Analyzing Multiple-Choice-Multiple-Response Items Using Item Response Theory

Trevor I. Smith,[1] Philip Eaton,[2] Suzanne White Brahmia,[3] Alexis Olsho,[4] Charlotte Zimmerman,[3] and Andrew Boudreaux[5]

[1]*Department of Physics & Astronomy and Department of STEAM Education, Rowan University, Glassboro, New Jersey 08028, USA*
[2]*School of Natural Sciences and Mathematics, Stockton University, Galloway, New Jersey 08205, USA*
[3]*Department of Physics, University of Washington, Seattle, Washington 98195, USA*
[4]*Department of Physics and Meteorology, United States Air Force Academy, USAF Academy, Colorado 80840, USA*
[5]*Department of Physics & Astronomy, Western Washington University, Bellingham, Washington 98225, USA*

Multiple-choice-multiple-response (MCMR) items allow students to select as many responses as they think are correct to a given question stem. Using MCMR items can provide researchers and instructors with a richer and more complete picture of what students do and do not understand about a particular topic. Interpreting students' MCMR responses is more nuanced than it is for single-response items. Unfortunately, many typical analyses of data from multiple-choice tests assume dichotomously-scored items, which eliminates the possibility of incorporating the rich information from students' response patterns to MCMR items. We present a novel methodology for using a combination of item response theory models to analyze data from MCMR items. These methods could be applied to inform scoring models that incorporate partial credit for various response patterns.

2022 PERC Proceedings edited by Frank, Jones, and Ryan; Peer-reviewed, doi.org/10.1119/perc.2022.pr.Smith

I. INTRODUCTION

The Physics Inventory of Quantitative Literacy (PIQL) is a 20-item multiple-choice test designed to assess students' physics quantitative literacy, i.e., reasoning quantitatively in ways typical of expert physicists [1]. One feature of the PIQL is the use of six multiple-choice-multiple-response (MCMR) items in addition to more typical single-response (SR) items. MCMR items may have more than one correct response, and allow students to choose all responses they think are correct. A benefit of using MCMR items on the PIQL is that researchers and instructors can get a richer picture of student understanding than is available with SR items alone, by probing fundamental quantitative reasoning as it interacts with physics concepts. Previous results have shown that students often choose at least one correct and one incorrect response simultaneously, which suggests their thinking is somewhere between completely novice-like and completely exert-like [1–4]. A major challenge in using MCMR items is deciding how such responses should be scored. (See Ref. [1] for more information about the PIQL and a more detailed discussion of the benefits and challenges of using MCMR items.)

One approach to scoring MCMR items is to categorize each response pattern as either completely correct (selecting all correct responses and no incorrect responses) or incorrect [1–3]. This type of dichotomous scoring allows the results from MCMR items to be combined with those of SR items in many typical quantitative analyses (e.g., classical test theory, factor analysis); however, this approach ignores the nuance and complexity of students' responses, which are the reason for using MCMR items.

To begin to address the need for a more nuanced scoring method, we have previously used a four-level categorization scheme that labels response patterns as Completely Correct, Some Correct, Both Correct and Incorrect, or Completely Incorrect [1–4]. Previous results using this four-level categorization have shown that at least 60% of students provide at least one correct response to each MCMR item on the PIQL, although this is often coupled with an incorrect response: 6%–45% of students select Both Correct and Incorrect response patterns [1]. This emphasizes the need to consider methods for assigning partial credit for responses that are not Completely Correct. Unfortunately, this four-level categorization does not give a definite answer to the question of how to score MCMR items. Is it better for a student to select only some of the correct response options, or all of the correct response options in combination with an incorrect option? How could these categories be translated into assigning partial credit for responses that are not Completely Correct?

In this paper we explore the use of item response theory (IRT) to analyze MCMR items, with an eye toward future work that could define partial credit for each response pattern. Smith, Louis, Ricci, and Bendjilali [5], and Eaton, Johnson, and Willoughby [6], have used IRT models for analyzing nominal (i.e., nondichotomous) data in order to rank responses according to how well they align with overall understanding for SR items on the Force and Motion Conceptual Evaluation (FMCE [7]) and the Force Concept Inventory (FCI [8]), respectively. Eaton, *et al.* proposed a method for assigning partial credit for incorrect responses based on their rankings [6]. Smith and Bendjilali have explored the relationship between IRT item parameters in Bock's nominal response model (NRM [9]) and student understanding to show that a student's selection of a specific incorrect response is an indicator of how well they understand the test material [10].

