Methods for utilizing Item response theory with Coupled, Multiple-Response assessments

Bethany R. Wilcox,[1] Katherine Rainey,[1] and Michael Vignal[1]

[1]*Department of Physics, University of Colorado, 390 UCB, Boulder, CO 80309*

Recent years have seen a movement within the research-based assessment development community towards item formats that go beyond simple multiple-choice formats. Some have moved towards free-response questions, particularly at the upper-division level; however, free-response items have the constraint that they must be scored by hand. To avoid this limitation, some assessment developers have moved toward formats that maintain the closed-response format, while still providing more nuanced insight into student reasoning. One such format is known as coupled, multiple response (CMR). This format pairs multiple-choice and multiple-response formats to allow students to both commit to an answer in addition to selecting options that correspond with their reasoning. In addition to being machine-scorable, this format allows for more nuanced scoring than simple right or wrong. However, such nuanced scoring presents a potential challenge with respect to utilizing certain testing theories to construct validity arguments for the assessment. In particular, Item Response Theory (IRT) models often assume dichotomously scored items. While polytomous IRT models do exist, each brings with it certain constraints and limitations. Here, we will explore multiple IRT models and scoring schema using data from an existing CMR test, with the goal of providing guidance and insight for possible methods for simultaneously leveraging the affordances of both the CMR format and IRT models in the context of constructing validity arguments for research-based assessments.

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

I. INTRODUCTION & BACKGROUND

Validated, research-based assessments are a ubiquitous tool in physics education research (PER) because they provide insight into student understanding of physics concepts. These assessments are often used to investigate the efficacy of instructional strategies, which can inform changes to instruction to improve students learning [1]. Validated assessment instruments have been developed for nearly all core physics content areas [2]. These instruments take a variety of formats including free-response (e.g., Ref. [3]), multiple-choice (e.g., Ref. [4]), and multiple-response (e.g., Ref. [5]). The format of an assessment has a number of important implications both for its use and validation. For example, free-response formats provide important insight into students' reasoning, but are time and resource intensive to score. Closed-response formats allow for automated scoring, but provide limited insight into students' reasoning.

An alternative format that maintains the automated scoring of a closed-response format while still providing some insight into students' reasoning is known as coupled, multiple-response (CMR) [6]. The CMR format is characterized by a multiple-choice question followed by a multiple-response followup that prompts students to select reasoning elements that support their answer to the multiple-choice question. In addition to providing information about students' reasoning, the CMR format also allows for flexibility in terms of scoring. For example, students can be scored based on both the correctness and consistency of their answers. As such, CMR questions are often scored polytomously.

Polytomous scoring, however, brings with it some constrains with respect to assessment validation. For example, many test theories assume dichotomous scoring schemes [7]. Most common test statistics do have polytomous versions, but these versions sometimes come with tradeoffs in terms of statistical power. One place where this tradeoff may be particularly important is with respect to Item Response Theory (IRT). IRT is a is a model-based approach to estimating item and test parameters which posits that a students' response to a particular item on an assessment should depend only on the difficulty of that item and the students' underlying latent "ability" [8]. A major advantage of IRT stems from the fact that, if you can craft items for which students' responses match the IRT item response function, the resulting estimates of students' latent ability and item difficulty will be independent [8]. In other words, statistics indicating how difficult the items are (individually or in aggregate) do not shift when the assessment is taken by student populations with a differing distribution of abilities, providing an avenue for producing test statistics that are independent of the test population. While there are many IRT models, we will focus here on the subset of IRT that is known as Rasch analysis. As discussed later, this is motivated in part to limit the impact of small N.

Both Rasch analysis and the CMR item format have affordances and (sometimes conflicting) constraints with respect to the development and validation of research-based assessments. In this paper, we explore methods for applying Rasch analysis to data from a CMR assessment called the Upper-level Statistical Mechanics and Thermodynamics Evaluation for Physics (U-STEP) [9]. The U-STEP is a newly developed assessment targeting both classical thermodynamics and statistical mechanics at the upper-division undergraduate level. The assessment features primarily CMR items that are scored polytomously based on both correctness and consistency of students responses. The goal here is **not** to establish a validity argument for the U-STEP using IRT. Rather, this analysis is designed to explore different Rasch models and identify different methods for mapping the scoring of the U-STEP onto these models with the goal of identifying promising strategies for leveraging the affordances of both Rasch analysis and the CMR format for future assessments.

