. Yet on average, Georgia, Qatar, and Armenia showed higher levels of random responders across scales, while Russia, Australia, Sweden, Kazakhstan, England, Canada, and Norway tend to show lower levels (see Figure 2).

Table 2

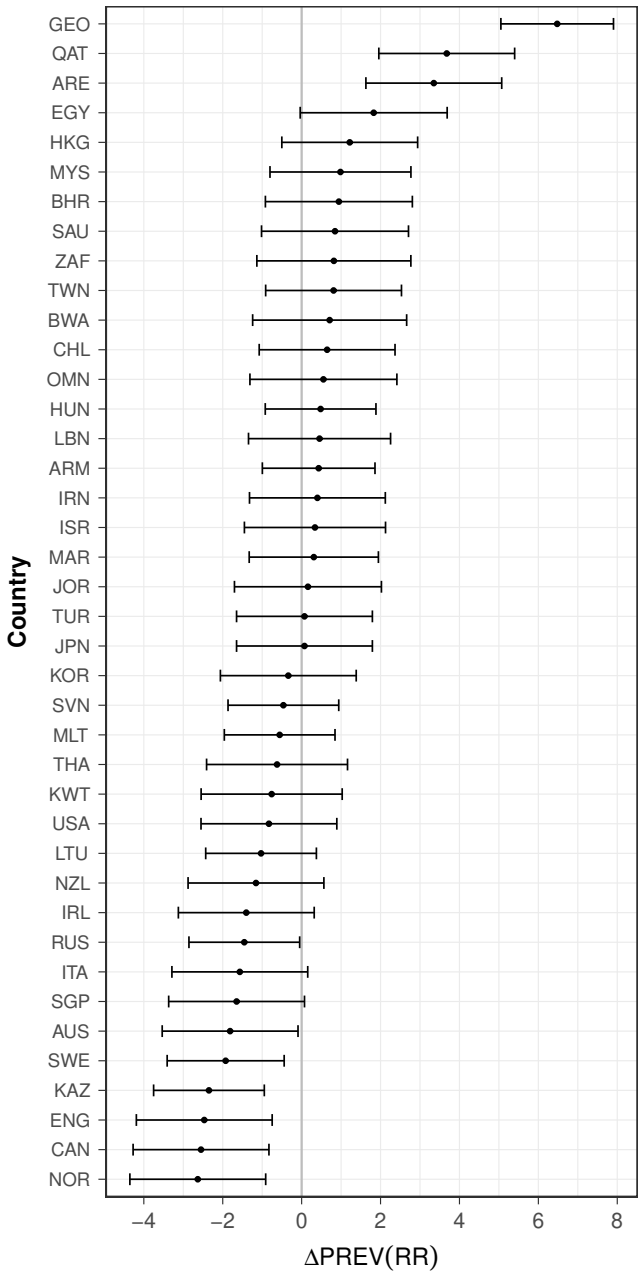
Overview of the number of excluded scales: Data collection procedures and/or lack of scale quality.

Integrated Science								Separated Science							
ISO	scales			excluded				ISO	scales			excluded			
	tot	N_{eff}	%	E	α^z	(E, α^z)	NA		tot	N_{eff}	%	E	α^z	(E, α^z)	NA
BHR	10	8	80%	1	1			ARM	19	18	95%	1			
BWA	10	7	70%	3				GEO	19	18	95%		1		
EGY	10	8	80%		1	1		LBN	19	9	47%	1	6		3
ISR	10	9	90%				1	MAR	19	12	63%	1	6		
JOR	10	8	80%		2			SWE	19	16	84%				3
KWT	10	9	90%	1											
MYS	10	9	90%		1										
OMN	10	8	80%		2										
SAU	10	8	80%			2									
ZAF	10	7	70%	2	1										
THA	10	9	90%		1										

Note. tot = total number of scales expected for a country within the corresponding science program; N_{eff} = the number of included scales for a country; % = percentage of included scales for a country. Quality criteria for the measurement model for the regular responders in the mixture: E = number of scales excluded due to the classification entropy being below .70; α^z = number of scales excluded due to two or more standardized item discrimination parameters being below .40; (E, α^z) = number of scales excluded due to the quality criteria for standardized item discrimination parameters and/or classification entropy not being met; NA = number of scales excluded due to the scale not being administered or the data not being available for public use. For the integrated science program, all scales were administered as normal and the quality criteria were met for the following 18 countries: AUS, CAN, CHL, TWN, ENG, HGK, IRN, IRL, ITA, JPN, KOR, NZL, NOR, QAT, SGP, TUR, ARE, and USA. For the separated science program, this was the case for the following 6 countries: HUN, KAZ, LTU, MLT, RUS, and SVN.

Figure 2

Differences in Prevalence of Random Responders across Countries.



Note. The vertical gray line represents the prevalence of random responders for an average country on an average scale under the baseline model $\mathcal{M}_0$ ($\hat{\beta}_0 = 8.9\%$). The black horizontal lines are 95% confidence intervals of the country-specific deviations in prevalence (ΔPREV(RR)) to that average.

Prevalence(RR_{cs}) = Scale Position × Questionnaire Length

It was expected that survey scales at a later position would display higher prevalence rates. The model results (see Table 3) indicated that the expected difference in prevalence rate as a function of differences in scale position was positive, yet not significantly different from zero ($\beta_1 = .12, \chi^2_{(\mathcal{M}0,\mathcal{M}1)}(1) = 1.28, p = .257$, $R^2_{scale} = 10.1\%$). The longer questionnaire was expected to display higher prevalence rates, yet no empirical support was found for this hypothesis ($\beta_2 = .06, \chi^2_{(\mathcal{M}0,\mathcal{M}2)}(1) = 0.01, p = .935$, $R^2_{scale} = 0\%$). Considering both scale position and questionnaire length jointly as predictors in the model, lead to similar results ($\chi^2_{(\mathcal{M}0,\mathcal{M}3)}(2) = 1.31, p = .521$, $R^2_{scale} = 10.2\%$), and no support for the hypothesized synergistic interaction was found either ($\chi^2_{(\mathcal{M}3,\mathcal{M}4)}(1) = .88, p = .349$, $R^2_{scale} = 16.9\%$).

Overall these results were not in line with expectations. However, when visualizing the data, an unexpected but impactful factor for the prevalence of random responders appears (see Figure 3). The black lines in Figure 3 show the country trends of the prevalence of random responders across scales in the student questionnaire. What becomes visible is that the prevalence rates show a systematic occurrence of several spikes throughout the survey in each of the countries. Two spikes occur for the integrated science program, and three more spikes (i.e., 5 in total) occur in the longer questionnaire of the separated science program. The locations of these spikes in prevalence are not randomly distributed but coincide with the locations of the confidence scales in the questionnaire. In the integrated science program the spike in prevalence occurs for both confidence scales (i.e., mathematics and science), and in the separated science program for all five confidence scales (i.e., mathematics, biology, earth science, chemistry, and physics). Given the clear impact of the confidence scales on the prevalence patterns, it makes sense to take this factor into consideration and to revisit our hypotheses adjusting for this unexpected confounder.

Table 3

Cross-classified mixed models of the prevalence of random responders as a function of scale position and questionnaire length.

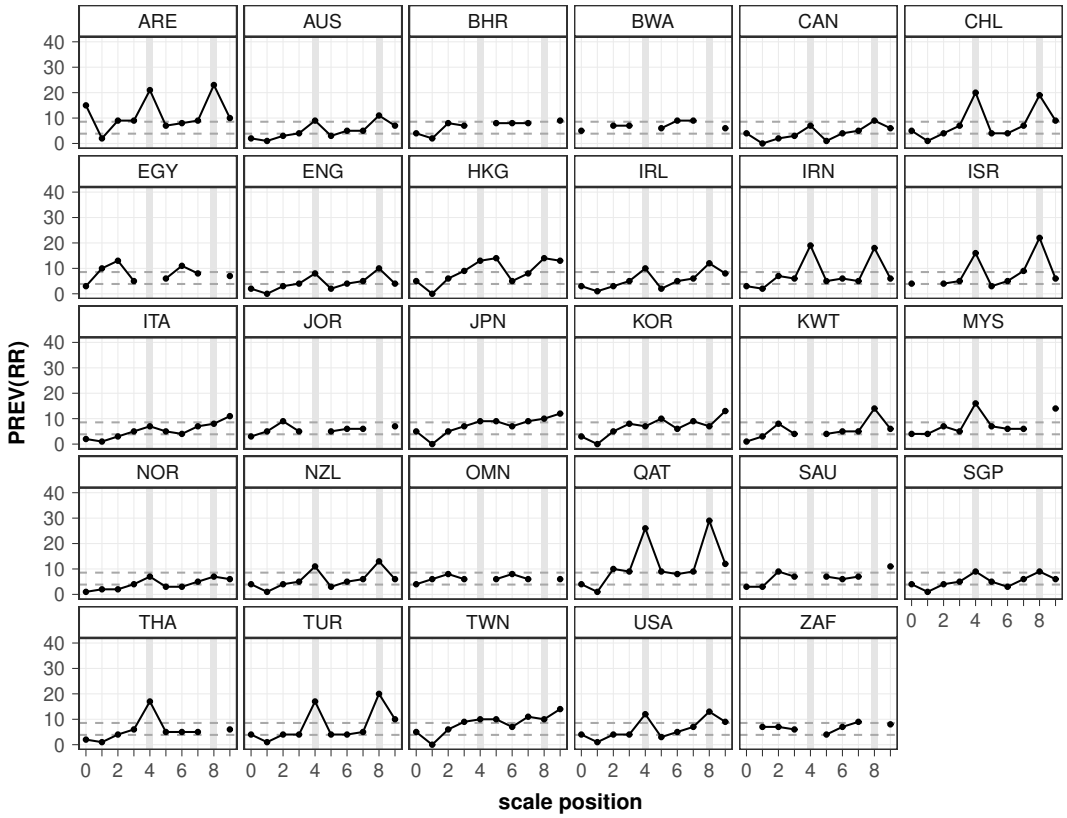
	$\mathcal{M}_0$		$\mathcal{M}_1$		$\mathcal{M}_2$		$\mathcal{M}_3$		$\mathcal{M}_4$	
	Parameter Estimates									
predictor	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE
β_0: intercept	8.90	1.14	7.92	1.39	8.88	1.18	7.93	1.39	6.54	2.00
β_1: position			0.12	0.11			0.13	0.11	0.39	0.30
β_2: length					0.06	0.81	-0.12	0.82	0.33	0.96
β_3: position×length									-0.17	0.19
	Variance Components									
$\sigma^2_{country}$	3.8		3.9		4.0		4.0		4.0	
σ^2_{scale}	25.5		22.9		25.5		22.9		21.2	
$\sigma^2_{residual}$	8.5		8.5		8.5		8.5		8.6	
R^2_{scale}			10.1%		0.0%		10.2%		16.9%	
npar	4		5		5		6		7	
-2LL	2428.2		2426.9		2428.2		2426.9		2426.1	

Note. npar = number of parameters; -2LL = deviance; R^2_{scale} = percentage of systematic variation in the prevalence of random responders across scales under $\mathcal{M}_0$ that can be attributed to differences in the predictor(s) in the corresponding model ($\mathcal{M}_1$ - $\mathcal{M}_4$). Located at position zero is the first substantive scale that followed after 14 more general background questions. Length is a binary variable differentiating between the shorter student questionnaire with 10 survey scales (i.e., length = 0) and the longer student questionnaire with 19 survey scales (i.e., length = 1).

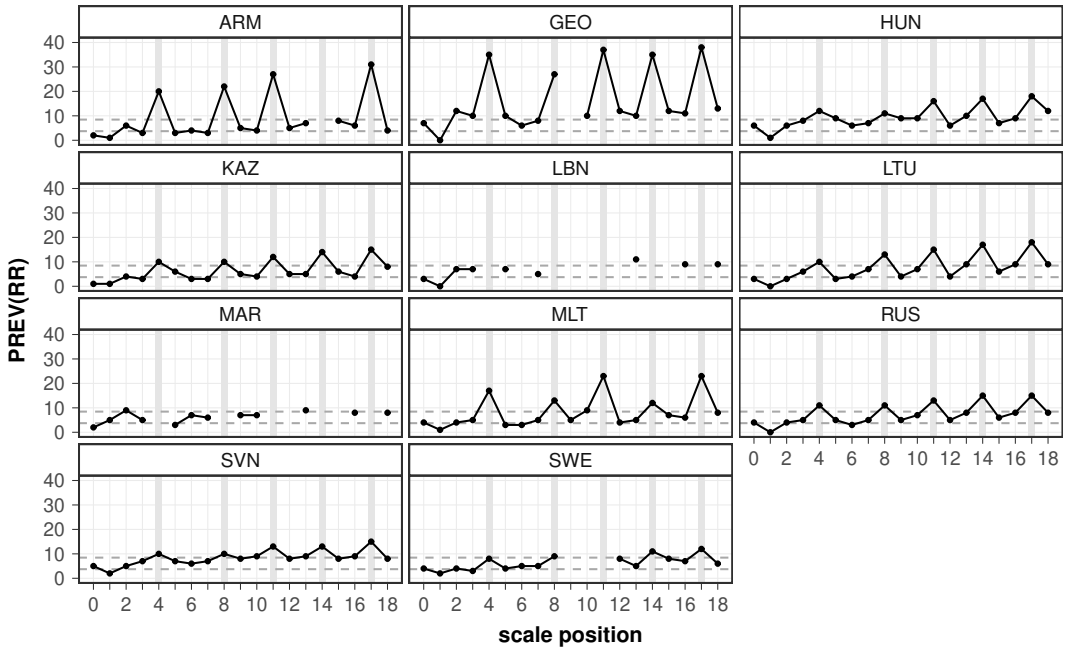
Figure 3

Observed Prevalence of Random Responders per Country across Scales by Science Program.

(a) *Integrated Science Program*



(b) *Separated Science Program*



Note. The solid black line represents the observed prevalence of random responders across scales. The dashed grey lines represent the average prevalence on the first and the last scale across countries in the corresponding questionnaire. The spikes in prevalence are related to the confidence scales; within the integrated science questionnaire located at position 4 and 8 and within the separated science questionnaire located at position 4, 8, 11, 14, and 17 (i.e., indicated by the grey vertical bars).

Table 4

Revisited with confidence in mind: Cross-classified mixed models of the prevalence of random responders as a function of scale position and questionnaire length.

	$\mathcal{M}_{0c}$		$\mathcal{M}_{1c}$		$\mathcal{M}_{2c}$		$\mathcal{M}_{3c}$		$\mathcal{M}_{4c}$	
	Parameter Estimates									
predictor	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE
β_0: intercept	6.14	0.54	4.57	0.51	6.22	0.58	4.70	0.52	3.53	0.61
$\beta_{confidence}$	9.89*	1.39	9.23*	1.20	9.88*	1.39	9.17*	1.19	8.79*	1.17
β_1: position			0.22*	0.05			0.23*	0.05	0.50*	0.10
β_2: length					-0.25	0.57	-0.67	0.57	-0.04	0.62
β_3: position×length									-0.22*	0.08
	Variance Components									
$\sigma^2_{country}$	2.1		2.1		2.1		2.1		2.1	
$\sigma^2_{confidence}$	30.9		31.3		30.9		31.3		31.7	
ρ	0.26		0.24		0.27		0.27		0.27	
σ^2_{scale}	3.6		1.3		3.6		1.3		0.9	
$\sigma^2_{residual}$	3.1		3.1		3.1		3.1		3.1	
ΔR^2_{scale}			8.7%		0.0%		9.0%		10.3%	
npar	7		8		8		9		10	
-2LL	2046.7		2035.4		2046.5		2034.0		2025.9	

Note. npar = number of parameters; -2LL = deviance; ρ = correlation between country-varying coefficient for confidence and the country-varying intercept; ΔR^2_{scale} = percentage of reduction in σ^2_{scale} under $\mathcal{M}_0$ uniquely attributed to the difference in the predictor(s) in the model beyond confidence. Calculated as the difference between the reduction attributed to the combined effect of confidence and the predictor(s) in the corresponding model ($\mathcal{M}_{1c}$ - $\mathcal{M}_{4c}$) and the reduction attributed to confidence on its own in $\mathcal{M}_{0c}$; * = $p < .05$. Confidence is a binary variable differentiating between non-confidence (i.e., confidence = 0) and confidence scales (i.e., confidence = 1). Located at position zero is the first substantive scale that followed after 14 more general background questions. Length is a binary variable differentiating between the shorter student questionnaire with 10 survey scales (i.e., length = 0) and the longer student questionnaire with 19 survey scales (i.e., length = 1).

Hypotheses Revisited with Confidence in Mind

To account for the spikes in prevalence, we added a binary predictor variable differentiating between non-confidence (i.e., confidence = 0) and confidence (i.e., confidence = 1) scales to the model. As Figure 3 also showed that the degree of irregularity for the confidence scales varied across countries, we allowed for a country-varying (random) coefficient for confidence with mean $\beta_{confidence}$ and variance $\sigma^2_{confidence}$ and potentially correlated with the country-varying intercept β_{0c}. Model results are summarized in Table 4.

The prevalence of random responders on an average non-confidence scale for an average country was estimated to be $\hat{\beta}_0 = 6.14\%$, whereas the corresponding prevalence for an average confidence scale was expected to be $\hat{\beta}_{confidence} = 9.89\%$ higher ($\chi^2_{(\mathcal{M}0,\mathcal{M}0c)}(3) = 381.51, p < .001$, $R^2_{scale} = 86.1\%$). Hence, there is clear statistical support for a systematic spike in the prevalence of random responders on the survey scales measuring confidence. When compared to the variation in prevalence across countries for non-confidence scales $\hat{\sigma}^2_{country} = 2.1$, the corresponding variation across countries for the difference between confidence and non-confidence scales is more sizeable $\hat{\sigma}^2_{confidence} = 31.3$. The latter result reflects the differences in height of the spikes in the different countries in Figure 3, whereas the baseline prevalence trends are more similar in nature. There was no clear pattern between country differences in prevalence heights for non-confidence scales and country differences in prevalence spike heights for confidence scales (i.e., $\hat{\rho} = .26$).

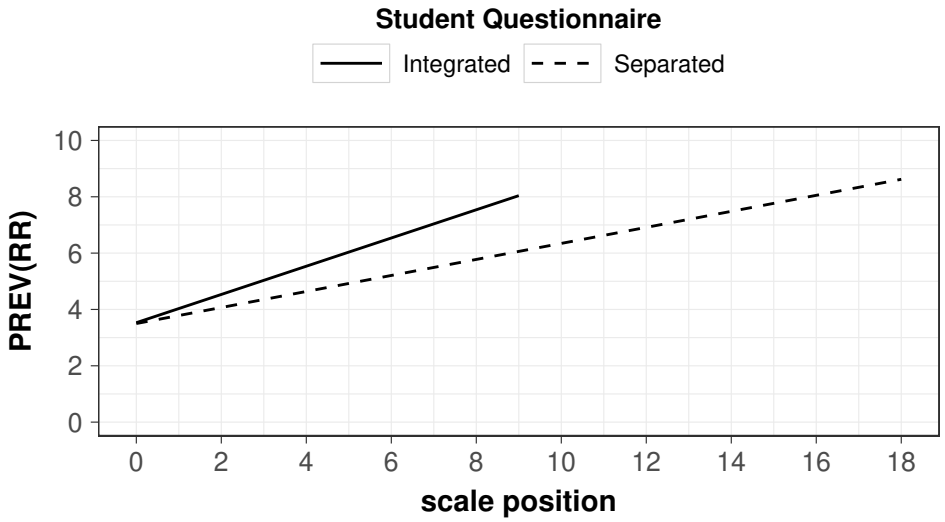
The addition of the new binary predictor effectively detrends the prevalence patterns across the student questionnaire for the unexpected spike pattern due to the confidence scales, allowing us to revisit the original hypotheses adjusting for that systematic distortion. As effect size measure ΔR^2_{scale} will be used, the difference between the model's R^2_{scale} and the reference R^2_{scale} under $\mathcal{M}0c$ the baseline model augmented with the new binary confidence predictor. Hence, this measure will quantify the unique contribution of scale position and questionnaire length to systematic variation across scales in the prevalence of random responders beyond what is accounted for by the confidence spike pattern.