These prior uses of IRT nominal models [9, 10] and nested-logit models [5, 6, 11] build on well-established uses of IRT to analyze SR items; however, using them for MCMR items is not entirely straightforward. Applying Bock's NRM to MCMR items would require coding each possible combination of responses as its own category: 346 categories across the six MCMR items, requiring estimates of 836 parameters in total across both SR and MCMR items. This would require a sample size of at least 8360 students (10 students for each parameter estimated [12]), and many of the categories defined by MCMR response patterns are likely to be chosen very rarely (if at all), making the results unstable.

Alternatively, one could treat each individual response as an independent item that is either selected (1) or not selected (0) and use a dichotomous model for each (e.g., the two-parameter logistic, 2PL [13]); however, this assumption of independence may not be valid for MCMR responses, some of which may express opposite ideas (e.g., "Energy was added ..." vs. "Energy was taken away ...").

We present a novel approach in which we treat each MCMR item as a combination of independent dichotomous "response-items" and grouped nominal response-items. We choose which responses to group together by examining within-item correlations between responses, and we report the goodness-of-fit of our IRT models using various fit statistics. Our current goals are to demonstrate how standard IRT methods may be used in novel ways to analyze MCMR items, and to emphasize the decisions that must be made in order to apply these methods. Future studies will apply these methods to a large data set from a broad, diverse student population to inform decisions about assigning partial credit for various combinations of responses.

We seek to answer the following research questions:

1. How strongly correlated are students' responses within each MCMR item on the PIQL?
2. How does the fit of an IRT model that includes nominal response-items compare with the fit of an IRT model with only dichotomous response-items?
3. What is the optimal combination of dichotomous and nominal response-items?

We expect that analyzing strongly correlated responses independently provides results that contain redundant information; therefore, grouping these responses as nominal response-items will likely improve the fit statistics of our IRT model. We also expect that including nominal response-items created from weakly correlated responses will have a minimal (or perhaps negative) impact on the model fit.

II. DATA SOURCES AND METHODS

Data for this study come from 3399 students enrolled in the three-course calculus-based introductory physics sequence at a large public research university in the USA. Data were collected by administering the PIQL online at the beginning of each course in five consecutive academic terms using PIQL v2.2 [1] (available on the PhysPort website [14]).

We addressed research question 1 by calculating polychoric correlations for each pair of MCMR responses to determine which are chosen together (or separately) more often than would be expected by random chance. Kubinger showed that polychoric correlations are more appropriate for dichotomous data (e.g., whether or not a student chose a specific response) than more traditional Pearson correlations [15]. Eaton, Frank, and Willoughby found that polychoric correlations stronger than 0.613 suggest that responses are linked by something more than a simple underlying latent trait [16].

We addressed research questions 2 and 3 by defining a base model for our IRT analyses in which all 14 SR items on the PIQL were modeled using Bock's NRM, and all 34 MCMR response options were assumed to be independent and analyzed dichotomously using a two-parameter logistic (2PL) model. With this "All2pl" model, two parameters are estimated for each response option.

We addressed research question 2 by replacing each within-item pair of dichotomous response-items with one nominal response-item whose options included all four combinations of responses: 00, 01, 10, 11. We compared the modified model with the base model using various fit statistics, including the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis fit index (TLI) [17–19]. Lower values of AIC, BIC, and RMSEA, and higher values of CFI and TLI, indicate better model fits. Each IRT model with a single nominal response-item and 32 dichotomous response-items required a sample size of at least 2140 students to estimate 214 parameters.

