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

Contact information: Saskia van Laar, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: saskia.vanlaar@maastrichtuniversity.nl

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

To give a brief overview of the thesis, the first part consists of two articles that are more methodological and focus on model fit evaluation with the comparative fit index and address the meaning of null baseline model comparisons. However, the second part of the thesis consists of four articles that include data analysis related to IEA studies. Specifically, the studies in this part addressed the issues of random response behavior in the TIMSS 2015 student questionnaire throughout different empirical studies.

Background

The central theme guiding these studies is that we know that the results of international large-scale educational assessments are widely used for research and educational policy (e.g., Hernández-Torrano & Courtney, 2021; Hopfenbeck et al., 2018). Yet, it is crucial to pay attention to the quality of these assessment results (e.g., Gustafsson, 2008) as the quality of the conclusions that are drawn is dependent on the quality of 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 not unique to international large-scale educational assessments in general, but 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 for an individual student). 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. Depending on the severity, if students would not respond accurately or thoughtfully, this could potentially lead to problems with the use and interpretation of the assessment results. Any type of inferences based on these assessments, either research conclusions or policy changes, are at risk of being invalidated. Thus from a quality assurance viewpoint, it is important to pay attention to undesirable or invalid response behavior.

Contributions. With these studies we tried to assess and identify invalid response behavior in the TIMSS 2015 student questionnaire and subsequently, the studies investigated the prevalence, impact, and characteristics of random responders, as well as where and how frequent random response behavior occurs. Ultimately, the studies hopefully increase awareness about the need for data quality monitoring for the survey part of international large-scale educational assessments, while the specific methodology used might offer a useful tool that can be used in the context of measurement validity and data quality.

Addressing Invalid Responding

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 ILSA's. Within this context, the two most common approaches to address the validity of item responses are using either self-report measures or response times.

Self-report Measures. In short, self-report measures are used to collect information about the quality of item responses on the assessment, often in relation to achievement. One drawback of this approach is that is an indirect and global measurement approach. With this approach, there is no focus on the actual responses given on the assessment by the students, students are merely asked to rate their behavior on the assessment and they are often used to try to say something about a test or questionnaire as a whole. This also implies there is no direct link with the actual responses that people want to draw conclusions about.

Response Times. The response time approach, also an indirect approach, is often applied in achievement testing, 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). While this approach allows to address the quality of each item separately, currently, the biggest disadvantage is that item-level response times are not always available. 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.

Response-based Approach. In our studies we have adopted a response-based approach. There are some other instances where response-based approaches are used in the context of ILSA's (for an example of the application of person-fit indices see e.g., Hopfenbeck & Maul, 2011). However, these measures can be very generic/aspecific in a sense that they pick up on multiple response patterns. In our adopted approach, we specifically tried to create a direct link between the conceptual definition of invalid response behavior and the quality of the item responses. It is important to notice that our approach is more locally oriented, using information related to the items of interest. Yet, while our approach uses the actual item responses as provided by the students, the response quality is addressed at scale-level.

Conceptual Framework

Definition of Random Responders

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. 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) 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). In the end it comes down to the idea that 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 about the student with respect to the assessment content (e.g., van Laar et al., 2024)