Accounting for the confidence spikes, a scale located five positions later in the questionnaire is expected to have about 1% higher prevalence of random responders than

the earlier scale ($\beta_1 = .22, \chi^2_{(\mathcal{M}0c,\mathcal{M}1c)}(1) = 11.31, p < .001$, $\Delta R^2_{scale} = 8.7\%$); a result supporting Hypothesis 1. Yet, no statistically significantly higher prevalence was found in countries teaching the separated science program when compared to those with the integrated science program ($\beta_2 = -.25, \chi^2_{(\mathcal{M}0c,\mathcal{M}2c)}(1) = .19, p = .666$, $\Delta R^2_{scale} = 0\%$), and hence no empirical support was found for Hypothesis 2 that the prevalence of random responders would be a function of questionnaire length. When looking jointly at scale position and questionnaire length, there was support found for an interaction ($\chi^2_{(\mathcal{M}3c,\mathcal{M}4c)}(1) = 8.11, p = .004$, $\Delta R^2_{scale} = 10.3\%$), yet not one of the hypothesized synergistic type. Figure 4 illustrates that in contrast to expectations the differences in prevalence between scales at subsequent positions are estimated to be larger in the shorter questionnaire than in the longer questionnaire. Notice that regardless of the length of the questionnaire, the prevalence estimate for an average survey scale at the first and last position in the respective questionnaire is estimated to be about 3.5% and 8.5%, respectively.

Figure 4

Prevalence of random responders as a function of scale position and questionnaire length in TIMSS 2015 under the cross-classified mixed model $\mathcal{M}_{4c}$.



Note. The TIMSS 2015 student questionnaire consisted of 10 survey scales for countries with an integrated science program, whereas it consisted of 19 survey scales for countries with a separated science program. For the model parameters of model $\mathcal{M}_{4c}$, see Table 4.

Discussion

The aim of the present study was to investigate the impact of two questionnaire characteristics, scale position and questionnaire length, on the prevalence of random responders in the TIMSS 2015 eighth-grade student questionnaire. Although random responders still provide responses to the items of a questionnaire scale, their responses can be seen as a type of non-response, as they would not lead to valid inferences on their actual attitudes or beliefs that were intended to be measured. It has been suggested that as students progress through a questionnaire they will experience for example boredom, disinterest, inattentiveness, or fatigue and consequently engage in random responding. Accordingly, a higher prevalence of random responders was hypothesized for scales at a later position in the questionnaire and for the longer version of the two questionnaires, and an even higher prevalence for later scales in the longer questionnaire (i.e., a synergistic interaction between scale position and questionnaire length).

Questionnaire Length. We found no clear difference in the prevalence of random responders between the longer student questionnaire administered in countries with a separated science program and the shorter student questionnaire administered in countries with an integrated science program. In a similar fashion, Boe et al. (2002) also didn't find an effect of questionnaire length when they looked at student response omission rates (labeled 'task persistence') in the TIMSS 1995 student questionnaire. A skeptical interpretation could attribute this finding to the difference in countries between the two questionnaire versions, but it has also been suggested that most educational and psychological questionnaires are just not long enough to find an effect on response quality to begin with (e.g., Bowling et al., 2022). In broader survey situations where there is a larger time and length difference, questionnaire length does seem to have an effect (e.g., Herzog & Bachman, 1981). Yet the mixed results in the literature with respect to questionnaire length suggest that actual effects will also depend on (i) the content or context of the specific questionnaire under consideration (e.g., Gibson & Bowling, 2020; Rolstad et al., 2011) and (ii) on the subjectively perceived length of the questionnaire instead of its actual length (Helgeson et al., 2002). Although there had been some complaints

reported by the students about the length of the student questionnaire (Martin et al., 2016), the current null finding with respect to questionnaire length does seem to suggest that the differences in test burden and testing time for the two versions of the TIMSS 2015 student questionnaires were kept within seemingly reasonable boundaries.

Scale Position. We found support for a scale position effect with a significantly higher prevalence of random responders for scales at a later position in the questionnaire compared to scales at an earlier position. Over the course of both questionnaires, the prevalence of random responders increased by 5%, from 3.5% on an average scale at the beginning to 8.5% at the end of the student questionnaire. The effect of scale position actually being stronger within the shorter version of the student questionnaire contrasted with the hypothesized synergism which would have implied the opposite trend. Galesic (2006) suggests that again students' relative perception of the questionnaire plays a role. Hence, students might consider scale position being considered relative to the perceived length of the questionnaire. Relatively speaking, with every additional scale in the shorter questionnaire more of the questionnaire has passed percent-wise (i.e., the progress signified by 1 scale is 10% in the shorter questionnaire compared to 5% in the longer questionnaire). This might have potentially influenced the students' subjective perception of how much they already had completed and how much was still left, and influenced how they would engage with further scales in the questionnaire[22].

The Case of the Confidence Scales. The most striking result with respect to the prevalence of random responders across the student questionnaire were the spikes in prevalence among all confidence scales (i.e., mathematics and science subject domains) with on average an extra 9% prevalence compared to other scales. This difference due to the specific scale character is double the size of the above-discussed 5% prevalence difference due to the maximal scale position difference. The implication of this finding is that random responder prevalence is not only depending on the 'endurance' of the students throughout the questionnaire. So what is so special about the confidence scales

[22]Note that students are only familiar with the version of the student questionnaire administered to them, they are not able to compare the length with the other version and as such have no baseline but their own perception.

that they elicit more random response behavior? Focusing on the characteristics of the confidence scales might provide some indications of what is going on.

First, the confidence scales are mixed-worded scales. It has been argued that reversed-worded items are more difficult to process (e.g., Marsh, 1986; Swain et al., 2008). Although the confidence scales are not the only mixed-worded scales in the student questionnaire, they do have the largest amount of reversed-worded items (e.g, 4 out of 9 reversed-worded items for confidence in mathematics compared to 2 out 9 items for the like-learning scales) which could contribute to a larger impact (e.g., Schmitt & Stults, 1985).

A second characteristic to consider is the type of items in the confidence scales. Because some of the items are related to self-concept (e.g., Michaelides et al., 2019), one could argue that items are more comparative in nature as opposed to more absolute/factual. Important here is that perceptions students have about themselves are always made in comparison to some standard, either internally (i.e., own performance in one subject with own performance in another subject) or externally (own performance with the performance of other students) (e.g., Marsh & Hau, 2004). Examples of items administered in the student questionnaire are "mathematics is harder for me than any other subject" or "mathematics is more difficult for me than for many of my classmates". Items that require comparisons, with additional changing or ambiguous standards and definitions of self, might just be more difficult to answer.

Both speculative explanations touch upon extra cognitive processing demands and perceived ambiguity or difficulty of the items in the confidence scales. This would be in line with the study by Baer et al. (1997) where the core reasons given by participants for random responding were difficulties in understanding items and difficulties in deciding on the response, in contrast to for instance lapses of concentration or boredom. Yet, these more abstract item characteristics are at the same time intertwined with the concrete scale contents 'confidence in a school subject'. On the upside, the fact that the confidence spikes generalized across different participating countries in TIMSS 2015 implies some generality of the underlying reasons.

Although TIMSS is low-stakes in all participating countries and there are solid standardized procedures for (back)translation of the different scales and administration of the questionnaire as a whole, this of course does not cancel out any further interplay with national context, socio-cultural aspects, language connotations, and differences in motivation and implicit communication surrounding TIMSS. Such contextual differences are reflected in the observed variability across countries in the average prevalence of random responders. Also when looking at the Confidence scales, the spikes in prevalence are for instance more outspoken in countries from the Middle-East region. Further research would need to dig into whether these scales are indeed eliciting more random responding or whether these questionnaire scales are being completely differently interpreted or approached by students in those regions than elsewhere[23].

The current study exploited the natural variation in scale position, questionnaire length, and scale characteristics found in the TIMSS 2015 student questionnaire, but to be able to clearly separate the influence of item characteristics and contents an experimental study would be called for in which item formulation of the questionnaire scales is systematically varied independently of scale contents. Yet, to implement such an experiment at a similar large-scale and level of generality as TIMSS 2015 might perhaps prove to be unrealistic. Complementary, we should also not dismiss the value of a more qualitative follow-up. Being classified as a random responder by the mixture IRT model does not imply that the student has deliberately responded randomly, but merely that the pattern of responses given is more random-like than it is consistent with the scale. Cognitive interviews and related techniques might provide insight into students' understanding and interpretation of the items in the confidence scales, into their processes to arrive at a response, but also into their feelings towards the scale contents in the questionnaire (e.g., Karabenick et al., 2007). Such research could potentially also shed light on other potential risk factors that have been put forward by Meade and Craig (2012) with respect to the general quality of responses (e.g., respondent interest, social contact,

[23]Note that among the 35 of 499 scale-country combinations not meeting the quality criteria for the application of the mixture IRT to random responder detection, 27 combinations involved confidence scales, of which 18 did not meet the standardized loadings criterion, implying weakness of the unidimensional measurement model for these cases.

and environmental distraction).

Conclusion

In sum, we conclude that one can indeed expect more students to engage in random responding on scales towards the end of the questionnaire in a large-scale educational assessment such as TIMSS. This seems likely related to more of a subjective relative evaluation for each individual, as in "aren't we there yet?", than to an objective physical criterion in terms of questionnaire length. Yet, when considering such response behavior, characteristics (item formulation and/or contents) of the questionnaire scales seem to be more crucial than expected. This implies that researchers and questionnaire designers want to better ensure that their target population is eager and willing to fully engage with the questions asked to increase response validity. The target population's subjective experience with the questionnaire can influence the quality of their responses given. We hope that the study's findings can contribute to convincing the organizations behind the international large-scale assessments in education of the value of investing in more extensive cognitive techniques and test panels. In general, an increased involvement of the target student population could benefit the design of the questionnaire scales.

References

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI–A. *Journal of Personality Assessment*, *68*(1), 139–151.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340.

Boe, E., May, H., & Boruch, R. (2002). *Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Achievement Differences at the National, Classroom, and Student Levels* (Research Report No. 2002-TIMSS1). University of Pennsylvania, Center for Research in Evaluation in Social Policy.

Bowling, N. A., Gibson, A. M., & DeSimone, J. A. (2022). Stop with the questions already! Does data quality suffer for scales positioned near the end of a lengthy questionnaire? *Journal of Business and Psychology.* Advance online publication.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(2), 718–738.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, *15*(1), 21–36.

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

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313–328.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, *36*(2), 410–420.

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, *5*(1), Article 18.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Helgeson, J. G., Voss, K. E., & Terpening, W. D. (2002). Determinants of mail-survey response: Survey design factors and respondent factors. *Psychology & Marketing*, *19*(3), 303–328.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549–559.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., De Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139–151.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49.

Marsh, H. W., & Hau, K.-T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/ex-

ternal frame of reference predictions across 26 countries. *Journal of Educational Psychology*, *96*(1), 56–67.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Merritt, S. M. (2012). The two-factor solution to Allen and Meyer's (1990) affective commitment scale: Effects of negatively worded items. *Journal of Business and Psychology*, *27*(4), 421–436.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*(3), 215–237.

Michaelides, M. P., Brown, G. T. L., Eklöf, H., & Papanastasiou, E. C. (2019). *Motivational Profiles in TIMSS Mathematics: Exploring Student Clusters Across Countries and Time*. Springer International Publishing.

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

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Boston College.

Muthén, L. K., & Muthén, B. O. (1998–2017). Mplus User's Guide. Eighth Edition.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.

R Core Team. (2020). R: A language and environment for statistical computing.

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101–1108.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191.

Snijders, T., & Bosker, R. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE Publications, Inc.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed likert items. *Journal of Marketing Research*, *45*(1), 116–131.

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

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.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205.

Yamamoto, K. (1989). Hybrid model of IRT and latent class models. *ETS Research Report Series*, *RR-89-41*.

Appendix A
Mplus syntax of mixture IRT model for the 'students value mathematics' scale in Norway

```
TITLE:
Norway_SQM20;

DATA:
file = "NOR_SQM20.dat";

VARIABLE:
  names = IDSCHOOL IDSTUD TOTWGT
    BSBM20A BSBM20B BSBM20C BSBM20D
    BSBM20E BSBM20F BSBM20G BSBM20H BSBM20I;
  missing = .;
  usevariables = BSBM20A BSBM20B BSBM20C BSBM20D
    BSBM20E BSBM20F BSBM20G BSBM20H BSBM20I;
  categorical = BSBM20A BSBM20B BSBM20C BSBM20D
    BSBM20E BSBM20F BSBM20G BSBM20H BSBM20I;
  idvariable = IDSTUD;
  weight = TOTWGT;
  cluster = IDSCHOOL;
  classes = c(2);

ANALYSIS:
  type = mixture complex;
  algorithm = INTEGRATION EMA;
  estimator = MLR;
  process = 3;
  starts = 400 100;

MODEL:
%overall%
  F BY BSBM20A-BSBM20I*;
  F@1;
  [F@0];
%c#1%
  F BY BSBM20A-BSBM20I*;
  F@1;
  [F@0];
  [BSBM20A$1-BSBM20I$1];
  [BSBM20A$2-BSBM20I$2];
  [BSBM20A$3-BSBM20I$3];
%c#2%
  F BY BSBM20A-BSBM20I@0;
  F@0;
  [F@0];
  [BSBM20A$1-BSBM20I$1@-1.09861228866811];
  [BSBM20A$2-BSBM20I$2@0];
  [BSBM20A$3-BSBM20I$3@1.09861228866811];

OUTPUT:
  stdyx;

SAVEDATA:
  file = cpr_NOR_SQM20.dat;
  format = free;
  save = cprobabilities;
```

Note. The item category threshold parameters in Class 2 (i.e., random responder class) are set on a logistic scale and correspond to cumulative response category probabilities of 25%, 50%, and 75% (i.e., 1/(1+exp(threshold))). A more detailed description of the model can be found in van Laar and Braeken (2022).

8 Article 5: Who

Chen, J., van Laar, S., & Braeken, J. (2022). *Who are those Random Responders on your Survey? The case of the TIMSS 2015 student questionnaire.* Manuscript under review.

Who are those Random Responders on your Survey? The case of the TIMSS 2015 student questionnaire

A general validity and survey quality concern with student questionnaires under low-stakes assessment conditions is that some responders will not genuinely engage with the questionnaire, often with more random response patterns as a result. Using a mixture IRT approach and a meta-analytic lens across 22 educational systems participating in TIMSS 2015, we investigated whether the prevalence of random responders on six scales regarding students' attitudes and beliefs in mathematics and sciences was a function of grade, gender, socio-economic status, spoken language at home, or migration background. Among these common policy-relevant covariates in educational research, we found support for small group differences in prevalence of random responders (OR $\geq$ 1.22)(average prevalence of 7%). In general, being a student in higher grades, being male, reporting to have fewer books, or speaking a language different from the test language at home were all risk factors characterizing random responders. The expected generalization and implications of these findings are discussed based on the observed heterogeneity across educational systems and consistency across questionnaire scales.

International large-scale assessments in education (ILSAE), such as IEA's Trends in International Mathematics and Science Study (TIMSS) or OECD's Programme for International Student Assessment (PISA), can provide input on current policy-relevant research questions with respect to inequality and inequity (e.g., Hopfenbeck et al., 2018). ILSAE tend to consist of both an achievement test component and a questionnaire component. The collected data allows for educational research that assesses potential differences in achievement and/or attitudes between, for instance, students of differing gender, socio-economic status, or migration background (e.g., Hopfenbeck et al., 2018), often in combination with a search for protective or risk factors with respect to such differences by comparing classroom practices and other contextual factors. In this way, ILSAE can help shape educational policy by clarifying standards and providing a wide basis of reference

comparisons for education systems, informing curriculum reforms, identifying investment targets based on poor performance in certain subject domains or by specific groups, and guiding resource allocation for optimization of classroom practices and teacher training (for a review, see e.g., Hernández-Torrano & Courtney, 2021).

Although a potential treasure trove, ILSAE have some inherent limitations such as providing less fine-grained learning achievement outcomes than the regular system of school exams (Clarke & Luna-Bazaldua, 2021) and relying on self-report measures for many relevant contextual factors or background variables (e.g., Hopfenbeck & Maul, 2011; Rutkowski & Rutkowski, 2010), and all this in a low-stakes assessment context (e.g., Eklöf, 2010). There is no immediate feedback nor negative or positive consequences for the students participating in the ILSAE. Hence, data quality and validity issues are of concern for everyone involved in these huge projects. A general concern is that not all students are providing genuine responses and that this might distort results to the extent that it could lead to misguided conclusions and educational policy recommendations. Random responding by students on questionnaire scales of the survey is one type of invalid response behavior that comes across as especially threatening or harmful. Random responding is described as providing "responses without meaningful reference to the test questions" (Berry et al., 1992, p.340) often ascribed to among others insufficient effort, carelessness, thoughtlessness, disengagement, or lack of seriousness and motivation on the part of the person responding to the survey (e.g., Huang et al., 2012). Hence, it is rather intuitive to understand the validity concerns (e.g., Cronbach, 1950; Messick, 1984) that having *random responders* on your survey would raise.

Although observable responses are still provided by the person, a random responder can be seen as causing a form of *nonresponse error*, because we end up lacking accurate data on the genuine attitude or information the person is surveyed about. Hence, as with nonresponse rates (e.g., Bethlehem, 2009; Cochran, 1951), low prevalence of random responders in the sample can be regarded as a quality indicator of both survey and corresponding survey data, whereas a high prevalence makes the quality of survey results open for critical debate. Similar to more traditional nonresponse (e.g., Groves &

Peytcheva, 2008; Hedlin, 2020), the biasing impact will not only depend on the prevalence but also on the underlying mechanism as commonly framed in terms of Rubin's (1976) framework of missing completely at random (MCAR), at random (MAR), or not at random (MNAR). Hence, it might be useful to think in similar terms about random responders when considering their potential impact. If minority groups or groups with other specific characteristics have a higher prevalence of random responders, such systematic disproportionate differences can jeopardize the representativeness of the sample, and if the propensity of engaging in random responding relates to the survey outcomes of interest, this can potentially skew, bias, and invalidate any inferences/conclusions based on the questionnaire scales (for a similar point on nonresponse, see e.g., Richiardi et al., 2013).

This Study

In this study, we performed an initial exploration of this validity issue for survey scales inquiring about students' attitudes towards mathematics and science in the TIMSS 2015 assessment (Martin et al., 2016). We conducted a study across 22 participating educational systems, comparing whether student groups —defined in terms of research- and policy-relevant covariate information on grade (age), gender, socio-economic status, spoken language at home, and migration background —differed in their odds of having been classified as a random responder on six TIMSS student questionnaire scales about students' attitudes and beliefs towards Mathematics and Science. Findings will inform about the potential differential prevalence of random responders among the student groups.

Identifying Random Responders. Detection methods for random responding are either based on auxiliary information at the item level such as item response times or are based on the actual item response pattern across a questionnaire scale. The response-time approach leads to an operationalization in terms of so-called 'rapid guessing', where an item response is given in too little time for the person to have actively processed the actual survey question (Wise, 2017). Although very fine-grained, this approach requires the availability and precise measurement of response time at the item level, as well as the setting of a reasonable threshold for when a response is considered 'too fast'. For sur-

veys where items on questionnaire scales are not presented one at a time, such auxiliary item-level information is not obvious to obtain (in contrast to achievement tests where it is more typical to show one problem at a time). The item response pattern approach requires methods to quantify unexpected variability across responses compared to a typical consistent pattern of responses across the questionnaire scale (e.g., Curran, 2016). This makes the approach less suitable for questionnaire scales that are not targeting a reflective construct (as compared to a more formative construct such as socio-economic status) and not feasible for single items (due to a lack of related items as a comparison base).