We addressed research question 3 by ranking the correlations between response pairs from strongest to weakest. Beginning with the most highly correlated (or anticorrelated) MCMR responses, we replaced the two dichotomous response-items with a single nominal response-item. If the model with the nominal response-item had better fit statistics than the base model, we kept those responses paired, otherwise we considered the responses to be independent and modeled them as separate dichotomous items. We repeated this process for each successive pair of MCMR within-item responses, building a model with more and more nominal response-items. We terminated this iterative process when all significant correlations had been included as nominal response-items. The most complicated model we tested included seven nominal response-items and 17 dichotomous response-items, and required a sample size of at least 2280 students to estimate 228 parameters.

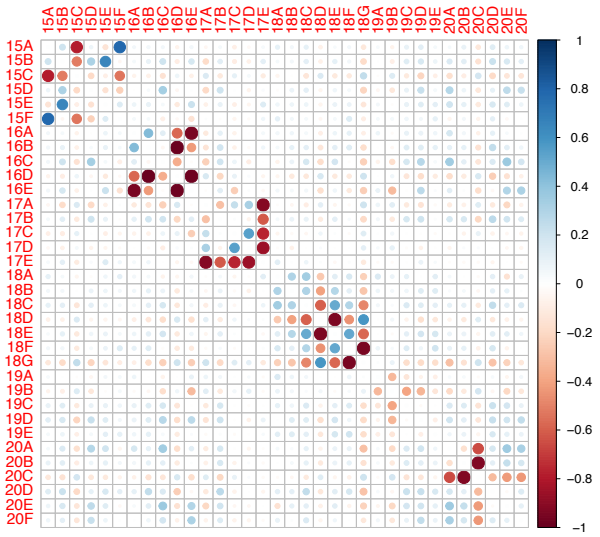


FIG. 1. Polychoric correlations between responses to MCMR items.

All analyses were performed using the R computing environment [20]. Polychoric correlations were computed using the POLYCOR package [21], and IRT analyses were performed using the MIRT package [22, 23]. To get a sense of the variability in the values of the IRT fit statistics, we used the MIRT function's option to generate random values for the initial parameter estimates (GenRandomPars = TRUE), repeating the estimation of each model 10 080 times for research question 2, and 20 064 times for research question 3. The median value of each statistic is reported, with uncertainty represented by the median absolute deviation of each distribution, scaled by a factor of 0.6745 to be comparable to the standard deviation [24].

III. RESULTS

A. Research Question 1: Correlations

Figure 1 shows the polychoric correlations between responses to MCMR items. The table is symmetric, with the rows and columns being labeled by the item and response option: e.g., "15A" represents response A to item 15. Figure 1 shows that only a small fraction of response pairs have correlations strong enough to be considered significant, all of which are within-item correlations (emphasized by their proximity to the diagonal). As such, we only considered within-item response pairs when building IRT models to address research questions 2 and 3.

The strongest correlations tend to be negative and often correspond to contradictory statements: for example, response 18D indicates that, "The force . . . is in the opposite

direction as . . . displacement," while 18E states, "The force . . . is in the same direction as . . . displacement." Moreover, only a minority of within-item response pairs display significant correlations, with item 19 having no significantly correlated pairs of responses.

B. Research Question 2: Grouping Responses

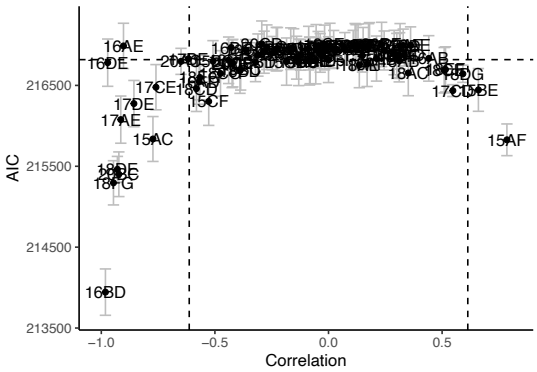
As mentioned above, we expect to see a relationship between the polychoric correlation between two response options, and the goodness-of-fit of an IRT model that includes those two responses as a single nominal response-item. The "All2pl" model represents the baseline, with all 34 MCMR response options analyzed as independent dichotomous response-items (using the 2PL model). Figure 2 shows the relationship between correlation and goodness-of-fit for two IRT fit statistics: AIC and CFI. In both cases, the value of the fit statistic for the All2pl baseline model is shown by a horizontal black dashed line, and the thresholds for significant correlations (± 0.613) are shown by vertical black dashed lines. Data labels represent the responses that are paired in each model: for example, the model labeled "16BD" includes responses 16B and 16D as a single nominal response-item, but leaves all other MCMR responses as independent dichotomous response-items.