Identifying Random Responders

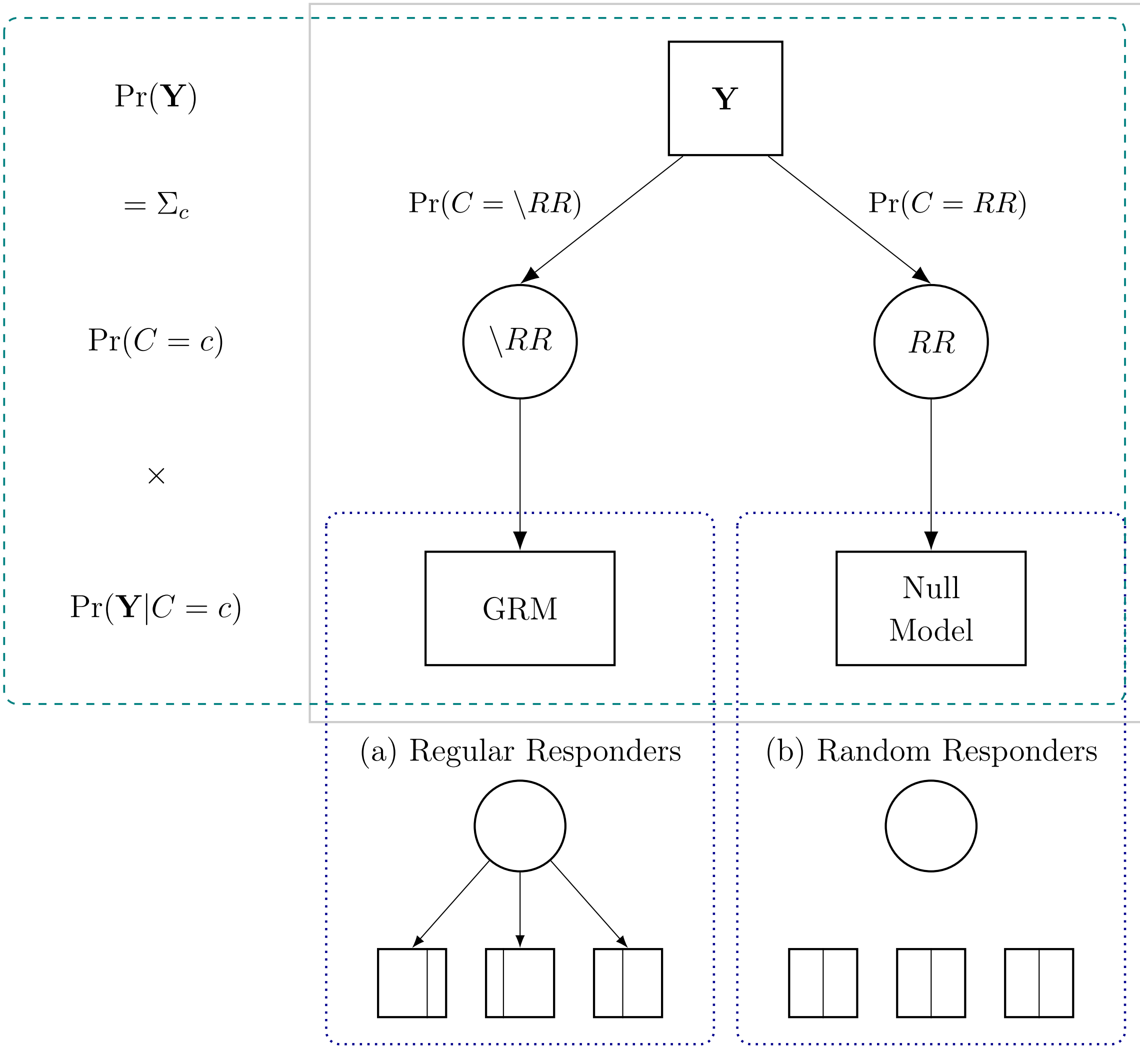
The studies are all built around the development of a mixture IRT model to distinguish between different types of responders. With this approach we make the implicit assumption that there are two groups of responders in our sample, regular responders (i.e., individuals who are genuinely answering to the questionnaire) and random responders (i.e., individuals who are choosing responses randomly throughout as if they disregard what is asked from them), and the students in both groups are expected to show different response behavior. While in practice knowledge on who is in which group is not available in the data, we can use the mixture model to classify students based on the class they belong to with the highest probability given their response pattern. The underlying idea is also represented in Figure 1 and provides a direct link with the observed item response patterns given on a survey scale.

Regular responders are expected to respond consistently according to their own opinions, abilities, or beliefs related to the assessment content. In terms of the mixture component model, these students follow a systematic measurement model (see panel (a) in Figure 1) 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. The random responders are expected to provide unsystematic random response patterns across the items, such that the responses no longer reflect a student's own opinions, abilities, or beliefs. Their mixture component model can be seen as an application of a null model, where all observed variables are assumed to be uncorrelated, but with the additional restriction that observed variables are uniformly distributed (i.e., each response category has an equal probability of being selected). In panel (b) of Figure 1 this is represented by the latent trait no longer being connected with the item responses and all squares being divided into equal (category) parts.

One important consideration here is that while 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, 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 indicate how or why random response patterns came to be.

Figure 1

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.
Symbols in panel (a) and (b) 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.

Research Questions

While all studies revolve around random responding in the TIMSS 2015 student questionnaire, they focus on different aspects of this behavior/phenomenon. Throughout the studies we focused on the following research questions: (1) How many of the students would qualify as random responders?; (2) Do scale-related inferences change when we exclude random responders? [Study 1]; (3) Do scale position and questionnaire length influence the prevalence of random responders? [Study 2]; (4) Which students are more likely to be identified as random responders? [Study 3]; (5) How consistent are students responding randomly? [Study 4].

Data

In the different studies we have used data from the TIMSS 2015 student questionnaire. The student questionnaire is one of the contextual questionnaires (e.g., home questionnaire, teacher questionnaires, school 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).