In absence of useful auxiliary information at the item level, we conducted scale level detection of random responding following a mixture item response theory (IRT) approach. More specifically, we used an extension of the HYBRID model by Yamamoto (1989) to the polytomous case for survey responses as proposed by van Laar and Braeken (2022). Hence, every student was classified as a random responder or a typical responder on the questionnaire scales under investigation.

Survey Scales. Among the survey scales present in TIMSS 2015, we focused on those related to students' attitudes and beliefs towards mathematics and science. This is an active and relevant area of research in education where there is a general worry about the decline in positive attitudes and beliefs with increasing age and grade or educational level (Potvin & Hasni, 2014). How these attitudes and beliefs relate to educational achievement varies on what exactly is surveyed. Students' confidence in mathematics or science tends to be positively related to achievement in the corresponding subject (Wigfield & Eccles, 2002), whereas achievement's relation with valuing the subject is typically weaker (Lee & Stankov, 2018). Educational stakeholders and governments are invested in these topics as a common educational policy objective aims to encourage students to choose more STEM-related subjects (Science-Technology-Engineering-Mathematics) in higher education to fill job market shortages in those areas and support technological innovation.

TIMSS 2015 surveyed both grade 4 and grade 8 students on their views on engaging teaching, their confidence, and how they like learning in each of the two subject domains

(Mathematics and Science) separately, and this in a multitude of educational systems across the world. The three type of scales were almost exactly the same across the subject domains and grades in both format and wording, and a thorough translation process was applied to support the international administration of the survey. Thus, this set of survey scales (3 types $\times$ 2 domains $\times$ 2 grades $=$ 12 scales) in TIMSS 2015 offered a good variety that helps to set the context for the potential generalization of the study's findings.

Covariates for the Differential Prevalence Study. When considering potential group differences in the prevalence of random responders, we followed the implicit hypothesis that if a participant needs to mentally push him/herself to read and respond to the items on a survey scale, the participant will be more inclined to answer randomly as a low-effort efficient reaction or due to misunderstanding of the survey question and/or response options. This implicit hypothesis and the relevance to educational policy were the two criteria that informed our choice of covariates to study. A third, more methodological criterion that came into play is that one wants to avoid having to rely on unreliable self-report group covariate information to define the groups of relevance. The group indicators that are based on self-report were restricted in this study to simple questions, early in the survey, that directly relate to a participant's identity and are expected to be more reliable and elicit higher veracity.

The TIMSS survey was administered to children in grade 4 as well as young adolescents in grade 8. Both *grade* populations responded to quite similar surveys, but they are not guaranteed to respond in a similar fashion. One can argue that questions about attitudes and beliefs towards mathematics and sciences might require more effort from those in the lower grades as it might be less obvious for them to relate to or understand the questions (e.g., Mellor & Moore, 2013). On the other hand, students in the higher grades are said to be more sceptical and critical towards time and effort investment affecting their response motivation (e.g., Rosenzweig et al., 2019; Silm et al., 2020). Hence, although a grade-differential prevalence of random responders sounds not too unreasonable to expect, it is less clear what direction this would take.

Although also available as a self-report measure, information on the gender of a student was directly available as registered by the TIMSS test administrators. With respect to potential *gender* differences in the prevalence of random responders, a literature review by DeMars et al. (2013) concluded that overall, when considering attendance, response times, and self-reported effort, females would be expected to put more effort into low-stakes tests than males. The review mostly covered achievement tests, but it sounds reasonable to extend a similar expectation to a survey context. Tentative explanations for such differential prevalence bring up gender-stereotyped personality trait differences in terms of conscientiousness and agreeableness (see also Bowling et al., 2016; Löckenhoff et al., 2014).

In education, the link between *socio-economic status* (SES) and educational outcomes (for an achievement-focused review, see e.g., Sirin, 2005) is a robust finding and reason for concern and research on educational inequalities and inequity. As a proxy for a student's SES, we used the self-reported estimate of the number of books at home. Based on a comparison with official register data in Sweden, Wiberg and Rolfsman (2021) recommended the use of this self-report measure, with the added benefit that it is simple and has low omission rates. In the survey non-response literature (e.g., Goyder et al., 2002), it is common to find lower non-response rates with higher SES, and this at all stages of the survey data collection. Reasons for this non-response trend are less clear, but speculated to be linked to socio-psychological factors. Following these findings, we expected to observe a similar difference in the prevalence of random responders between low and high SES groups.

Spoken language at home might be another potential factor related to the differential prevalence of random responders. When the language of the survey is different from the language the student speaks at home, this might require more effort, both cognitively in terms of ease of understanding as well as mentally in terms of engagement/relating to the survey. In the context of achievement tests for young adults, Goldhammer et al. (2017) observed that a difference between test and home language was related to more disengagement as measured by more rapid-guessing. Hence, also for the prevalence of

random responders, we expected a similar difference to apply.

We also considered *migration background*, an issue that is often of prime interest for policymakers. Based on the self-reports on whether their respective parents were born in the country where the survey was administrated, a crude student migration background index was constructed. General expectations on the relation of this covariate to the prevalence of random responding are hard to make as the contextual factors surrounding immigration will heavily differ depending on the educational system.

Furthermore, we will map and report resulting patterns of student group differences in the prevalence of random responders across the different *educational systems* participating in TIMSS2015, but, lacking a well-justified theory on such cross-system differences, no further hypotheses were made.

In sum, the key research question addressed by this study is 'who are the random responders on the students' attitudes and beliefs in mathematics and science survey scales of TIMSS 2015?'. More specifically, we investigated whether being classified as random responder instead of typical responder is associated with student characteristics such as grade, gender, SES, spoken language at home, or migration background.

Method

TIMSS is an international large-scale assessment of mathematics and science, which has been conducted normally every four years since 1995. TIMSS 2015 provides the sixth assessment of trends in the fourth grade and/or eighth grade of fifty-seven educational systems and seven benchmarking participants, including assessments of mathematics and science achievement as well as context questionnaires collecting background information (Mullis & Martin, 2013).

The student questionnaire is part of the context questionnaires and is given to each student who takes part in the assessment, with some questions identical for the fourth-graders and eighth-graders. The student questionnaire for eighth grade has an integrated version and a separated version, depending on the implemented science program in the educational system. The integrated version is for those with science as a single or general subject, while the separated version is for those where science is separated into different

subjects, including biology, earth science, chemistry, and physics.

Sample

We considered the educational systems that participated in both the mathematics and the science assessment of TIMSS 2015, with both grade four and grade eight students, and that were not one of the added benchmarking participants. Furthermore, to retain close comparability of student questionnaires between grades four and eight, we only included educational systems with an integrated science program. This ensured that student questionnaires are consistent in terms of questionnaire length, scale items, and scale position. In total, 22 educational systems[24] meet these inclusion criteria: Australia (AUS), Bahrain (BHR), Canada (CAN), Chile (CHL), Chinese Taipei (TWN), England (ENG), Hong Kong SAR (HKG), Iran, Islamic Rep. of (IRN), Ireland (IRL), Italy (ITA), Japan (JPN), Korea, Rep. of (KOR), Kuwait (KWT), New Zealand (NZL), Norway (NOR), Oman (OMN), Qatar (QAT), Saudi Arabia (SAU), Singapore (SGP), Turkey (TUR), United Arab Emirates (ARE), and United States (USA).

TIMSS's target sample size for the number of students to be reached within an educational system is $n = 4000$ (if student population size and other practicalities permit). For the set of educational systems in this study, sample sizes ranged from 3593 grade 4 students in Kuwait to 21177 in the United Arab Emirates, and from 3759 grade 8 students in Saudi Arabia to 18012 in the United Arab Emirates. Table A1 and Table A2 in the Appendix summarize these and other descriptive statistics.

Measures

The measures used in this study were all part of or based on items in the TIMSS 2015 student questionnaire. The student questionnaire covers basic background questions about the students and their home situation, and it includes questions about the students' school experiences, attitudes, and beliefs with respect to school subjects and homework.

[24] Their corresponding (ISO) code will be used as the label in figures and tables.

Survey Scales: Students' Attitudes and Beliefs in Mathematics and Science

The six survey scales measured three types of student attitudes and beliefs on two subject domains (mathematics and science): Like Learning Mathematics (variables: 'ASB01A'-'ASB01I' in grade 4, 'BSBS17A' - 'BSBS17I' in grade 8), View on Engaging Teaching in Mathematics Lessons (variables: 'ASB02A'-'ASB02J' in grade 4, 'BSBS18A' - 'BSBS18J' in grade 8), Confidence in Mathematics (variables: 'ASB03A'-'ASB03I' in grade 4, 'BSBS19A' - 'BSBS19I' in grade 8), Like Learning Science (variables: 'ASB04A'-'ASB04I' in grade 4, 'BSBS21A' - 'BSBS21I' in grade 8), View on Engaging Teaching in Science Lessons (variables: 'ASB05A'-'ASB04J' in grade 4, 'BSBS22A' - 'BSBS22J' in grade 8), and Confidence in Science (variables: 'ASB06A'-'ASB06G' in grade 4, 'BSBS23A' - 'BSBS23H' in grade 8). These were Likert scales of between 7 and 10 items using four categories ranging from 'agree a lot', over 'agree a little'/'disagree a little', to 'disagree a lot'. The Like Learning scales (abbreviated as Like-M and Like-S) contain items related to how the student perceives and enjoys the subject and are also referred to as measuring intrinsic motivation (e.g., Michaelides et al., 2019). The Confidence scales (abbreviated as Conf-M and Conf-S) contain items related to students' self-concept with respect to the subject domain (e.g., Michaelides et al., 2019). The View scales (abbreviated as View-M and View-S) contain items related to how the student perceives their teacher's interaction with both the subject and the students.

Covariates: Five Student Characteristics

Five student characteristics were considered as covariates potentially related to the prevalence of random responders on the questionnaire scales in an educational system in TIMSS 2015 (for descriptive statistics by grade and per educational system, see Table A1 and Table A2).

Grade. The student's grade is a non-student-reported variable based on whether the student was part of the grade four or grade eight administration of the student questionnaire (TIMSS provides separate datasets per grade by country). This grade variable was dummy coded, with grade four coded as zero, and grade eight coded as one. Note that some educational systems, for reasons related to curriculum or the current state

of education, decided to participate with different grades than four and eight (i.e., Norway, England, and New Zealand participated with grade five and grade nine). Regardless, these grades will still be labeled four and eight during the analyses. There were no missing data for this background variable.

Gender. For gender, we used the non-student-reported variable 'ITSEX' from the Student Tracking Form, which is filled out by the test administrators (e.g., Martin et al., 2016). This gender variable was dummy coded, with female coded as zero, and male coded as one. The male-to-female student ratio was about 50/50 in both grades of all participating educational systems (the biggest imbalance was 54% to 46% in Hong Kong). There were no missing data for this background variable.

Self-reported Socio-Economic Status (SES). The students reported an estimated number of books at home on an ordered scale of five categories: "None or very few (0-10 books)", "Enough to fill one shelf (11-25 books)", "Enough to fill one bookcase (26-100 books)", "Enough to fill two bookcases (101-200 books) and "Enough to fill three or more bookcases (more than 200 books)". The five categories of this number of books variable ('ASBG04' and 'BSBG04' in grade 4 and grade 8 student questionnaire, respectively) were recoded to a scale ranging from 0 to 4. The distribution of the number of books variable varied widely across educational systems and grades. For example, in Korea 29% of fourth-graders and 25% of eighth-graders reported having 101-200 books, and 44% and 39% reported having more than 200 books at home, respectively. In contrast, only 11% of fourth-graders reported having more than 100 books in Chile and only 10% of eighth-graders in Kuwait. In most educational systems no more than 5% of the students did not provide a response to this survey question, with the exception of fourth-grade students in Saudi Arabia (9%), and students of both grades in Kuwait (10%).

Self-reported Language at Home. The students reported their frequency of speaking the language of the achievement test and student questionnaire at home on an ordered scale of four categories. This language variable (i.e., 'ASBG03' in the grade 4 student questionnaire and 'BSBG03' in the grade 8 student questionnaire) was dummy coded, collapsing the categories "never" and "sometimes" to be coded as zero, and collapsing the

categories "almost always" and "always" to be coded as one. The proportion of students considering themselves to speak (almost) always the language of the test at home varied largely across the educational systems, from only 19% in Kuwait to 100% in Korea. In most educational systems, eighth graders reported more often than fourth graders to (almost) always speak the language of the test at home, with an average between-grade difference of 9 percentage points. On average, about 5% of the students in an educational system did not respond to this survey question, with the largest proportion of missingness (up to 10%) in Kuwait.

Self-reported Migration Background. Students were asked whether their mother was native-born and whether their father was native-born. Both the father variable (i.e., 'ASBG06A' in grade 4 and 'BSBG09A' in grade 8)' and the mother variable (i.e., 'ASBG06B' in grade 4 and 'BSBG09B' in grade 8) had three response categories: "Yes", "No", and "I don't know". A dummy variable was created based on whether the student reported, on at least one of the two variables, their parent to be foreign-born. A combination of one native-born and either an omitted or "I don't know" response resulted in a missing score on this dummy variable; the same holds for a combination of responses only consisting of an omitted or "I don't know" response. The proportion of students reporting to have at least one foreign-born parent varies widely across educational systems, from as low as 1% in Korea to as high as 66% in the United Arab Emirates. On average about 6% of the students in an educational system missed a score on the migration dummy, with the largest proportion of missingness (up to 22%) for grade 4 students in Taiwan and the United States.

Outcome: Classification as Random Responder

Following a mixture item response theory (IRT) approach (Sen & Cohen, 2019), we classify a student as a random or as a typical responder, for each of the six survey scales considered in this study, using an extension of the HYBRID model by Yamamoto (1989) to the polytomous case for survey responses as proposed by van Laar and Braeken (2022). Classification is based on the maximum posterior class membership probability of a mixture model consisting of two classes. The approach assumes that there are two

distinct, yet unobserved latent groups of responders in the population expressing different response behavior on the survey scale: the class of 'random responders' and the class of 'typical responders' (see Figure 1). In the class reflecting the typical responders, a student is assumed to provide responses across items in a consistent fashion according to their value on the underlying common latent trait (see Figure 1a). In the class reflecting the random responders, a student is assumed to provide unrelated responses across items in a more haphazard fashion (see Figure 1b). More specifically, this comes down to a mixture of (i) a graded response model (Samejima, 1969) for ordered item responses and (ii) a null model with independent item responses that have an equal chance of falling in either of the possible response categories. Note that because the class model for random responders has only fixed known parameters, the mixture model only has one extra parameter to estimate compared to a conventional graded response model, being the mixture class weight which can be seen as the prevalence estimate of random responders in the population.

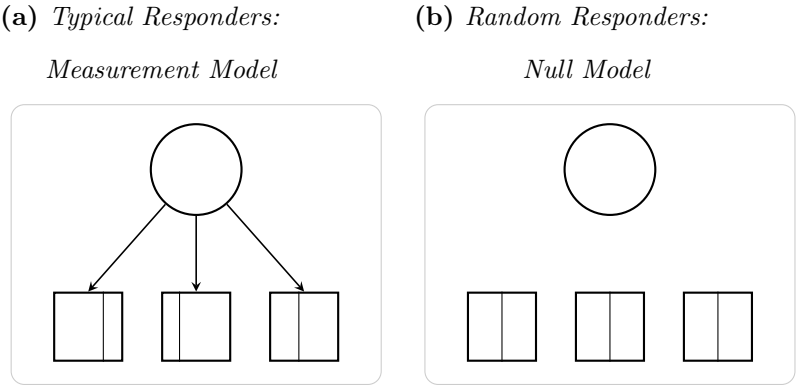
Estimation. The mixture IRT model was estimated separately for each scale per educational system in each grade. Models were estimated in Mplus Version 8.2 (Muthén & Muthén, 1998–2017) through the MplusAutomation package for R version 0.7-3 (Hallquist & Wiley, 2018). We accounted for the total student weights in the TIMSS sampling design and used full-information maximum likelihood estimation with robust standard errors and the expectation-maximization acceleration algorithm with a standard of 400 random starts, 100 final stage optimizations, and 10 initial stage iterations. For each model, the resulting classification variable was a dummy variable with a typical responder being coded zero and a random responder coded as one. These dummy variables were the main outcome variable for further analyses in the current study.

Quality Check. If the mixture model for a specific country-scale combination failed either of two quality checks, the corresponding outcome variable was set to missing. First, the measurement model for the typical responders in the mixture was inspected to ensure that it reflected a clean unidimensional model (i.e., compatible with the assumed common trait for the survey scale). This criterion was not met when two or more standardized item

discrimination parameters (i.e., factor loadings) were below .40. Secondly, a classification entropy of at least .70 was required to ensure that the mixture model was able to provide a good enough distinction between the two latent groups of responders.

Figure 1

Mixture IRT model Framework to Define and Operationalize Random Responders in terms of Independence and Uniformity of Item Responses.



Note. Symbols follow standard path diagram conventions, with squares representing observed variables (i.e., item responses); circles, latent variables (i.e., trait to be measured by the scale of items); arrows indicating dependence relations; vertical lines, response category thresholds. Typical responders: conditional independence given the latent trait; Random responders: mutual independence with uniformly distributed response categories (cf. squares divided into equal parts and no relation with circle or other squares). Reprinted under the terms of CC-BY-NC from "Random responders in the TIMSS 2015 student questionnaire: A threat to validity?" by S. van Laar and J. Braeken, 2022, *Journal of Educational Measurement*.

Statistical Analysis

Odds ratios (OR) were computed as an effect size measure comparing whether the odds of having been classified as a random responder on a specific survey scale are different between the student groups identified by the respective covariate. Odds ratios of 1.22, 1.88, and 3.00 were interpreted as small, medium, and large effect sizes, respectively (Olivier & Bell, 2013) (for negative dependence, the corresponding inversed values are .82, .53, and .33). Computations were student-weighted in accordance with the TIMSS sampling design and run via the R-package 'survey' (Lumley, 2010). For the grade covariate, the data was combined across grades, per educational system by scale combination,

to allow for a comparison between grade 4 and grade 8 students. For the grade covariate the odds ratio was computed for each scale based on the across-grades pooled dataset per educational system; For the four other covariates, the odds ratio was computed within each grade, per educational system by scale combination. Hence, a total of 1188 (i.e., $(1 + 4 \times 2) \times 22 \times 6$) odds ratio estimates were obtained.

To summarize the abundance of results, we made use of meta-analytic tools (e.g., Borenstein et al., 2021) via the R-package 'metafor' (Viechtbauer, 2010). Confidence intervals of the average log odds ratio (i.e., log(OR)) were computed under the random effects meta-analytic model with educational systems taking the role of the independent 'studies'. These confidence intervals were supplemented by corresponding prediction intervals for a randomly selected individual system estimate; the width of the prediction intervals relative to the confidence interval reflects the amount of heterogeneity in effect size among the educational systems. The further away the prediction interval stretches from the confidence interval, the more different the effect sizes across systems are. We briefly summarized noticeably system-specific patterns in the text and included forest plots in the Appendix that display the individual estimates per covariate for each educational system, per grade by survey scale combination. All analysis scripts were run under R version 4.0.0 (R Core Team, 2020).