The AIC results in Fig. 2(a) show a clear trend: models that include pairs with small correlations have AIC values very near (and often above) the All2pl value; models with paired responses that have larger correlations tend to have lower (i.e., better) AIC values. The results for the BIC statistic (not shown) are virtually identical to the AIC. These results suggest that adding a nominal response-item formed from a random pair of within-item responses will not significantly improve the AIC or BIC fit statistics.

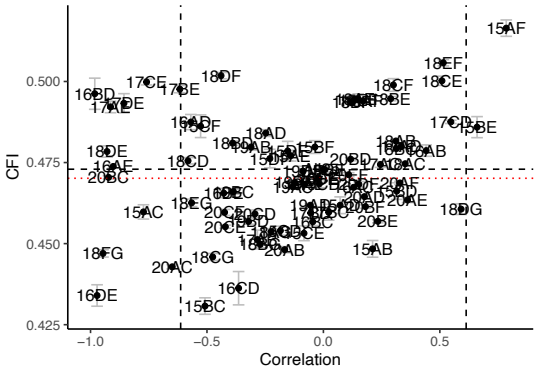
The CFI results shown in Fig. 2(b) are not as clear as those for the AIC. Some of the models with weakly correlated response-items make the model worse (lower CFI) and others make it better (higher CFI). Similarly, some models with strongly correlated response-items seem to make the fit worse. The red dotted line in Fig. 2(b) shows the median CFI value for all models with a only a single pair of responses grouped into a nominal option. This median is below the All2pl baseline, which suggests that adding a nominal response-item formed from a random pair of within-item responses is not likely to yield a significant positive effect. The results for the TLI and RMSEA statistics are similar to those for the CFI, with the median value being worse than the All2pl model, but without a distinct shape to the data.

C. Research Question 3: Building an Optimal Model

The strongest correlation is between 16B and 16D, and Fig. 2 shows that the 16BD model significantly improves both fit statistics. The next strongest correlations are between 16E



(a)



(b)

FIG. 2. Fit statistics vs. correlation for IRT models with one MCMR nominal response-item each. The black horizontal dashed line is the All2pl baseline value; the vertical dashed lines represent correlations of ± 0.613. (a) AIC is shown (BIC has the same shape). (b) CFI is shown (TLI and RMSEA have the same general result); the red dotted line is the median value for all models. Error bars represent the median absolute deviation for 10 080 runs of each IRT analysis.

and other responses to item 16: 16E states that, "None of these are correct," (referring to responses A–D) so it is not surprising that it would be anticorrelated with the others. Figure 3 shows the impact of incorporating the correlations involving 16E to the 16BD IRT model: the model labeled "16noE" in Fig. 3 removes 16E from the analysis entirely, with choosing 16E considered equivalent to NOT choosing 16A, B, C, or D. Each subsequent model (moving to the right on each plot in Fig. 3) includes the groupings of the previous models. (The 17noE model is similar to 16noE, with the none-of-the-above response 17E being removed from the analysis.) The final model, labeled 17CD in Fig. 3, includes seven nominal response-items and 17 dichotomous response-items, as opposed to the 34 dichotomous response-items included in the All2pl model.