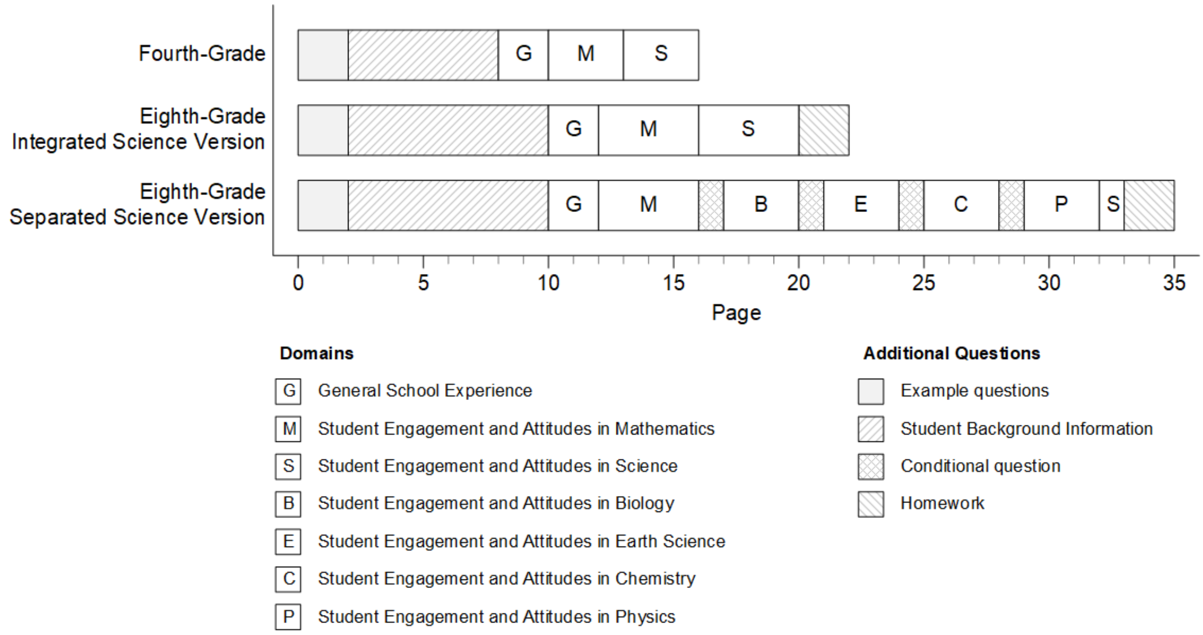
We specifically focus on the student questionnaire, because in practice, the student questionnaire is used by a lot of researchers to answer a wide variety of research questions and it serves an important role in putting the achievement results into context. In addition, the student questionnaire also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

Up to this point, the student questionnaire also seems to have 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.

Furthermore, we also utilized the fact that the TIMSS assessment is applied to two different age groups, and that the eighth-grade student questionnaire comes in two versions, which allows for the investigation of the role of questionnaire characteristics on response behavior. Figure 2 shows an overview of the scale structure of the student questionnaire(s) in the different grades. The fourth-grade student questionnaire contains eight standalone scales, while the student questionnaire for the eight-grade contains either 10 or 19 standalone scales. 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 2

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.

Highlights of Studies

Study 1 (van Laar & Braeken, 2022)

Given the limited information surrounding random responding in international large-scale educational assessments, the aim of the first 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 eighth-grade student questionnaire (integrated science version). To this end, the mixture IRT model approach was used to identify those students who would qualify as random responder in five different countries (six groups) on four different scales ('Confidence in' and 'Value of' Mathematics and Science) . 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%. 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). 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 in results and conclusions.

Study 2 (van Laar & Braeken, 2024)

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.

Study 3 (Chen et al., 2023)

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, the 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 several policy-relevant covariates in educational research (i.e., 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.

Study 4 (van Laar et al., 2024)

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 time, 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 and whether random responding is more a systematic trait-like behavior or more state-like. 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 and 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.

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.

Chen, J., van Laar, S., & Braeken, J. (2023). *Who are those Random Responders on your Survey? The case of the TIMSS 2015 student questionnaire. Large-scale Assessments in Education*, *11*, Article 377. https://doi.org/10.1186/s40536-023-00184-6.

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.

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.

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.

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

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.

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. https://doi.org/10.1111/jedm.12317.

van Laar, S., & Braeken, J. (2024). Prevalence of random responders as a function of scale position and questionnaire length in the timss 2015 eighth-grade student questionnaire. *International Journal of Testing*, *24*, 24–52. https://doi.org/10.1080/15305058.2023.2263206.

van Laar, S., Chen, J., & Braeken, J. (2024). *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*. *Measurement: Interdisciplinary Research and Perspectives*, *22*(1), 90–107. https://doi.org/10.1080/15366367.2023.2203972.

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., Soland, J., & Bo, Y. (2020). The (non)impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57–77.