Results

Prevalence of Random Responders

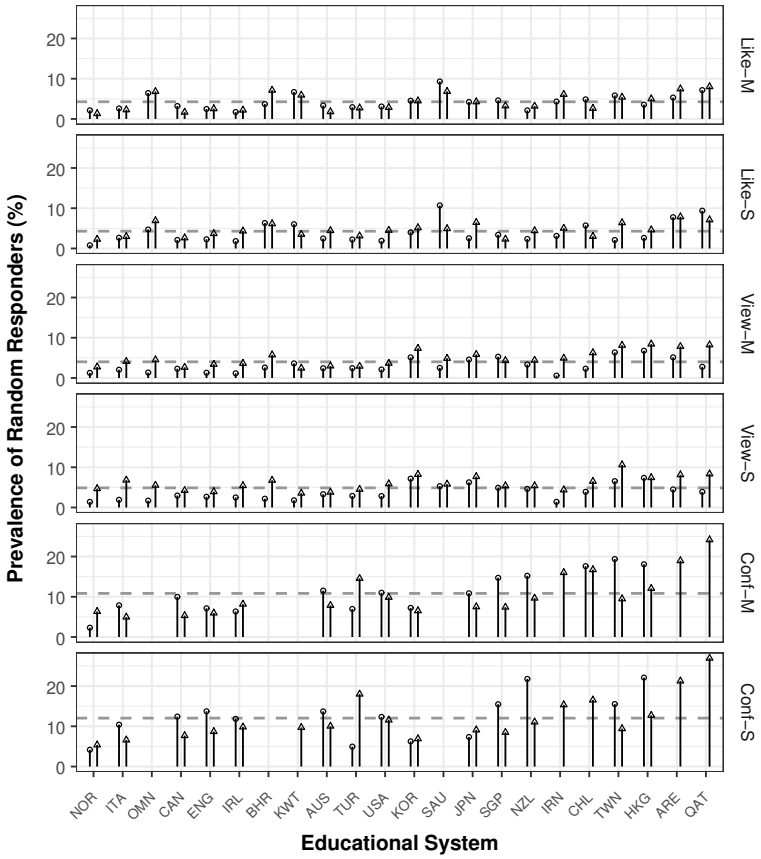
As mentioned before, we had two quality checks to determine whether the resulting classification following the mixture IRT approach to detect random responders could be relied on for further analyses. For the Like and the View scales in both grades and both Mathematics and Science, the random responder classification passed the quality checks for all educational systems without exception. This was not uniformly the case for the Confidence scales. In grade 4, the classification for seven and eight educational systems (out of 22) did not pass the quality checks for Mathematics (i.e., ARE, BHR, IRN, KWT, OMN, QAT, SAU) and Science (i.e., ARE, BHR, CHL, IRN, KWT, OMN, QAT, SAU),

respectively. In grade 8, this was the case for four and three educational systems (out of 22) in Mathematics (i.e., BHR, KWT, OMN, SAU) and Science (i.e., BHR, OMN, SAU), respectively. Notice that it was mostly the same subset of educational systems that did not pass the quality checks for the Confidence scale; mainly due to the questionnaire scale not adhering in those systems to the anticipated unidimensionality of the construct. For the corresponding educational systems not passing the quality checks, no further analyses linking the random responder classification to covariates will be performed, such that they will further appear as missing in the summary graphics and statistics reported.

For countries that passed the quality checks, the average prevalence of having been classified as a random responder on the Like and View scales was around 4%, ranging from 1% to 11% across educational systems and grades (see Figure 2), while the average prevalence on the Confidence scales was somewhat higher at about 11%, ranging from 2% to 27%). The overall average prevalence (across scales, grades, and educational systems) was around 7%.

Figure 2

Estimated Prevalence of Having Been Classified as a Random Responder on the six Questionnaire Scales across Educational Systems.



Note. Circles and triangles represent grade four and grade eight, respectively. Educational systems were ordered by across-grade-and-scale average prevalence, with the gray dashed line being the across systems and across grades average for the scale.

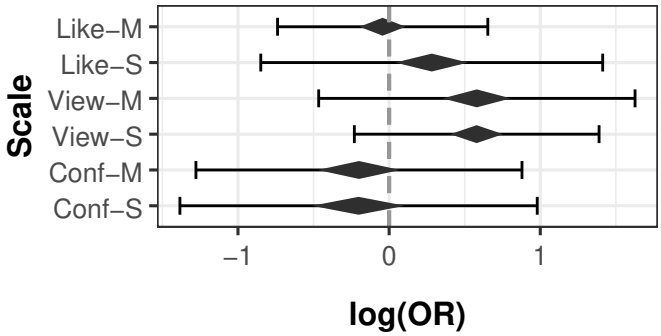
Random Responder = f(Grade)

The relation between having been classified as a random responder and grade differed across the six scales. On average across the 22 educational systems, grade eight students had significantly higher odds of having been classified as a random responder than grade four students on both View scales (OR = 1.79, small to medium effect size) and the Like Science scale (OR = 1.33, small effect size), whereas no such support was found on both Confidence scales and the Like Mathematics scale (see Figure 3, the confidence intervals (black diamonds) of View-M, View-S, and Like-S exceed zero; the confidence intervals of Conf-M, Conf-S and Like-M include zero). The width of the prediction intervals in

Figure 3 did imply heterogeneity among the educational systems. For instance, Iran showed the most obvious grade difference in the prevalence of random responders (OR = 2.84, medium effect size), especially on the View Mathematics scale (OR = 8.67, large effect size), while Singapore showed an opposite grade difference (i.e., grade 8 < grade 4) in five of the six scales (OR = 0.68, small effect size).

Figure 3

Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Grade.



Note. The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval for a randomly sampled educational system. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix: Figure A1 and Tables A1-A2. A positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher/lower for grade 8 than for grade 4 students. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' attitudes and beliefs in Mathematics and Science.

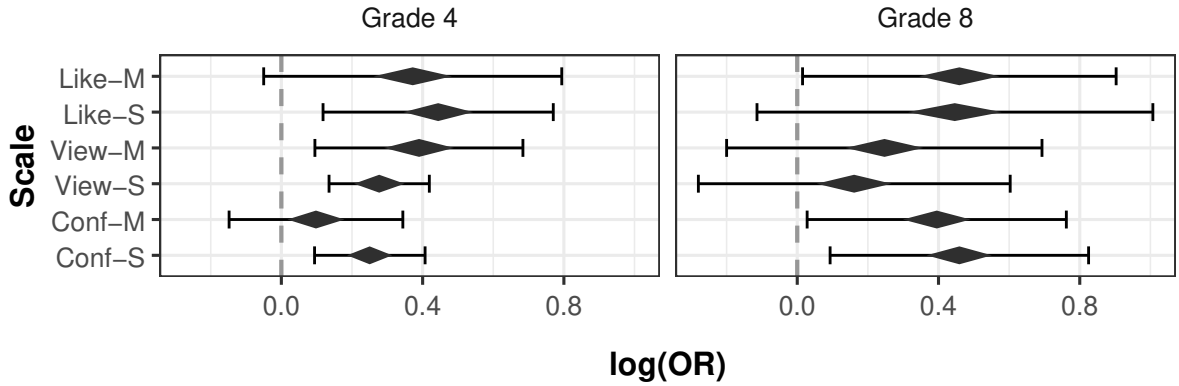
Random Responder = f(Gender)

On average across the 22 educational systems, male students had significantly higher odds of having been classified as a random responder, and this on all six scales and in both grades (see Figure 4, all confidence intervals (black diamonds) exceed zero). The average odds ratio for the six scales ranged from 1.10 to 1.58, with a median of 1.46 (i.e., log(OR) = .38), corresponding to a significant but small effect size, and hence gender difference in the prevalence of random responders. Although the gender difference applied quite generally, the width of the prediction intervals in Figure 4 implied heterogeneity among

the educational systems. For instance, Chile and the USA were the educational systems where the gender difference was almost absent (i.e., log(OR)$\approx$ 0), whereas Saudi-Arabia and Oman were two educational systems with a more pronounced gender difference in the prevalence of random responders (i.e., medium OR effect sizes). For Norway, there was no support for a gender difference for either View scale, but it had the highest observed gender difference among systems on the Like Mathematics scale (average OR = 2.51 across grades).

Figure 4

Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Gender.



Note. The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval for a randomly sampled educational system. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix: Figure A2 and Tables A1-A2. A positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher/lower for male than for female students. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' attitudes and beliefs in Mathematics and Science.

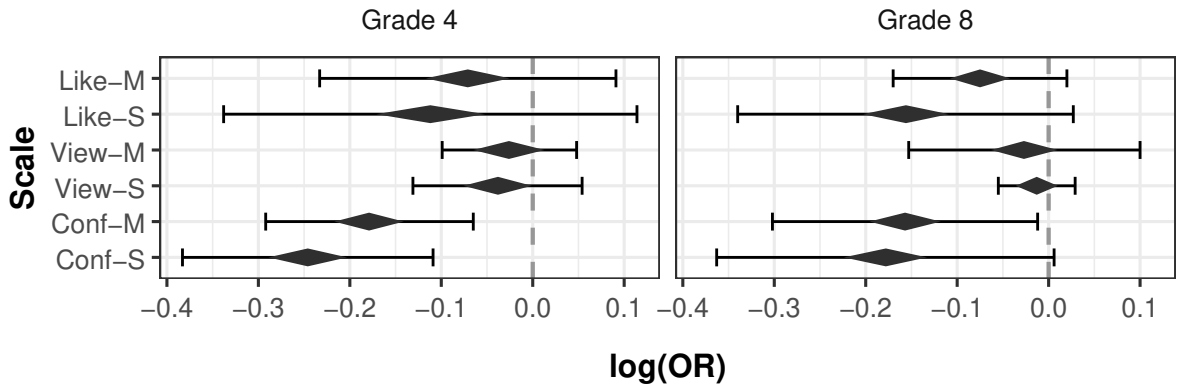
Random Responder = f(Number of Books at home [SES])

On average across the 22 educational systems, students with a higher self-reported number of books at home had significantly lower odds of having been classified as a random responder on the Like scales (average odds ratio across grades: OR = .93 for Mathematics; OR = .87 for Science) and the Confidence scales (average odds ratio across grades: OR = .85 for Mathematics; OR = .81 for Science), but no support for such

relation was found on the View scales (see Figure 5). Note that the number of books covariate had 5 ordered categories, and the interpretation here was for only one category difference, hence the difference between students with the most (more than 200 books) and the fewest (0-10 books) self-reported number of books at home was expected to be four units. For instance, the median of the across-systems average odds ratios for the six scales was .91 (i.e., log(OR) = -.09) in one unit difference, leading to a small effect size of OR = .70 when comparing the two scale-extremes (i.e., $\exp(-.09 \times 4) = \exp(-.09)^4$). The prediction intervals indicated that most educational systems showed that students reporting to have more books at home had significantly lower odds of having been classified as a random responder on the Confidence scales. Yet, the width of the prediction intervals in Figure 5, for these and the other four scales, did imply heterogeneity among the educational systems. For instance, Chile and Saudi Arabia were the educational systems where the number of books difference was almost absent (i.e., average log(OR)$\approx$ 0), while England and New Zealand had the largest OR effect sizes among systems (average OR = .82 and .83, respectively). For Ireland, there was no support for a number-of-books difference for the Like Mathematics and View Science scales, but it had the highest observed number of books difference among systems on the Confidence in Science scale (average OR = .68). At the individual educational system level, the confidence intervals for fourth grade are generally wider than for eighth grade (see Appendix: Figure A3).

Figure 5

Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Number of Books at Home.



Note. The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval for a randomly sampled educational system. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix: Figure A3 and Tables A1-A2. Number of Books at Home is coded 0=None or very few (0-10 books) / 1=Enough to fill one shelf (11-25 books) / 2=Enough to fill one bookcase (26-100 books) / 3=Enough to fill two bookcases (101-200 books) / 4=Enough to fill three or more bookcases (more than 200 books), hence a positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher for students who reported having more/fewer books at home. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' attitudes and beliefs in Mathematics and Science.

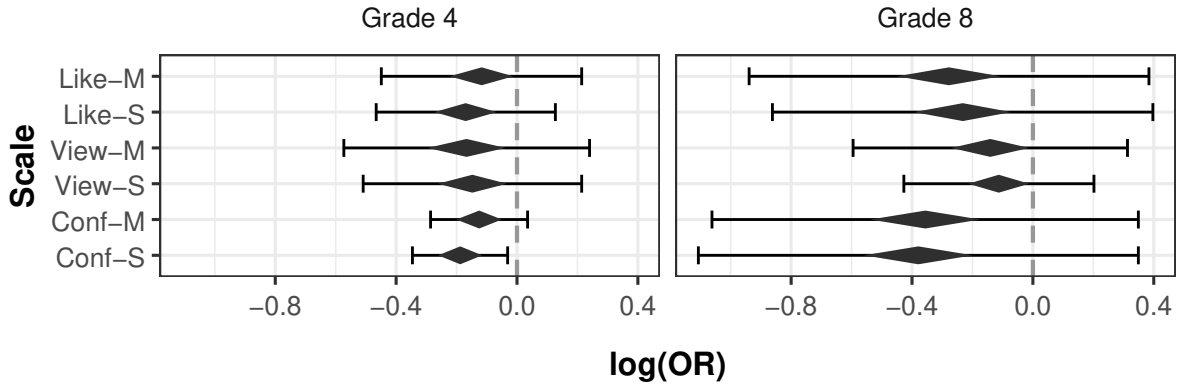
Random Responder = f(Language at Home)

On average across the 22 educational systems, students who more often speak the test language at home had significantly lower odds of having been classified as random responders than those who don't speak the same language at home, and this on all six scales and in both grade four and grade eight (see Figure 6, all confidence intervals are below zero). The average odds ratio for the six scales ranged from .68 to .89, with a median of .84 (i.e., log(OR) = -.17), corresponding to an ignorable to small effect size. The width of the prediction intervals in Figure 6 did imply heterogeneity among the educational systems, with prediction intervals even wider and more negative effect sizes for individual systems in grade eight than in grade four. For instance, Japan showed the most obvious language-related prevalence difference (average OR = .37, medium effect

size). Note that the language covariate had extreme distributions in some educational systems, such as in Japan and Korea where very few students reported speaking any other language at home (1-2% of fourth and eighth graders in Japan and close to 0% of eighth-graders in Korea), contributing to wider confidence intervals in these systems. Qatar's grade eight was the only educational system where speaking the same language had a positive relation to having been classified as a random responder (average OR = 1.46, small effect size).

Figure 6

Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Spoken Language at Home.



Note. The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval for a randomly sampled educational system. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix: Figure A4 and Tables A1-A2. Language at Home is coded 1 = Always or almost always speak <language of test> at home / 0 = Sometimes or never speak <language of test> at home, hence a positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher for students more/less frequently speaking <language of test> at home. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' attitudes and beliefs in Mathematics and Science.

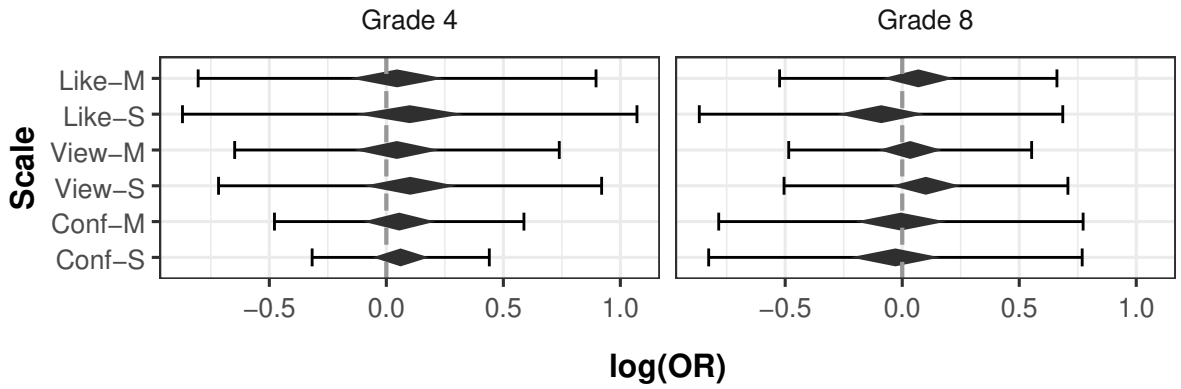
Random Responder = f(Migration Background)

On average across the 22 educational systems, no significant relation was found between having at least one foreign-born parent (versus both native-born parents) and having been classified as random responders on all six scales for both grades. How-

ever, the width of the prediction intervals in Figure 7 did imply heterogeneity among the educational systems, with different directions of effect sizes for individual systems (see Appendix: Figure A5). For instance, the United Arab Emirates and Qatar showed the strongest negative migration background prevalence differences (i.e., students with at least one foreign-born parent had significantly lower odds of having been classified as random responders than those with both native-born parents, average OR = 0.55 and 0.56, respectively, medium effect sizes), whereas Turkey and Iran showed the strongest positive migration background prevalence differences (i.e., students with at least one foreign-born parent had significantly higher odds of having been classified as random responders than those with both native-born parents, average OR = 2.49 and 1.90, respectively, medium effect sizes). Note that some educational systems such as Japan and Korea had few students with migration backgrounds (i.e., under 5%), contributing to wider confidence intervals in these systems.

Figure 7

Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Migration Background.



Note. The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval for a randomly sampled educational system. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix: Figure A5 and Tables A1-A2. Migration background is coded 1=At least one foreign-born parent / 0=Both native-born parents, hence a positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher/lower for students with than without migration background. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' attitudes and beliefs in Mathematics and Science.

Discussion

Although observable responses are still provided, a random responder can be seen as causing a form of *nonresponse error*, in that we end up lacking accurate data on the genuine attitude or information the student is surveyed about. Similar to more traditional nonresponse, a low prevalence of random responders can be seen as a quality indicator for both the survey and response data resulting from the survey. We found an overall prevalence of random responders ranging from 1% to 27%, with an average of 7% across educational systems for the six TIMSS 2015 scales measuring students' attitudes and beliefs in mathematics and science. Hence, supporting the quality of international large-scale assessments in comparative educational research, this prevalence is relatively low. Yet this 7% average does represent some of those students that typically make up for the stereotypical anecdotes that are underlying general concerns about whether

students provide genuine valid responses to the questionnaire in these typical low-stakes assessments. The range of prevalence estimates is comparable to numbers found in the literature for self-report inventories in other fields (e.g., Credé, 2010; Steedle et al., 2019).

Differential Prevalence of Random Responders. Similar to nonresponse (e.g., Richiardi et al., 2013), the impact of the prevalence of random responders crucially depends on who they are, these random responders. If minority groups or groups with other specific characteristics have a higher prevalence of random responders, such systematic disproportionate differences can jeopardize the representativity of the sample and if the propensity of engaging in random response behavior relates to the survey outcomes of interest this can potentially skew, or at worst invalidate, inferences/conclusions based on the questionnaire scales. The key research objective in this study was to investigate whether random responders were disproportionately present in groups defined by research- and policy-relevant covariates. We used a mixture IRT approach to classify students as random responders and meta-analysis summaries to present our results for each of six questionnaire scales across 22 educational systems and two grades.

We found a small to medium grade difference in prevalence for the View scales (and the Like-S scale), with grade eight students having higher odds of having been classified as a random responder than grade four students on average across educational systems. This was counter to our implicit hypothesis that assumed the questionnaire to be less taxing for the students in the later grade, but it could very well be consistent with a higher intrinsic motivation of younger kids versus young adolescents, similar to the observed decline for achievement tests in students' expectancies and task values (e.g., Rosenzweig et al., 2019). We found a small gender difference in prevalence, with male students having higher odds of having been classified as a random responder than female students. This seems in line with the stereotype expectation that girls are more diligent and that boys would put in less effort in low-stakes situations (e.g., DeMars et al., 2013). Context-wise, a small SES difference in prevalence was found for all scales except the View scales, with students reporting having fewer books at home also having higher odds of having been classified as a random responder than students reporting having more books at home. This SES

difference is in line with findings in the more general nonresponse literature (Goyder et al., 2002). A small to ignorable language difference in prevalence was also found, with students speaking a language at home different from the test language having higher odds of having been classified as a random responder than students with matching language, with the trend being more pronounced in grade four than in grade eight. This is in line with a priori expectations following the ease of understanding and mental engagement, and consistent with findings in the rapid guessing literature (e.g., Goldhammer et al., 2017). For immigration background, no empirical support for a difference in prevalence was found using the crude self-reported parents' birthplace indicator.