With each added pair/group, the AIC and BIC statistics improve. The RMSEA, CFI, and TLI statistics improve for all but two of the models shown (16noE and 20ABC). Adding additional nominal response-items to model 17CD resulted in a worse fit for all statistics. For models with mixed results

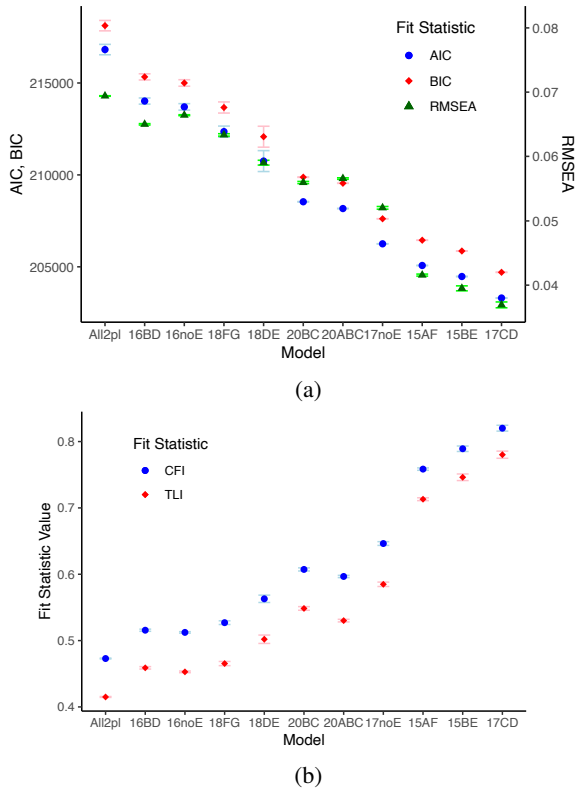


FIG. 3. Fit statistics for IRT models with the most highly correlated responses to MCMR items grouped. (a) AIC, BIC, RMSEA (lower is better). (b) CFI, TLI (higher is better). Error bars represent the median absolute deviation for 20 064 runs of the IRT analysis.

(better for some statistics but worse for others), we made a judgement by looking at the specific response options. The 16noE model fit may be better or worse than 16BD; as mentioned above, response 16E is the "None of these are correct" option. Combining three responses for 20ABC may be a better or worse fit than combining only two responses for 20BC; 20A, B, and C all deal with an object's speed, while 20D, E, and F deal with the direction of its velocity. Due to the relationships between the response options within items 16 and 20, we chose to keep both the 16noE and the 20ABC models. Model 15ACF (not included in Fig. 3) also showed mixed results; we rejected the model because 15A and 15F are mathematically equivalent (incorrect) equations, but 15C is a different (correct) equation. We only considered this type of choice when the trends in the fit statistics were mixed.

As mentioned in Sec. I, using the NRM for all MCMR response patterns would require a much larger data set, and could yield unstable results. To explore the possible usefulness of an IRT model using the NRM for all MCMR items, we created a data set that included only response patterns that had been selected by at least 20 students (the minimum sample size for each response category [12]). This left us with 128 response categories spread over the six MCMR items. This is significantly fewer than the 346 total possible combinations, but still required a sample size of at least 4000 students to estimate a total 400 parameters across all SR and MCMR items. This was larger than our available data set, and left us needing to select a larger minimum value. A minimum of 60 selections, for example, gave a model with 60 MCMR response categories that could work with our data (a minimum sample size of 2640 students), but the selection of this minimum value is quite arbitrary. Additionally, eliminating response patterns based on their frequency in the data set potentially removes meaningful patterns that are conceptually worth keeping, even if they are not popular. We do not believe these results would be representative of our data set.

IV. DISCUSSION

The results from Fig. 3 show that IRT models that include the most strongly correlated MCMR response pairs as nominal response-items more closely fit our data set than the baseline All2pl model in which all MCMR response options are included as independent dichotomous response-items. The results from Fig. 2 show that these improvements to the fit statistics are well beyond what would be expected from substituting any pair of dichotomous response-items with a single nominal response-item. Together, these results demonstrate the efficacy of our approach to using IRT to analyze data from MCMR items: using polychoric correlations between response options to identify pairs and groups to combine into nominal response-items for IRT analyses. This novel approach generalizes IRT analyses of nominal data to be able to incorporate all possible patterns of response options to MCMR items.