Generalizability. The findings of this initial study indicate that who are random responders is not entirely random. The obvious caveat remains that there still might be other crucial covariates than those considered here on which the two groups might systematically differ. As noted in the introduction, some of that covariate information might not always be as easy to measure reliably and validly. One should especially be aware of the catch-22 risk of using self-report measures to characterize responders that might not genuinely report back on those indicators. Furthermore, some of the available covariate indicators might be suboptimal: the number of books for SES or parents' birthplace for migration background might not necessarily be the optimal indicators in all cultures or not all younger kids might in fact be able to reliably provide such information. Thus it would be good not to generalize the null findings for the latter covariates beyond the specific operationalization used in this study.

With respect to the scales, only the View scales showed a grade difference in prevalence and almost no SES difference in prevalence, whereas the Confidence scales had a higher overall prevalence of random responders (i.e., on average 11%). Note that the Confidence scales also tended to fail quality checks for mostly the Middle East countries, indicating larger measurement issues there for the majority of students. Altogether these findings do indicate that whatever the mechanisms are underlying random responding, these won't be all generic or uniformly applicable across scales. This implies a crucial role for scale contents and for how students (i.e., the target population) engage with or understand the

questionnaire scale contents.

The observed heterogeneity across educational systems implies that context does matter. Whereas on average a difference in prevalence between two covariate groups might be absent, it might still apply to an individual educational system. For example, a high SES difference in prevalence was observed in England and New Zealand, and grade eight students in Qatar that spoke the same language at home as the test surprisingly had higher odds of having been classified as a random responder than those who did not. The latter finding is likely due to the somewhat atypical immigrant population in Qatar compared to other systems in our study. Similarly, when considering language and migration background, the native culture was so dominant in Japan and Korea that the minority groups were very small, leading to somewhat larger but also more uncertain prevalence differences than elsewhere.

Handling Random Responders. Having been classified as a random responder does not necessarily mean that one consciously and purposefully provides random responses. The classification has only a direct binding to the observed response pattern and not to the underlying intentions or response process. Random response patterns can equally arise due to incidental inattention or lack of understanding of the question or uncertainty about the applicability of response options, and so on. In this sense, it is perhaps more natural to qualify the responses given as nonresponse instead of as definite invalid. Hence, we recommend similar approaches as used in the handling of missing data, to deal with data from random responders (e.g., Meng, 2012). This would imply sensitivity analyses comparing inferences with and without the inclusion of the detected random responders and techniques such as multiple imputation on a rich feature set of relevant covariates and survey design variables to comply with a missing-at-random working assumption. Note that the latter does not mean completely at random (for which we have indications it is not), but conditional on the relevant covariate group differences as suggested in explorative studies like the current study.

Conclusion

Similar to missingness rates, prevalence rates of random responders don't tell the whole story, as their influence will depend on the underlying mechanism: the other variables involved and who in effect provides the nonresponses. This study has shown the prevalence of random responders on questionnaire scales in international comparative educational research to be a function of common policy-relevant covariates. Therefore, we call for two actions: (i) For individual researchers using data from the questionnaires of the international large-scale assessments in education, a default practice of sensitivity analyses and robustness checks; (ii) For the larger testing organizations (e.g., OECD or IEA), the default inclusion of a wide arsenal of survey quality indicators including not only prevalence but also relations to covariates, and this for a larger set of non-response behavior including random responders.

References

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340.

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective*. Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis, 2nd edition*. John Wiley & Sons Inc.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218–229.

Clarke, M., & Luna-Bazaldua, D. (2021). *Primer on large-scale assessments of educational achievement*. World Bank.

Cochran, W. G. (1951). General principles in the selection of a sample. *American Journal of Public Health*, *6*(41), 647–653.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

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.

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

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering per-

son and item characteristics. *Large-scale Assessments in Education*, *5*(1), Article 18.

Goyder, J. C., Warriner, K., & Miller, S. (2002). Evaluating socio-economic status (ses) bias in survey nonresponse. *Journal of Official Statistics*, *18*, 1–12.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *The Public Opinion Quarterly*, *72*(2), 167–189.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Hedlin, D. (2020). Is there a 'safe area' where the nonresponse rate has only a modest effect on bias despite non-ignorable nonresponse? *International Statistical Review*, *88*(3), 642–657.

Hernández-Torrano, D., & Courtney, M. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *PloS ONE*, *9*, Article 17.

Hopfenbeck, T. N., Lenkeit, J., Masri, Y. E., 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.

Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing*, *11*(2), 95–121.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, *65*, 50–64.

Löckenhoff, C. E., Chan, W., McCrae, R. R., Fruyt, F. D., Jussim, L., Bolle, M. D., Paul T. Costa, J., Sutin, A. R., Realo, A., Allik, J., Nakazato, K., Shimonaka, Y.,

Hřebíčková, M., Graf, S., Yik, M., Ficková, E., Brunner-Sciarra, M., de Figueora, N. L., Schmidt, V., ... Terracciano, A. (2014). Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology*, *45*(5), 675–694.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley; Sons.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College.

Mellor, D., & Moore, K. A. (2013). The Use of Likert Scales With Children. *Journal of Pediatric Psychology*, *39*(3), 369–379.

Meng, X.-L. (2012). You want me to analyze data I don't have? Are you insane? *Shanghai Archives of Psychiatry*, *24*(5), 297–301. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4198883/

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*(3), 215–237.

Michaelides, M. P., Brown, G. T. L., Eklöf, H., & Papanastasiou, E. C. (2019). *Motivational Profiles in TIMSS Mathematics: Exploring Student Clusters Across Countries and Time*. Springer International Publishing.

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Boston College.

Muthén, L. K., & Muthén, B. O. (1998–2017). Mplus User's Guide. Eighth Edition.

Olivier, J., & Bell, M. (2013). Effect sizes for 2×2 contingency tables. *PloS ONE*, *8*(3), e58777.

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at k-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education*, *50*(1), 85–129.

R Core Team. (2020). R: A language and environment for statistical computing.

Richiardi, L., Pizzi, C., & Pearce, N. (2013). Commentary: Representativeness is usually not necessary and often should be avoided. *International Journal of Epidemiology*, *42*(4), 1018–1022.

Rosenzweig, E. Q., Wigfield, A., & Eccles, J. S. (2019). Expectancy-value theory and its relevance for student motivation and learning. *The Cambridge Handbook of Motivation and Learning* (pp. 617–644). Cambridge University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, *42*(3), 411–430.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191.

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.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453.

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice*, *38*(2), 101–111.

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wiberg, M., & Rolfsman, E. (2021). Students' self-reported background SES measures in TIMSS in relation to register SES measures when analysing students' achievements in Sweden. *Scandinavian Journal of Educational Research* Advance online publication. https://doi.org/10.1080/00313831.2021.1983863

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of Achievement Motivation* (pp. 91–120). Academic Press.

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

Yamamoto, K. (1989). Hybrid model of IRT and latent class models. *ETS Research Report Series*, *RR-89-41*.

Appendix A

Table A1

Distribution of the covariates for fourth-grade students.

Educational System	Gender: female	Gender: Male	SES: 0-10 books	SES: 11-25 books	SES: 26-100 books	SES: 101-200 books	SES: >200 books	SES: missing	Language: never/sometimes	Language: (almost) always	Language: missing	Migration: native-born parents	Migration: foreign-born parent	Migration: missing	Sample Size: n
ARE	48	52	19	30	27	10	9	4	45	52	4	25	61	15	21177
AUS	49	51	8	18	35	21	16	1	15	84	2	52	37	11	6057
BHR	50	50	20	29	25	11	11	3	31	66	3	52	36	12	4146
CAN	49	51	10	21	38	17	13	2	24	74	2	45	42	12	12283
CHL	49	51	31	33	22	6	5	3	10	87	3	84	8	8	4756
ENG	51	49	10	22	34	18	14	3	17	81	2	52	34	13	4006
HKG	46	54	14	20	32	18	16	1	29	70	1	36	47	17	3600
IRN	49	51	40	27	18	6	6	2	33	65	2	78	12	10	3823
IRL	47	53	9	20	33	20	16	1	12	83	5	69	26	5	4344
ITA	49	51	17	35	28	10	8	1	16	83	1	75	21	4	4373
JPN	50	50	12	29	37	13	8	0	2	98	0	94	3	3	4383
KOR	48	52	4	4	18	29	44	0	8	92	0	96	2	2	4669
KWT	51	49	28	31	17	7	6	10	65	25	10	50	30	20	3593
NZL	49	51	11	19	33	19	17	2	16	83	1	47	40	13	6322
NOR	49	51	7	22	38	19	12	2	15	84	1	70	27	4	4329
OMN	50	50	28	28	23	9	10	4	36	60	4	71	19	11	9105
QAT	51	49	20	28	27	11	11	3	46	52	1	30	56	14	5194
SAU	49	51	32	27	17	7	8	9	19	74	7	66	16	18	4337
SGP	48	52	10	21	37	18	13	0	51	49	0	47	44	9	6517
TUR	49	51	22	33	28	8	5	4	15	80	5	84	7	8	6456
TWN	49	51	19	25	29	13	13	0	40	59	1	65	13	22	4291
USA	51	49	13	23	33	15	13	3	20	76	3	51	27	22	10029

Table A2

Distribution of the covariates for eighth-grade students.

Educational System	Gender: female	Gender: Male	SES: 0-10 books	SES: 11-25 books	SES: 26-100 books	SES: 101-200 books	SES: >200 books	SES: missing	Language: never/sometimes	Language: (almost) always	Language: missing	Migration: native-born parents	Migration: foreign-born parent	Migration: missing	Sample Size: *n*
ARE	50	50	19	29	28	11	10	2	35	63	1	26	66	8	18012
AUS	51	49	11	18	26	20	20	6	7	92	2	51	38	11	10338
BHR	48	52	25	29	26	10	9	2	26	73	1	58	36	6	4918
CAN	51	49	11	21	30	18	17	3	13	85	3	53	42	5	8757
CHL	48	52	25	38	25	7	5	1	5	94	2	93	3	4	4849
ENG	51	49	17	22	28	16	15	2	5	93	2	65	29	6	4814
HKG	47	53	18	25	31	13	13	0	16	84	0	36	52	11	4155
IRN	48	52	27	33	22	8	10	0	33	67	0	95	3	2	6130
IRL	50	50	15	22	28	19	15	1	10	86	4	68	30	2	4704
ITA	49	51	16	25	25	16	18	1	11	88	1	81	17	2	4481
JPN	51	49	12	21	32	17	18	0	1	99	0	96	2	2	4745
KOR	47	53	7	7	22	25	39	0	0	100	0	98	1	0	5309
KWT	50	50	32	30	19	6	4	9	74	19	8	58	30	13	4503
NZL	51	49	14	18	29	19	17	2	7	91	2	57	36	7	8142
NOR	50	50	10	20	29	20	20	1	6	93	1	77	21	2	4697
OMN	48	52	24	33	25	9	8	2	33	66	1	76	22	2	8883
QAT	50	50	22	29	26	12	10	1	30	69	1	32	63	5	5403
SAU	51	49	37	30	19	6	7	2	27	73	1	79	15	6	3759
SGP	49	51	18	27	31	14	11	0	35	65	0	56	40	3	6116
TUR	48	52	16	35	30	11	8	1	10	90	0	95	3	2	6079
TWN	49	51	20	23	27	13	16	0	9	91	0	85	10	5	5711
USA	50	50	17	21	29	17	15	1	9	89	1	68	25	7	10221

Figure A1

System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Grade.

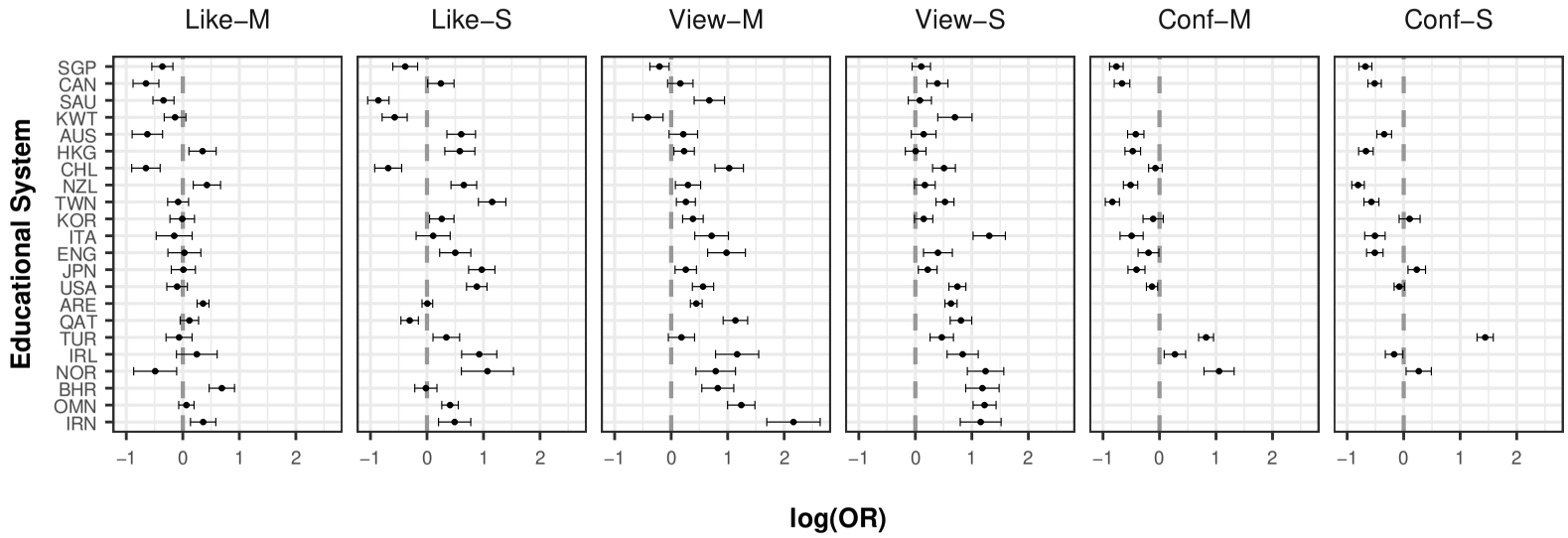


Figure A2

System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Gender.

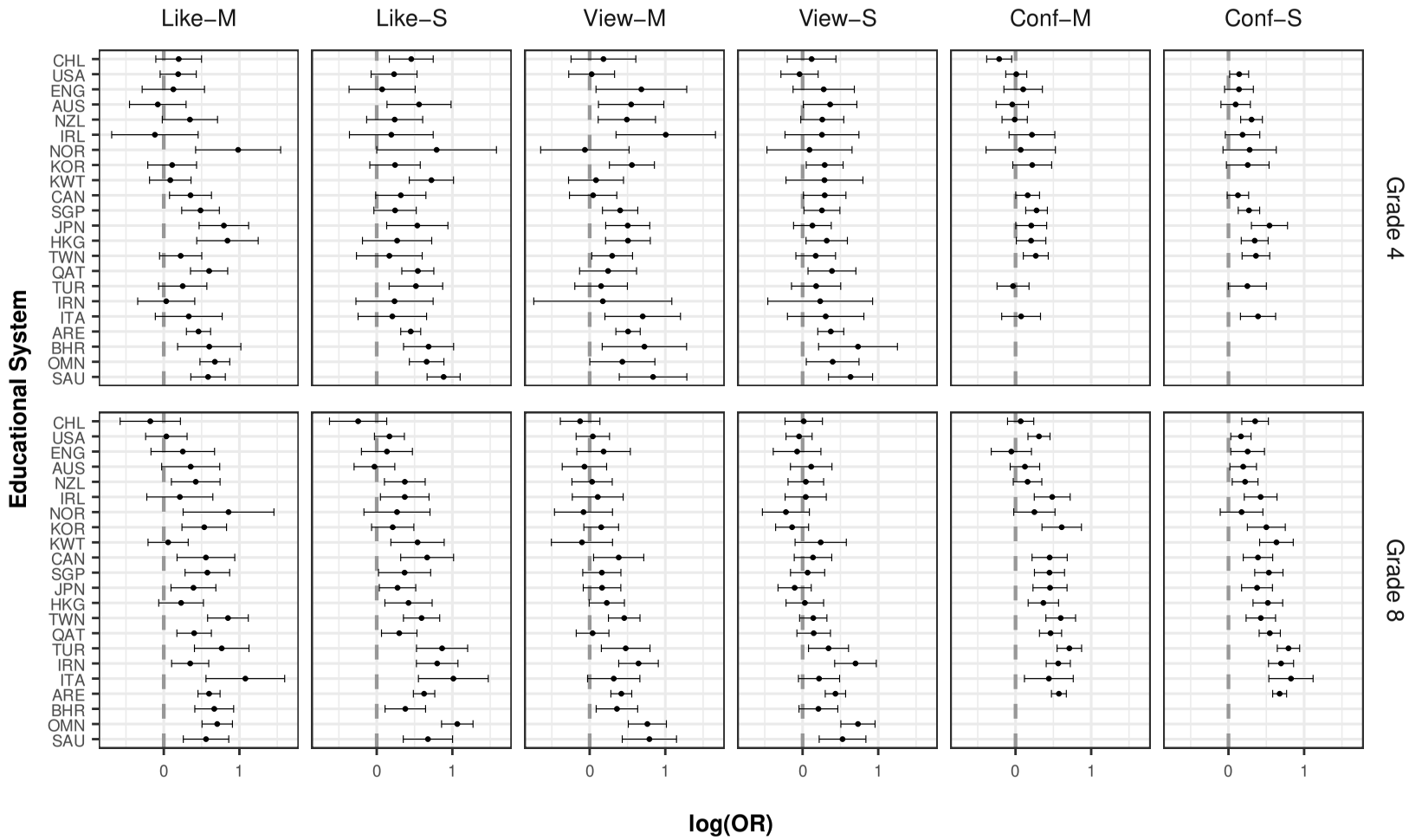


Figure A3

System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Number of Books at Home.

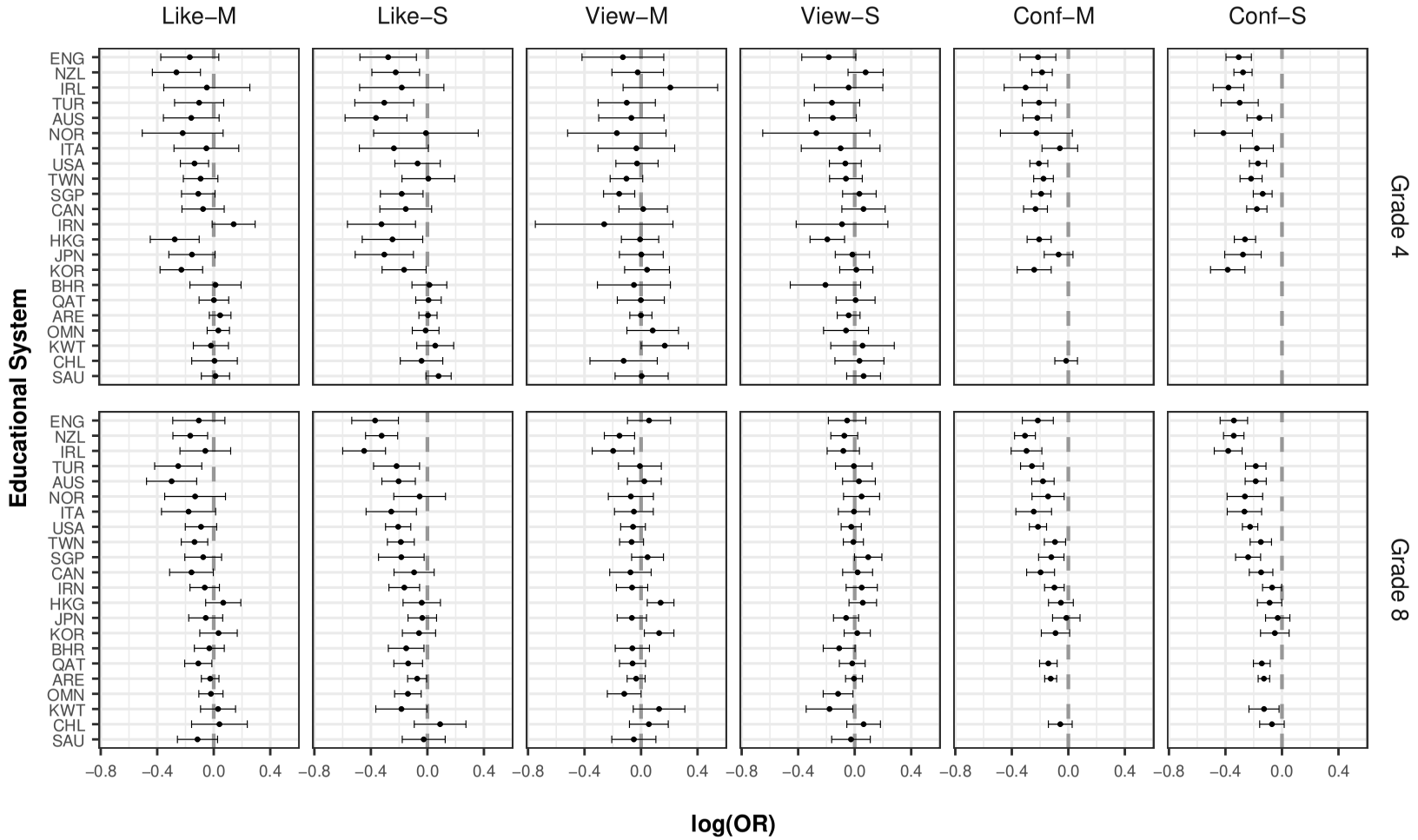


Figure A4

System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Spoken Language at Home.

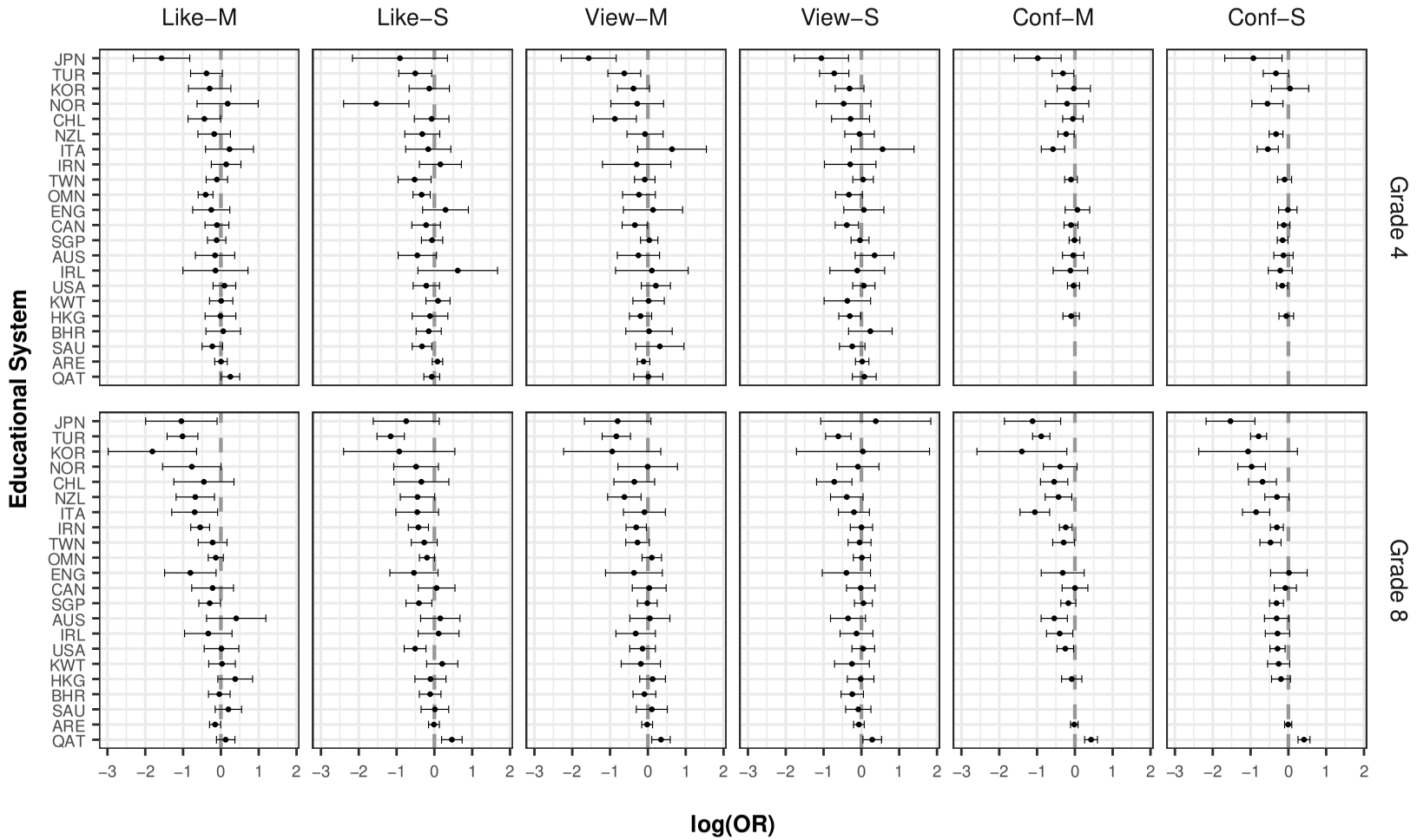
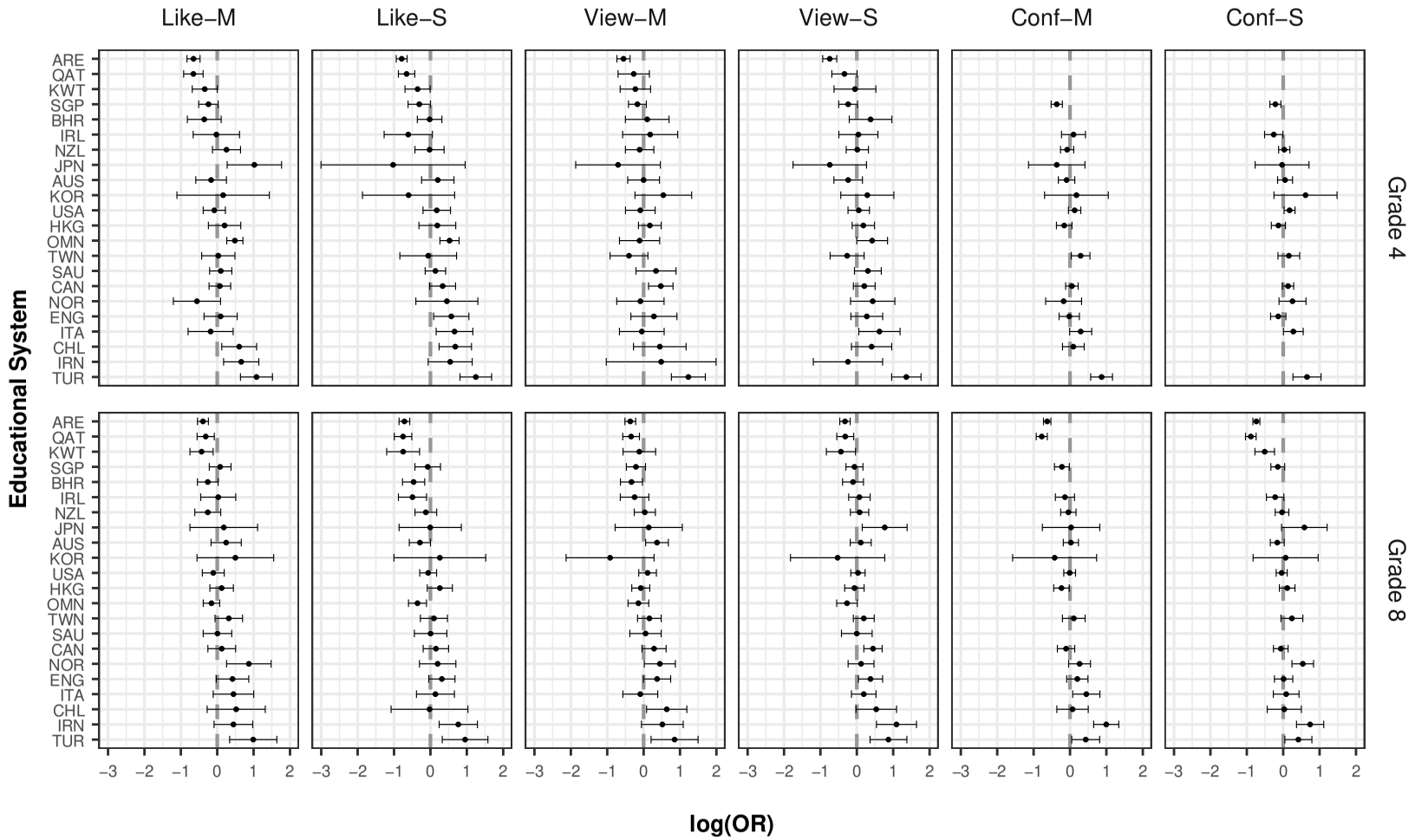


Figure A5

System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Migration Background.



9 Article 6: How often

van Laar, S., Chen, J., & Braeken, J. (2022). *How Randomly are Students Random Responding to your Questionnaire? Within-Person Variability in Random Responding across Scales in the TIMSS 2015 eighth-grade Student Questionnaire.* Manuscript under review.

How Randomly are Students Random Responding to your Questionnaire? Within-Person Variability in Random Responding across Scales in the TIMSS 2015 eighth-grade Student Questionnaire

Questionnaires in educational research assessing students' attitudes and beliefs are low-stakes for the students. As a consequence, students might not always consistently respond to a questionnaire scale, but instead provide more random response patterns with no clear link to items' contents. We study inter-individual differences in students' intra-individual random responding profile across 19 questionnaire scales in the TIMSS 2015 eighth-grade student questionnaire in seven countries. A mixture IRT approach was used to assess students' random responder status on a questionnaire scale. A follow-up latent class analysis across the questionnaire revealed four random responding profiles that generalized across countries: A majority of consistent non-random responders, intermittent moderate random responders, frequent random responders, and students that were exclusively triggered to respond randomly on the confidence scales in the questionnaire. We discuss implications of our findings in light of general data-quality concerns and the potential ineffectiveness of early-warning monitoring systems in computer-based surveys.

A large research base in the educational sciences is built on studies using questionnaires to survey students' values, beliefs, and attitudes towards school subjects such as mathematics and science (Eccles & Wigfield, 2002; Linnenbrink & Pintrich, 2002; Osborne et al., 2003; Potvin & Hasni, 2014). This research base includes both smaller-scale individual research-team studies and larger-scale international comparative studies such as OECD's Program for International Student Assessment (PISA) or IEA's Trends in International Mathematics and Science Study (TIMSS). The research results are typically used to evaluate and contextualize educational practice and inform educational policy.

Yet, research on such attitudinal constructs is low-stakes for the students as it has no direct consequences or relevance for themselves (in contrast to for instance cognitive

achievement tests or exams). At the heart of educational and psychological measurement, even before Cronbach's (1946) treatise on response sets and validity, there is a core concern that students might not always respond accurately or thoughtfully to a questionnaire, but instead shift to responding with the lowest effort (e.g., Curran, 2016; Eklöf, 2010; Huang et al., 2012) such that their item responses and scale scores might no longer accurately reflect the constructs that the questionnaire scales were intended to assess (e.g., Messick, 1984). One of the ways this can express itself is that instead of the expected consistent item response pattern on a questionnaire scale, a more random response pattern is provided with no clear link to items' contents.

Individuals engaging in random response behavior on a questionnaire scale can potentially distort inferences on basic item statistics, reliability, dimensionality, and intercorrelations within and between constructs (e.g., Credé, 2010; Huang et al., 2012; Liu et al., 2019; Maniaci & Rogge, 2014; Meade & Craig, 2012). Random responding can be seen as a type of nonresponse; even though responses are observed, genuine information on the actual response that the individual would have given, if they would have responded in a regular non-random fashion, is missing. Hence, the underlying process giving rise to the nonresponse is crucial both for understanding the phenomenon as well as for assessing its expected impact and how to handle it in data analyses (cf. M(C/N)AR missingness framework, Rubin, 1976). Typically, scale and item means are biased towards their midpoint and residual item variances tend to be inflated, whereas scale and item covariances within and between other constructs can be biased in either direction or remain unaffected. Higher impact can be expected with increased prevalence and when regular consistent responders tend to score further away from the midpoint.

Random responding is speculated to occur due to among others carelessness, insufficient effort, disengagement, or lack of motivation and seriousness on behalf of the respondent (e.g., Huang et al., 2012). To the extent that it are almost always the same individuals that are random responding on scales throughout the questionnaire, the validity threat is mostly located within the student responding. By implication, random responding would then be mostly beyond our reach unless we manage to figure out in-

dividualized incentives that convince these students to genuinely engage with the survey in a low-stakes context. On the other hand, to the extent that individuals systematically vary the extent of their random response behavior throughout the questionnaire, the validity threat could be due to specific triggers in their questionnaire progress or scale content or type. The latter aspects would provide possible pathways to redesign and modify the questionnaire to dampen the triggers and reduce the general validity threat. In contrast, when random responding occurs more incidentally among individuals throughout the questionnaire, there are no clear levers for reducing its prevalence, but the impact of such random responding can also be expected to be minimal as it would conform to a completely-at-random nonresponse pattern (cf. MCAR, Rubin, 1976).

Thus, the distinct patterns in within-person variation in random responding become especially important (Molenaar, 2004, see also) if we want to extend conclusions about individuals based on limited information as in a data quality monitoring or screening system. For example, if a student is identified as a random responder on one scale, what does this imply for the rest of the questionnaire? Can their responses on other scales in the questionnaire still be trusted or are they all to be considered invalidated? By studying the *inter-individual differences in intra-individual random response behavior across the questionnaire scales in a survey*, we aim to further clarify to what extent random responding would be a systematic biasing factor that potentially threatens and distorts inferences made from the survey data and shed further light on potential factors triggering such random response behavior.

Individual differences in random responding across questionnaire scales

There are different ideas of how response behavior is actualized over the course of a questionnaire at the individual level. Most of these ideas can be traced back to ancient-old discussions in the general field of individual differences such as the trait versus state (e.g., Schmitt & Blum, 2020) or person versus situation (e.g., Fleeson, 2004) debates.

Trait perspective. A person-central or trait perspective would prescribe that individuals have the tendency to respond to a questionnaire in a consistent manner (e.g., favoring a certain response option, speeding, or guessing) and that this response behavior

is reflective of underlying personality traits and relatively stable across time (e.g., Messick, 1991). Cronbach (1950) indicates that especially within a singular-content questionnaire this consistency of response behavior should become clear. This does not imply that individuals will respond perfectly consistent, as it has generally been considered that no trait is perfectly stable. Yet, from this viewpoint, it is expected that if respondents would have engaged in random responding on one scale they would also have an increased probability of being identified as a random responder on any of the other scales in the questionnaire, and limited within-person variability in random responding across the questionnaire is implied.

Bowling et al. (2016) provide tentative evidence on temporal stability and correlations with personality traits, which would be consistent with a trait-based individual differences perspective. Other findings in the literature cast doubt on whether such individual consistency in random responding across the questionnaire is a realistic pattern to expect. For a sequence of achievement tests in a USA context, Soland and Kuhfeld (2019) conclude that rapid guessing is not longitudinally stable, but some cross-sectional correlations with other trait measures were observed. For a low-stakes scientific reasoning test, Wise et al. (2009) note that only a small percentage of college students appeared to have engaged in random response behavior on the majority of the questionnaire. Similarly, self-reports in personality research indicated that of the 52% of college students who reported themselves to have engaged in some level of random responding on the MMPI-2 questionnaire, only 3% indicated to have responded randomly to 'many' or 'most' of the items (Berry et al., 1992). Given such findings and the fact that most surveys in large-scale educational research are also non-singular in contents, we expect that only for a limited number of individuals a universally applicable 'random responder' trait applies in an educational survey context.

State perspective. Alternatively, random response behavior could be more of a temporary state that expresses itself only in specific parts of a questionnaire regardless of the content. In the personality assessment literature, the perspective of so-called 'back-random' responding has especially gained much attention (e.g., Clark et al., 2003; Gallen

& Berry, 1997; Pinsoneault, 2007). Once an individual's internal 'cognitive resources' have been depleted and/or they are no longer willing to actively engage with the questionnaire, the individual switches from regular response behavior to expressing random response behavior and carry on to do so for the remainder of the questionnaire (Bowling et al., 2021; Clark et al., 2003). Notions of boredom, disinterest, inattentiveness, or fatigue are indicated as potential underlying drivers of the phenomenon. From this viewpoint, it is expected that once individuals are identified as random responder on a scale, they will also have a higher probability to be identified as random responder on the scales following that scale in the questionnaire.

For a low-stakes information literacy test, Wise (2006) found that several participants switched to random response behavior over the course of the test and persisted to do so for most of the remainder of the test (see also, Cao & Stokes, 2007). Similarly, self-reports in personality research asked about "the proportion of test questions which you were unable to pay attention to and answered randomly" (Berry et al., 1992, p.341) and 42–52% of the individuals across different samples indicated the most common place for them was towards the end of the questionnaire (from response options: mostly in the first part; mostly in the middle part; mostly in the last part; scattered throughout). For achievement testing, Ackerman and Kanfer (2009) concluded that subjective test fatigue was better predicted by individual differences in personal motivation than by mere physical differences in test length. If back-random responding is indeed more of a personal motivation issue, then the low-stakes character of many assessments in educational research can be considered to be a facilitating factor for back-random responding. Whether and the point at which this within-person shift to random response behavior can be observed, will vary across the questionnaire from person to person depending on their general engagement with the questionnaire. Thus, good questionnaire design would target the survey to the intended population, inquire about aspects that speak to this population, and allow generous time to complete the survey; In such an ideal situation, back-random responding should theoretically be a rare phenomenon.

Situation perspective. Whereas the previous perspectives seek the triggers for

random response behavior mostly internal to a person, one could also posit that external triggers could play a role. Cronbach (1950) indicated that response sets are more stable within singular-contents questionnaires and that response sets become more influential as items become more difficult or ambiguous. Similarly, Baer et al. (1997) indicated that the two main reasons for giving random responses were difficulty in understanding the question or in deciding on the response alternative. From this viewpoint it is expected that if respondents would engage in random responding on one scale they would also have an increased probability of being identified a as random responder on a similar scale in the questionnaire, but not a dissimilar scale; where similarity is either in contents or response type. This implies limited within-person variability across similar scales, but large within-person variability between dissimilar groups of scales.

Idiosyncratic perspective. Another alternative is that random response behavior is more unsystematic in nature, an extremely volatile state, meaning that the reasons for random responding on one scale and not on another are rather idiosyncratic to the individual. This would be reflected by individuals switching behavior multiple times and engaging in random responding rather haphazardly throughout the questionnaire. Switching behavior might not be uncommon in practice. Baer et al. (1997) observed that of the 73% of young adolescents who reported themselves to have engaged in some level of random response behavior on a personality inventory, the majority of respondents indicated this behavior to be scattered across the questionnaire. Similarly, Berry et al. (1992) found that 18–32% of the individuals across different samples reported having engaged in random responding in a random fashion. From this viewpoint, the probability of being identified as a random responder on one scale would be independent of someone's response behavior on the other scales and consequently, it would not be possible to make any predictions about the validity of the complete set of responses on all scales in the questionnaire.