More work is needed before this method may be used to make decisions about scoring students' responses to MCMR items. Careful choices would have to be made to translate IRT results into definitions for assigning partial credit, and the analyses would have to be replicated and shown to be generalizable across multiple student populations. The students in our data set do not (and cannot) represent the overall physics student population in the USA in terms of race and ethnicity. In addition, these students (in general) have had more exposure to prior instruction in both physics and mathematics than many introductory physics students [25]. As such, our data sources are not broad or diverse enough to make strong conclusions about rankings of response options or determining partial credit models; however, this work serves as a proof-of-concept for using IRT methods in novel ways to analyze data from MCMR items.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grants No. DUE-1832836, No. DUE-1832880, and No DUE-1833050. One author holds an NRC Research Associateship award at Air Force Research Laboratory.

[1] S. White Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics, Phys. Rev. Phys. Educ. Res. **17**, 020129 (2021).

[2] T. I. Smith, P. Eaton, S. White Brahmia, A. Olsho, A. Boudreaux, and C. Zimmerman, Toward a valid instrument for measuring physics quantitative literacy, in *Physics Education Research Conference 2020*, PER Conference, edited by S. Wolf, M. Bennett, and B. Frank (Virtual Conference, 2020) pp. 496–502.

[3] T. I. Smith, P. Eaton, S. W. Brahmia, A. Olsho, A. Boudreaux, C. DePalma, V. LaSasso, S. Straguzzi, and C. Whitener, Using psychometric tools as a window into students' quantitative reasoning in introductory physics, in *Physics Education Research Conference 2019*, PER Conference, edited by Y. Cao, S. Wolf, and M. Bennett (Provo, UT, 2019) pp. 560–566.

[4] T. I. Smith, S. White Brahmia, A. Olsho, and A. Boudreaux, Developing a reasoning inventory for measuring physics quantitative literacy, in *Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education.*, edited by A. Weinberg, D. Moore-Russo, H. Soto, and M. Wawro (Oklahoma City, OK, 2019) pp. 1181–1182.

[5] T. I. Smith, K. J. Louis, B. J. Ricci, and N. Bendjilali, Quantitatively ranking incorrect responses to multiple-choice questions using item response theory, Phys. Rev. Phys. Educ. Res. **16**, 010107 (2020).

[6] P. Eaton, K. Johnson, and S. Willoughby, Generating a growth-oriented partial credit grading model for the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **15**, 20151 (2019).

[7] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, Am. J. Phys. **66**, 338 (1998).

[8] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The Physics Teacher **30**, 141 (1992).

[9] R. D. Bock, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, Psychometrika **37**, 29 (1972).

[10] T. I. Smith and N. Bendjilali, Motivations for using the item response theory nominal response model to rank responses to multiple-choice items, Phys. Rev. Phys. Educ. Res. **18**, 10133 (2022).

[11] Y. Suh and D. M. Bolt, Nested Logit Models for Multiple-Choice Item Response Data, Psychometrika **75**, 454 (2010).

[12] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (The Guilford Press, 2008).

[13] R. D. Bock and R. D. Gibbons, *Item Response Theory* (Wiley, Hoboken, NJ, 2021).

[14] S. White Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, PhysPort Assessments: Physics Inventory of Quantitative Literacy (2021).

[15] K. D. Kubinger, On artificial results due to using factor analysis for dichotomous variables, Psychology science **45**, 106 (2003).

[16] P. Eaton, B. Frank, and S. Willoughby, Detecting the influence of item chaining on student responses to the Force Concept Inventory and the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **16**, 20122 (2020).

[17] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 10137 (2018).

[18] J. Yang, C. Zabriskie, and J. Stewart, Multidimensional item response theory and the force and motion conceptual evaluation, Phys. Rev. Phys. Educ. Res. **15**, 20141 (2019).

[19] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 10124 (2018).

[20] R Core Team, R: A Language and Environment for Statistical Computing (2020).

[21] J. Fox, polycor: Polychoric and Polyserial Correlations (2019).

[22] R. P. Chalmers, Multidimensional Item Response Theory (mirt) (2022).

[23] R. P. Chalmers, mirt: A Multidimensional Item Response Theory Package for the R Environment, Journal of Statistical Software **48**, 1 (2012).

[24] R. R. Wilcox, *Introduction to robust estimation and hypothesis testing*, 3rd ed. (Academic Press, Boston, 2012) p. 75.

[25] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 20106 (2020).