The literature is scarce and inconclusive on which patterns of within-person variability will be dominant, or even present or absent. Hence, the core research question is exploratory in nature and can be regarded as a step in charting this unknown territory.

To study this research question we selected the 2015 cycle of the Trends in Mathematics and Science Study (TIMSS 2015: Martin et al., 2016) as it collected responses from large random samples of students, in multiple educational systems across the world, to a large-scale student survey with many questionnaire scales covering students' attitudes and beliefs towards relevant subjects that are popular with educational researchers and policy makers alike.

The many methodological approaches to detect random response behavior that rely on auxiliary resources or require long scales with many items (e.g. Rupp, 2013), are not applicable to most education survey research as both features are typically impractical and as a result absent. In achievement testing, an operationalization utilizing reaction time information on the item level to identify what is labeled a 'rapid guess' (see e.g., Wise et al., 2009) has gained traction. A rapid guess is framed as a response given within such a limited time span that it is clear that the individual did not spend sufficient time to consider and process the question asked, and as a consequence provided an essentially random response. Unfortunately, even when the survey would be computer-based, item-level reaction times are unattainable as items of a questionnaire scale are typically presented all at once on the screen. Given that there is also no single correct response on a survey item, aberrant responses are less obvious unless they form a very systematic pattern (e.g., diagonal responding across the items of a questionnaire scale) and we can for instance also not identify students that perform below chance level. The use of bogus items or instructed response items (Breitsohl & Steidelmüller, 2018; Leiner, 2019) is also not commonplace in educational survey research, and the debate is not yet settled on whether these tools are even an ethical practice or effective. In the end, one needs to resort to the pattern of actual item responses given to the different questionnaire scales in the survey. Given the absence of auxiliary elements for the TIMSS 2015 student questionnaire, we will employ a mixture item response theory (IRT) approach (van Laar & Braeken, 2022) to explicitly model the possibility of two underlying yet unobserved groups in the population, students engaging in regular response behavior versus students engaging in more random response behavior across the items of a questionnaire scale.

This means that random response behavior is operationalized at scale-level directly based on item responses given on that scale by the student. This is perhaps a more coarse operationalization than what would be possible otherwise with auxiliary information, but this is compensated by the presence of up to 19 scales in the TIMSS 2015 student questionnaire allowing for a sufficient range to explore within-person variability.

Method

TIMSS is an international large-scale assessment of mathematics and science, which has been conducted every four years since 1995. TIMSS 2015 provides the sixth assessment of trends in the fourth grade and/or eighth grade of fifty-seven educational systems and seven benchmarking participants. TIMSS 2015 includes assessments of mathematics and science achievement as well as context questionnaires collecting background information (Martin et al., 2016). The data used in this study stems from the student questionnaire for the eighth-grade students. The student questionnaire covers basic background questions about the students and their home situation, but also questionnaire scales about the students' school experiences, attitudes, and beliefs with respect to school subjects and homework. TIMSS's target sample size for the number of students to be reached within an educational system is $n = 4000$ (if student population size and other practicalities permit).

The assessment of the students in TIMSS 2015 was separated into three sections. The students first have the achievement tests, with 45 minutes of testing time per section (i.e., mathematics and science) with a 30-minute break in between. After the achievement tests a second break followed after which the student questionnaire was administered to every student that took part in the TIMSS 2015 achievement test. The testing time for the student questionnaire was set at 30 minutes. The total testing time for an eighth-grade student in the TIMSS 2015 assessment (i.e., all 3 sections) is then 120 minutes in total plus the time for the two breaks. The times were set such that in principle students do not need to rush to complete a section. Students were not allowed to leave the room or start with a new section even if they had already completed the task within the set time frame (Martin et al., 2016). Hence, there is no reward for rushing through the assessment

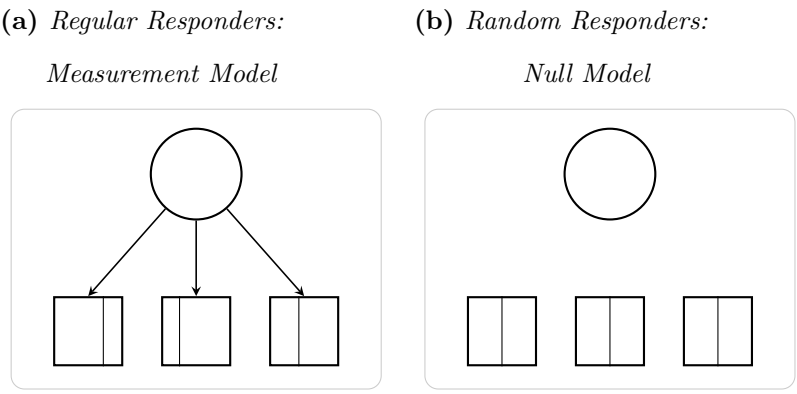
as students had to remain seated in class and everyone also gets the same break time.

Outcome: Random Responder Status

A mixture item response theory model framework (van Laar & Braeken, 2022) was adopted to operationalize and define the target outcome variable of interest, the random responder status of a student on a particular scale in the TIMSS 2015 student questionnaire. The approach assumes that there are two distinct, yet unobserved latent groups of responders in the population expressing different response behavior on a questionnaire scale: regular or non-random responders and random responders (see Figure 1).

Figure 1

Mixture IRT model Framework to Define and Operationalize Random Responders in terms of Independence and Uniformity of Item Responses.



Note. Symbols follow standard path diagram conventions, with squares representing observed variables (i.e., item responses); circles, latent variables (i.e., trait to be measured by the scale of items); arrows indicating dependence relations; vertical lines, response category thresholds. Reprinted under the terms of CC-BY-NC from "Random responders in the TIMSS 2015 student questionnaire: A threat to validity?" by S. van Laar and J. Braeken, 2022, *Journal of Educational Measurement*.

The regular responders are expected to respond consistently according to their own opinions and beliefs related to the questionnaire content of the items on the scale, in line with a traditional latent variable measurement model (see Figure 1a, the 'circle' is the common cause of the 'squares') such as the graded response model (Samejima, 1969). In contrast, the random responders are expected to provide responses that do not reflect their opinions and beliefs, but are more haphazard, in line with a null model implying

mutually independent item responses that have an equal chance of falling in either of the possible response categories (see Figure 1b, the 'squares' are mutually disconnected, nor influenced by the 'circle'; all squares are divided into uniformly equal category parts).

Under the mixture IRT model, the likelihood of a person p's item response pattern $\mathbf{y}_p$ (see Equation 1) is written as a weighted sum of the two mentioned model expressions: the joint probability of the observed item response pattern given the person's latent trait value under the graded response model multiplied by $\Pr(\backslash RR)$ the prior probability for a person to be a member of the regular responder group plus the joint probability of the observed item response pattern under the null model multiplied by $\Pr(RR)$ the prior probability for a person to be a member of the random responder group.

$$\begin{aligned} \mathcal{L}(\mathbf{Y}_p = \mathbf{y}_p) = & \\ & \Pr(\backslash RR) \quad \textstyle\prod_i \Pr(Y_{pi} = y_{pi} | \theta_p, \backslash RR) \\ & + \\ & \Pr(RR) \qquad \textstyle\prod_i \Pr(Y_{pi} = y_{pi} | RR) \end{aligned} \tag{1}$$

Notice that this mixture model has only one additional to-be-estimated parameter compared to the regular measurement model. The part of the model accommodating the possibility of random responders in the population, only has fixed parameters as item response probabilities are known and assumed to be uniformly equal across categories and items. Given that the mixture weights sum up to one by definition (i.e., $\Pr(RR) + \Pr(\backslash RR) = 1$), only one extra parameter needs to be estimated. $\Pr(RR)$ can be interpreted as a model-based estimate of the prevalence of random responders on the questionnaire scale. The resulting estimated model can be used to classify individuals according to their individual item response pattern in one of the two classes based upon their maximum posterior class membership probability. Thus, on the particular questionnaire scale, an individual student is (classified as) a random responder $RR_p = 1$ if $\Pr(RR_p = 1 | \mathbf{y}_p) = \frac{\Pr(RR) \prod_i \Pr(Y_{pi} = y_{pi} | RR)}{\mathcal{L}(\mathbf{Y}_p = \mathbf{y}_p)} > .5$, and $RR_p = 0$ otherwise. For each scale in the questionnaire, such a mixture IRT model will be estimated and used to compute the random responder status of the individual students having responded to that scale,

resulting in a binary profile of random responder status across scales in the questionnaire for each individual.

If the mixture model for a questionnaire scale failed either of two quality checks, the corresponding random responder status for that scale was set to missing for all students: (1) When the measurement model of the regular responder class had two or more standardized item discrimination parameters (i.e., factor loadings) below .40, the scale was considered unscalable for the majority population (i.e., no clean unidimensional scale structure); (2) When classification entropy dropped below .70 we concluded that the mixture model was unable to provide a good enough distinction between the two latent groups of responders.

Study Design: Sample & Student Questionnaire

Sample

Inclusion criteria. We study the students participating in TIMSS 2015 in the set of countries that have a so-called separated science program where all four subjects (i.e., biology, chemistry, earth science, and physics) are taught as independent subjects in the curriculum (instead of as part of one big integrated science subject). This choice is motivated by the useful features it brings to our study design: The student questionnaires in these countries contain extra scales and additional structure, as now students' values and attitudes were asked towards the four different science subjects instead of the single integrated science subject in other countries. The higher number of scales is beneficial for the study of intra-individual variability across scales and the additional questionnaire structure allows investigating whether subject matter or scale-specifics could be potential triggers for random responding.

Exclusion criteria. Although Malta and Sweden follow a separated science program, their students do not necessarily follow all four science subjects, and hence these countries were excluded from our sample. Lebanon and Morocco were excluded from the sample as the random responder mixture classification did not meet the required quality criteria for the majority of questionnaire scales. The latter points to larger discrepancies in those countries such as the scales not being unidimensional and/or specific items being

unscalable.

Effective sample. Applying inclusion and exclusion criteria to TIMSS 2015 results in the following set of seven countries (ISO-code) in our study: Armenia (ARM), Georgia (GEO), Hungary (HUN), Kazakhstan (KAZ), Lithuania (LTU), Russia (RUS), and Slovenia (SVN). In both Armenia and Georgia, a single (but different) questionnaire scale did not meet the classification quality criteria and here the random responder status RR_p was set to missing for all students on that scale in the corresponding country (i.e., Confidence in chemistry for Armenia and Like learning earth science for Georgia).

TIMSS Student Questionnaire

The random responder status of a student will be estimated for 19 scales in the TIMSS student questionnaire. These scales were each intended to reflect a unidimensional construct and contained between 7 to 10 Likert items for which a student needed to indicate to what extent s/he agrees with the given statement or indicate how often a specific situation has occurred to them on a 4-point response scale, ranging from 1 (*agree a lot* or *at least once a week*) to 4 (*disagree a lot* or *never*). The scales cover constructs such as students' sense of belonging, bullying, value of mathematics, value of science, and a set of three scales on like learning, views on engaging teaching, and confidence in each of five school subjects (mathematics, biology, earth science, chemistry, and physics). Note that the set of starting questions on students' background and home educational resources will not be considered in our analyses as those were single items of varying response formats that did not form a reflective scale.

Statistical Analysis

To determine the random-responder status of each student on the different scales, the confirmatory mixture IRT model of van Laar and Braeken (2022) for ordered polytomous indicators was run independently per scale-by-country combination (for sample Mplus syntax, see Appendix A). To determine latent class random responder profiles across the whole questionnaire, an exploratory sequence of unstructured latent class models for binary indicators was fitted independently per country, and the number of classes (i.e., profiles) was determined by means of BIC (e.g., Nylund et al., 2007). For gen-

eralization and interpretability, we match-aligned the resulting classes across countries. The expectation is that each country will show at the minimum a majority class with a close-to-consistent profile of non-random responding across the questionnaire, whereas expectations for the number and profile-type of additional classes are less clear.

Model-implied random responder rates will be computed to visualize the latent classes and these profiles will be supplemented by class-specific within-person statistics summaries. For the latter, we first compute for each individual a set of statistics based on their random responder profile (i.e., the binary sequence of their random responder status across scales in the questionnaire). The number of runs (i.e., a sequence of constant random responder status across subsequent scales) and the maximum run length would inform about the individual within-questionnaire consistency. Their switching behavior is more directly quantified through the 1st order transition probabilities giving the probability of (not) being a random responder on the current scale given that you were (not) a random responder on the previous scale (i.e., $\Pr(RR_{\text{scale}} = 1|RR_{\text{previous scale}} = 1)$ and $\Pr(RR_{\text{scale}} = 0|RR_{\text{previous scale}} = 0)$. The number of Guttman errors (Guttman, 1950) in the binary sequence formed by the profile informs about the level of within-person sequential inconsistency and a high number of errors would be incompatible with the earlier mentioned back-random responding profile.

Both the mixture IRT models and the latent class models were estimated using full-information maximum likelihood in Mplus Version 8.2 (Muthén & Muthén, 1998–2017) through the MplusAutomation package for R version 0.7-3 (Hallquist & Wiley, 2018), with robust standard errors and the expectation-maximization acceleration algorithm with a standard of 400 random starts, 100 final stage optimizations, and 10 initial stage iterations. All analyses accounted for the TIMSS sampling design by applying the total student weights in Mplus for the models and through the survey R package (Lumley, 2020) for the descriptive statistics. Analysis scripts were run under R version 4.0.0 (R Core Team, 2020).

Results

Descriptives

About 4500 students filled in the TIMSS 2015 student questionnaire in each of the seven countries, with sample sizes ranging from $n = 4028$ in Georgia to $n = 4917$ in Armenia. The prevalence of random responders among students in these countries varied across the scales, with the minimum RR prevalence observed on the Student Bullying scale (across-countries median prevalence = 0%) and the maximum RR prevalence on the Students Confident in Physics scale (across-countries median prevalence = 17%). On average, Georgia had the highest median within-person RR proportion across scales (11% or 2 out of 18 scales), while at least half of the students were not identified as RR on any scale (median within-person RR proportion = 0%) in 4 out of 7 countries (i.e., Kazakhstan, Lithuania, Russia, Slovenia).

Missing random responder status RR_p**.** When no item responses on an entire scale were observed for an individual student, there was also no data to assign a posterior class membership to the student and the student's random responder status on that scale was set to missing. About 4% of the students did not have a random responder status on 1 to 2 scales, and for another 4% this was the case on 3 or more scales. Across countries, slightly higher missingness percentages were observed for Georgia (12% missed 1 to 2 scales and 6% missed 3 or more scales) and Armenia (6% missed 1 to 2 scales and 6% missed 3 or more scales). Reasons underlying the missing responses are unknown and could be ascribed to a multitude of factors leading the student to either purposefully or accidentally skipping a page and as such an entire scale of the questionnaire. The missingness was generally observed to be randomly distributed across scales with missingness percentages on average below 6%. The exception was Georgia where missingness was concentrated on scales linked to the earth science subject (up to 10% of students were missing their random responder status on these scales). Yet, across all countries the majority of students, on average about 92% (ranging from 82% in Georgia to 97% in Lithuania and Russia), had a random responder status ($RR_p = 1$: yes/0 : no) for each of the administered scales in the TIMSS 2015 student questionnaire. Thus, for the

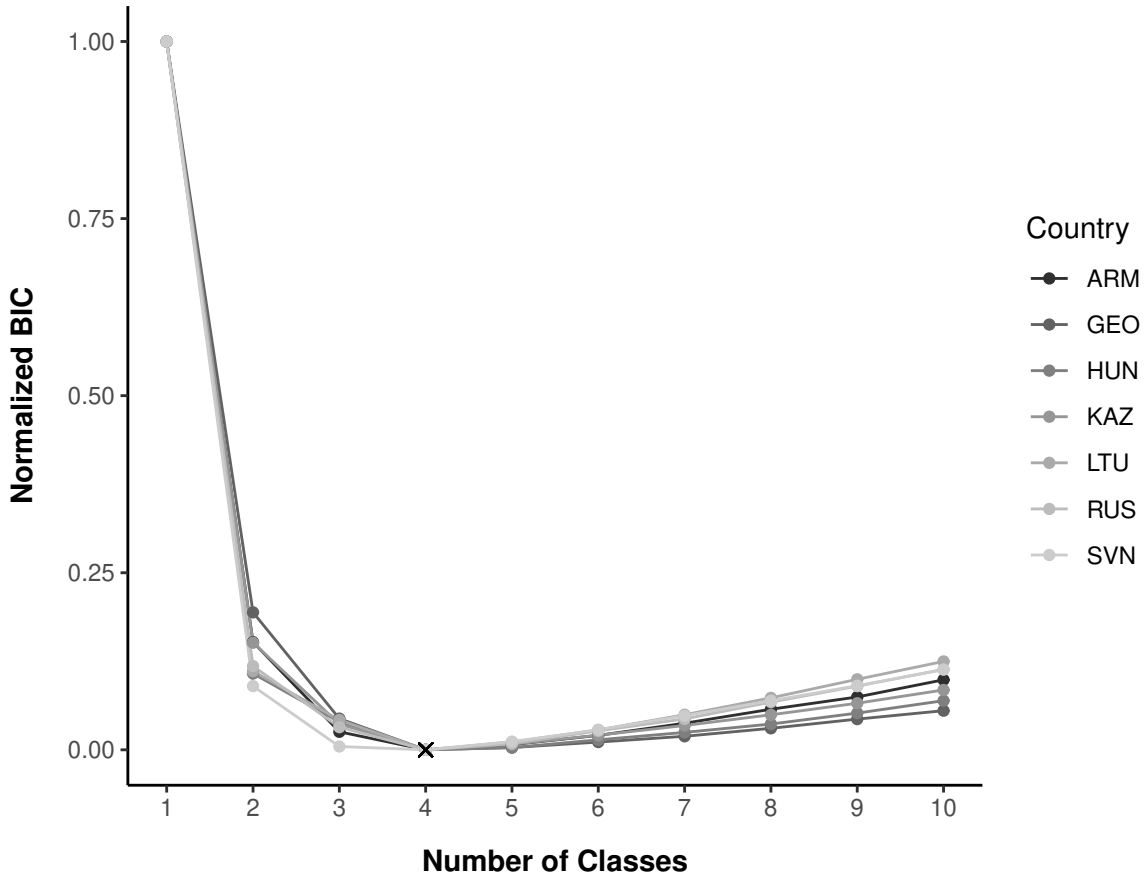
subsequent latent class analyses modeling random responder profiles across scales, a student's missing random responder status on a scale will be treated as missing at random. When computing the within-person descriptive statistics, a student's random responder status pattern across scales is used as is, skipping missing classifications (i.e., this implies that a missing status on a scale does not break a run). This treatment brings a slight within-person consistency bias, but given the low amount and randomness of missings, the inferential impact can be expected to be limited.

Latent Class Profiles of Random Responder status across the questionnaire

When determining the number of latent classes, the normalized BIC plot showed uniformly across countries a huge drop after one class followed by a quadratic inverse U-shape with the minimum at 4 (see Figure 2). To complement the relative perspective offered by the normalized BIC plot, we also computed model-weights based on raw BIC values (Wagenmakers & Farrell, 2004). Model weights for the 4-class solution were close to the boundary value of 1, clarifying that the 4-class solution is also the single preferred solution. Entropy values for the 4-class models ranged from .73 to .82, indicating that students could be classified in a rather clear crisp fashion across the classes. These model comparison results support the hypothesis of population heterogeneity in terms of random responding across the scales in the student questionnaire and suggest there to be four distinct latent profiles in each of the seven countries.

Figure 2

Normalized Bayesian Information Criteria as a function of the Number of Latent Classes.



Note. The Bayesian information criteria (BIC) were normalized, where in each country the normalized BIC = [BIC-min(BIC)]/[max(BIC)-min(BIC)]. As a result, the latent class model with the highest BIC has a value of 1 and that with the lowest BIC has a value of 0. The model with the lowest BIC (indicated by the cross symbol in the plot) has the better balance between goodness-of-fit to the data and model complexity and is to be selected for inference and generalization purposes.

Figure 3 displays the resulting random responder status probability profiles for the four-class solution in each of the seven countries. For each class, the vertical axis represents the probability of having a positive random responder status on a given scale in the student questionnaire. On the horizontal axis, the respective scales are listed in order of occurrence in the questionnaire and can be grouped (cf. dotted gray vertical lines) in terms of the particular subject domain (cf. single gray letter) they relate to.

The four classes corresponded to four distinct random responder across-scales profiles

that could be neatly matched across countries. Table 1 provides per class an overview of relevant within-person statistics that can help to further characterize what type of within-person random responder across-scales patterns can be observed in each of the four classes.

Consistent non-random responders. The profile for the majority class (Figure 3: lightest gray with diamond points) indicates close to zero probabilities of a positive random responder status for all scales, suggesting that this class bundles students that are (almost) never classified as a random responder on any of the scales. This is further corroborated by an across-students average within-person maximum run length of a zero random responder status close to the total of scales in the questionnaire, a first-order transition probability $\Pr(RR_{\text{scale}} = 0 | RR_{\text{previous scale}} = 0)$ close to 1, and hardly any Guttman errors (Table 1). All these average within-person statistics imply large within-person consistency and point to generally (close-to) all-zeroes random responder status patterns for students in this class.

Random responders triggered exclusively by the Confidence scales. The profile for a second class runs rather parallel with the majority class, were it not for the substantially higher probabilities of positive random responder status on the five subject-specific Confidence scales (Figure 3: darkest gray with square points). Notice that this pattern indeed repeats across countries, although the peaks at the confidence scales do vary in height across countries (in Armenia the Confidence in Chemistry scale is absent as it did not meet quality criteria). This profile suggests that this class bundles students that are exclusively classified a as random responder on the confidence scales, and not elsewhere in the questionnaire. This is further corroborated by students having on average a positive random responder status on 57% of the Confidence scales, but on only 6% of the other scales in the questionnaire.

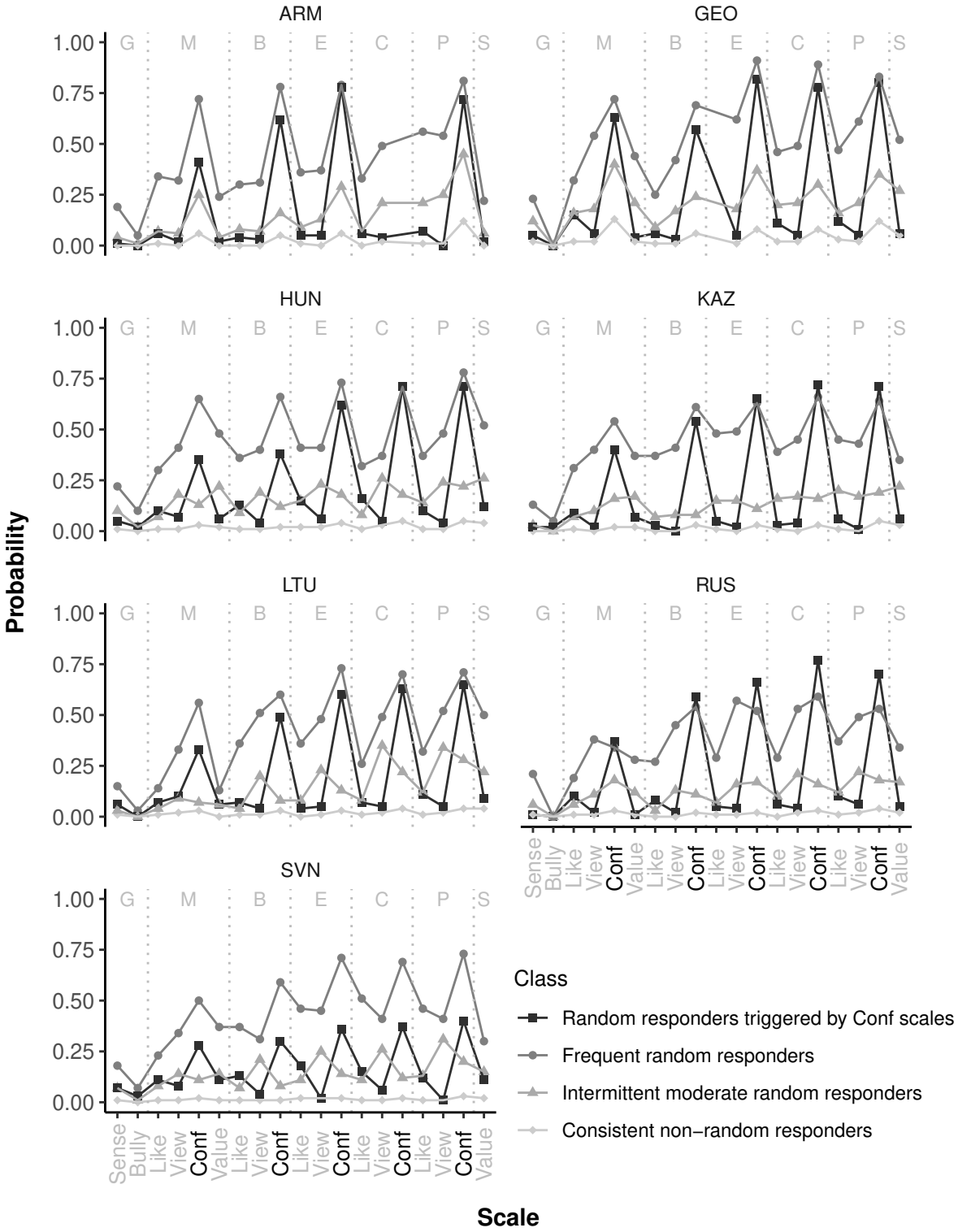
Frequent random responders. The response probability profile for the minority class (Figure 3: dark gray with circle points) has the highest probabilities of a positive random responder status for all scales, suggesting that this class bundles students that are frequently classified as a random responder on any of the scales (i.e., on average

on about 43% of the scales). The individual student patterns in this class are far from consistent, with transition probabilities close to 50/50 in either way, and many Guttman errors (see Table 1).

Intermittent moderate random responders. The profile for the third class indicates non-zero but low probabilities of a positive random responder status for most scales (Figure 3: light gray with triangle points). Individual students have on average on about 3 scales (i.e., 16% of the questionnaire) a positive random responder status, spread out across two runs, and resulting in on average 4 Guttman errors (see Table 1). Together with the response profile of this class, these within-person statistics results imply that random responding in this class is more intermittent across the questionnaire and without clear systematic trends across students in this class.

Figure 3

Random Responder Status Probability Profiles for the Four-Class Solution.



Note. $\Pr(RR = 1)$ = class-specific probability of a positive random responder status (i.e., student classified as a random responder by the mixture IRT model for the scale). The scales on the horizontal axis appear in order of occurrence in the TIMSS 2015 student questionnaire. The dotted vertical lines divide questionnaire scales by subject: G = General, M = Mathematics, B = Biology, E = Earth Science, C = Chemistry, P = Physics, S = Science.

Table 1

Class Sizes and Class Averages of Within-Person Statistics across seven Countries.

	Consistent non-RR	RR Conf	Frequent RR	Intermittent RR
Class Size	3295	490	186	641
Class Proportion	[.53, .76]	[.07, .24]	[.03, .07]	[.10, .18]
Pr(RR=1)	[.01, .04]	[.17, .24]	[.40, .51]	[.15, .22]
Number of Runs (RR=1)	[0, 1]	[3, 4]	[4, 5]	[2, 3]
Number of Runs (RR=0)	[1, 2]	[4, 5]	[4, 5]	[3, 4]
Maximum Run Length (RR=1)	[1, 1]	[1, 2]	[3, 5]	[1, 2]
Maximum Run Length (RR=0)	[15, 18]	[6, 9]	[4, 5]	[8, 10]
$\Pr(RR_{\text{scale}} = 1 \mid RR_{\text{previous scale}} = 1)$	[.00, .03]	[.05, .12]	[.43, .58]	[.15, .25]
$\Pr(RR_{\text{scale}} = 0 \mid RR_{\text{previous scale}} = 0)$	[.96, .99]	[.67, .81]	[.49, .59]	[.77, .85]
Guttman Error	[0, 1]	[5, 6]	[6, 8]	[4, 6]

Note. The median class size across countries is reported. The reported intervals provide the range of average values across the seven countries. A run is liberally defined here as a within-person sequence of constant random responder status across scales. A missing status for a scale was ignored and considered to not break up a run.

Back-random responders. None of the class profiles corresponded to the pattern you would expect under back-random responding. Among all 32086 students, there were only 590 students who showed a non-zero proportion of random responding in combination with zero Guttman errors, a set of within-person statistics that would surface under back-random responding. Yet of those 590 students, 506 responded randomly only to the last scale (and 55 to the two last scales). Similarly, among all 32086 students, only 34 and 114 students had 4 out of the last 6 scales and 3 out of the last five scales a random responder status and a non-responder status elsewhere. These findings seem to suggest that the occurrence of back-random responding is at most a rarity in our sample with the TIMSS 2015 student questionnaire.

Discussion

In this study, we explored intra-individual variability in random responding behavior across questionnaire scales in the TIMSS 2015 student questionnaire. The objective was to clarify to what extent random responding is a more systematic trait-like behavior or more state-like as in activated once a personal threshold has been breached (cf. back-random responding) or when triggered by specific contents or type of a scale, or more haphazard due to more idiosyncratic instances.

Our latent class analyses uniformly converged on a four-class solution with four distinct random responding profiles that generalized well, both in class size as well as in profile character, across the seven countries under study. Do note that the sample contains eighth-grade students in countries with a separated science program in their school curriculum and that are mostly located in Eastern Europe. Hence, there are potentially some shared contextual influences that need to be taken into account when extrapolating results outside this age group or towards other countries elsewhere.

The identified majority class reflects within-person profiles in which no random responding occurs on scales throughout the questionnaire. Although the students have nothing to gain or lose from filling in the low-stakes questionnaire, the majority appears to respond to the scales in a rather construct-consistent manner. This is a reassuring finding for TIMSS 2015, and by extension also for the potential of other low-stakes educational surveys.

In contrast, the identified minority class reflects within-person profiles that randomly responded on almost half of the scales in the questionnaire. This class profile can be speculated to correspond well to explanations of random responding in terms of carelessness, insufficient effort, or lack of motivation and seriousness on behalf of the respondent (e.g., Huang et al., 2012). The relatively high frequency of random responding also questions the general trustworthiness of the delivered responses on the questionnaire by those students, even on scales for which the student was not classified as a random responder.

A slightly larger class shows an intermittent random responding pattern across the questionnaire with a much more moderate frequency of occurrence, about 3 scales or 16%

of the questionnaire. Here, random responding can be considered more incidental due to undefined idiosyncratic features and occasional lapses of engagement by the student. To the extent that this is indeed a reflection of completely-at-random events, data quality and inferences should remain relatively unharmed.

In contrast to the former classes for which there were no obvious systematic observable triggers for the random response behavior, a fourth class reflected within-person profiles where the students responded randomly but exclusively to the confidence scales. Such a systematic pattern cannot be ascribed to momentary lapses in engagement or insufficient effort, nor to response type artefacts (4-point Likert items were used uniformly across the questionnaire), but we ought to look at item contents. Participants in a study by Baer et al. (1997) also reported that their core reasons for random responding were not lapses of concentration or boredom, but mostly difficulties in understanding items or deciding on the response. A similar phenomenon could be at play here. Perhaps the students in this latent class genuinely find it uncomfortable to publicly disclose their confidence in school subjects? Examples of items on such a confidence scale are for instance "mathematics is harder for me than any other subject" or "mathematics is more difficult for me than for many of my classmates". Students' perceptions about themselves are always made in comparison to some standard, either internally (i.e., own performance in one subject with own performance in another subject) or externally (own performance with the performance of other students) (e.g., Marsh & Hau, 2004). Items that require comparisons, with additional changing or ambiguous standards and definitions of self, might just be more difficult to answer or could result in internal inconsistencies in perception for certain individuals. Hence, one cannot exclude the possibility that these students in fact provided genuine valid responses from their individual viewpoints. This type of more systematically triggered random responding could be considered more harmful than the intermittent random responding and, in our particular case, this raises questions for the validity and data quality of the confidence scales in the questionnaire.

We found no support for so-called back-random responding profiles (e.g., Clark et al., 2003; Gallen & Berry, 1997), where students are assumed to switch from regular

responding to random responding once they reached their 'threshold'. The lack of back-random responding implies that explanations in terms of a full-blown depletion of internal cognitive resources or alternatively a firm conscious decision to no longer actively engage with the questionnaire are not applicable. TIMSS states that there is ample time for the student to fill in the questionnaire, that the general task demands of the assessment are not out of bounds, and that there is also no benefit in rushing through the survey (one has to stay in class anyhow for the allotted time). Hence, this non-speeded no-rush low-stakes character of the questionnaire potentially sets (part of) the context to the null-finding on back-random responding, and we caution against generalizing this finding to educational surveys under more rushed speededness conditions.

Currently, we were restricted to implementing a scale-level mono-method assessment of random responding (van Laar & Braeken, 2022) which operationalized random responding as providing item responses on a scale more alike patterns resulting from a random responder reference group than alike the consistent response pattern by the 'regular' population. This does mean that some aberrant response patterns won't be picked up (e.g., straightlining on scales with unidirectional items). At the inferential level, the current approach also ignores the classification uncertainty in random responder status assessment as the maximum posterior binary membership classification is used as a binary outcome. To allow a broader grip and strengthen the detection of random responders, auxiliary data is needed that allows for a multi-method approach using for instance bogus items (e.g., "I am not in grade eighth"), instructed-response items (e.g., "Please mark slightly agree"), duplicate items (cf. so-called lie-scales in personality questionnaires), and the provision of individual survey completion speed indicators (e.g., Leiner, 2019).

Conclusion

Whereas non-compliance and manipulation checks have become more and more a default part of experimental design, similar data quality checks or monitoring procedures are not yet commonplace in educational science research using questionnaires and surveys. Ethical questions remain as this monitoring needs to be communicated to the participants and potentially creates an atmosphere of mistrust and ambiguity which might also

aversively affect the quality of response. From a perspective of implementing a monitoring system, the intermittent character of random response behavior in one of the classes does not hold much promise for the value of an early warning monitoring system and one might even wonder whether the added value of monitoring outweighs the potential negative impact of the 'big brother is watching you' impression that such a monitoring system might bring along to the students. The confidence-exclusive class also sketches that, in a questionnaire context where there is no objectively correct answer, the label 'random' might also be a misnomer; let alone imagine the potential mishap when one would ascribe it to insufficient effort and actively communicate it as such to a participant. Given these practical complications and the finding that the majority of students do not end up as random responders in the survey, we would advise against implementing an active monitoring system with early warnings for the students. Instead, we would advocate for a passive monitoring system, including survey completion speed indicators, to support post-survey response data quality checks and including student-level diagnostic indicators in the publicly available datasets of the survey to allow secondary data analysts to run proper sensitivity checks to assure robustness of their research findings. Next to enabling such sensitivity checks, we should also not underestimate proper survey design and here one can step things up in educational research by valuing the power of cognitive labs and the feedback of survey panels that are not filled with 'academic experts' on the constructs to be measured, but with members of the actual target group. The information these panels bring can potentially help detect, before any large-scale implementation, the unintended triggers to random response behavior in the questionnaire.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163–181.

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI–A. *Journal of Personality Assessment*, *68*(1), 139–151.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218–229.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(2), 718–738.

Breitsohl, H., & Steidelmüller, C. (2018). The impact of insufficient effort responding detection methods on substantive responses: Results from an experiment testing parameter invariance. *Applied Psychology*, *67*(2), 284–308.

Cao, J., & Stokes, S. L. (2007). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*(2), 209–230.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment*, *15*(2), 223–234.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

Eccles, J., & Wigfield, A. (2002). Motivational beliefs, values and goals. *Annual Review of Psychology*, *53*(1), 109–132.

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

Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, *13*(2), 83–87.

Gallen, R. T., & Berry, D. T. R. (1997). Partially random MMPI-2 protocols: When are they interpretable? *Assessment*, *4*(1), 61–68.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 66–90). Princeton University Press.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229–248.

Linnenbrink, E. A., & Pintrich, P. R. (2002). Motivation as an enabler for academic success. *School Psychology Review*, *31*(3), 313–327.

Liu, T., Sun, Y., Li, Z., & Xin, T. (2019). The impact of aberrant response on reliability and validity. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 133–142.

Lumley, T. (2020). Survey: Analysis of complex survey samples [R package version 4.0].

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.

Marsh, H. W., & Hau, K.-T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Educational Psychology*, *96*(1), 56–67.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach* (pp. 161–200). Lawrence Erlbaum.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*(3), 215–237.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218.

Muthén, L. K., & Muthén, B. O. (1998–2017). Mplus User's Guide. Eighth Edition.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079.

Pinsoneault, T. B. (2007). Detecting random, partially random, and nonrandom Minnesota Multiphasic Personality Inventory-2 protocols. *Psychological Assessment*, *19*(1), 159–164.

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: a systematic review of 12 years of educational research. *Studies in Science Education*, *50*(1), 85–129.

R Core Team. (2020). R: A language and environment for statistical computing.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3–8.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97.

Schmitt, M., & Blum, G. S. (2020). State/Trait Interactions. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 5206–5209). Springer International Publishing.

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

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205.

Appendix A

Sample Mplus syntax of the mixture IRT model for the 'students value mathematics' scale in Kazakhstan.

```
TITLE: Kazakhstan_SQM20;

DATA: file = "KAZ_SQM20.dat";

VARIABLE:
  names = IDSCHOOL IDSTUD TOTWGT
    BSBM20A BSBM20B BSBM20C BSBM20D
    BSBM20E BSBM20F BSBM20G BSBM20H BSBM20I;
  missing = .;
  usevariables = BSBM20A BSBM20B BSBM20C BSBM20D
    BSBM20E BSBM20F BSBM20G BSBM20H BSBM20I;
  categorical = BSBM20A BSBM20B BSBM20C BSBM20D
    BSBM20E BSBM20F BSBM20G BSBM20H BSBM20I;
  idvariable = IDSTUD;
  weight = TOTWGT;
  cluster = IDSCHOOL;
  classes = c(2);

ANALYSIS:
  type = mixture complex;
  algorithm = INTEGRATION EMA;
  estimator = MLR;
  process = 3;
  starts = 400 100;

MODEL:
%overall%
  F BY BSBM20A-BSBM20I*;
  F@1;
  [F@0];
%c#1%
  F BY BSBM20A-BSBM20I*;
  F@1;
  [F@0];
  [BSBM20A$1-BSBM20I$1];
  [BSBM20A$2-BSBM20I$2];
  [BSBM20A$3-BSBM20I$3];
%c#2%
  F BY BSBM20A-BSBM20I@0;
  F@0;
  [F@0];
  [BSBM20A$1-BSBM20I$1@-1.09861228866811];
  [BSBM20A$2-BSBM20I$2@0];
  [BSBM20A$3-BSBM20I$3@1.09861228866811];

OUTPUT: stdyx;

SAVEDATA:
  file = cpr_KAZ_SQM20.dat;
  format = free;
  save = cprobabilities;
```