II. RASCH ANALYSIS

In this section, we provide a brief primer on some of the important aspects of Rasch analysis. The mathematical formalism of the Rasch model (also known as the one-parameter IRT model) assumes the probability of a student answering an item correctly is determined by a latent trait of the student and item difficulty. The Rasch model posits the following model for the probability that a particular student will respond to a particular item correctly:

$$P(\theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}} \tag{1}$$

Here, θ_j represents the latent trait of interest and has historically been referred to as the "ability"[1] of student j and b_i is the difficulty of item i. There are several important assumption implicit in the Rasch model including the assumption that all item discriminations are equal [10]. Additionally, the base Rasch model assumes dichotomously scored items, though there are polytomous Rasch models [11]. Finally, Rasch analysis assumes a unidimensional test.

Rasch analysis requires that student responses match the model in Eqn. 1, and as with IRT models the fit of the Rasch model to the data cannot be assumed, but instead must be evaluated. A variety of fit statistics can provide statistical evidence for model fit and identify mis-fitting items. One benefit of the Rasch model over other IRT models with more parameters is that it requires smaller sample to establish model fit and provide reliable parameter estimates. Though model fit depends on both sample size and test length, some have suggested a sample size of at least 200 respondents and 15 items (the same number of items on the U-STEP) provides sufficient fit statistics [12].

[1] Note, "ability" refers to the latent trait that the statistical models quantify. Fundamentally, however, it is a measure of performance as opposed to innate ability. This term is used for consistency with the existing literature. However, this term is potentially problematic, particularly with respect to the interpretation of performance gaps between subgroups of students.

TABLE I. Information about the two semesters of pilot administrations of the U-STEP used in this study. In all cases, N here refers to the number of students enrolled rather than the number of responses. Note the average of institutional response rates does not include classes with 0% response rate (N=1 for Spring and Fall 2020).

		$N_{students}$ per class			$N_{students}$	Response Rate	
	$N_{institutions}$	avg.	min.	max.	total	overall	avg.
Fall 2020	10	23	8	86	227	75%	73%
Spring 2021	18	19	3	110	349	82%	91%

III. CONTEXT & METHODS

Data from this study come from two semesters of U-STEP data collection in upper-division thermal physics courses. As mentioned previously, the U-STEP has 15 items, most of which are CMR items. The development and validation of the U-STEP using Classical Test Theory are reported in Refs. [9, 13, 14]. An exploratory factor analysis of data from the U-STEP indicated that the U-STEP's factor structure is dominated by a single factor, and thus, the instrument appears statistically unidimensional.

For this study, we utilize data from two semesters of the pilot data used to establish a validity argument for the U-STEP, which are summarized in Table I. Between the two semesters, we collected a total of $N = 455$ students responded to the U-STEP from 27 distinct institutions. In soliciting pilot sites, we placed emphasis on participation from multiple institution types serving various student populations. This provided us with a more representative sample of students. For example, our Fall 2020 sample was 20% women and 59% White, while the Spring 2021 sample was 25% women and only 50% White. This is a comparable representation of women in our sample compared to representation in the field of physics; it is also a lower representation of White students than that in physics more generally (and a closer reflection of the representation of non-Hispanic White people in the US) [15].

Students' responses were filtered to identify any responses that could be considered invalid for some reason. For example, all responses that took less than 5 min from start to finish were dropped because 5 minutes is roughly the amount of time necessary to skim through the exam in full - suggesting that these students did not take the diagnostic seriously. Additionally, any students who skipped or only partially responded to 5 or more of the items on the U-STEP were also dropped from the data set. After these drops, the total number of responses retained for each semester was $N = 164$ for Fall 2020, and $N = 277$ for Spring 2021, well exceeding the threshold required to achieve reliable fit statistics.

In all courses, the U-STEP was taken online and typically outside of normal class time and without use of notes or online resources. For example, many instructors included the completion of the U-STEP as a portion of their final homework assignment for the semester and assigned participation credit to the students as a portion of the homework assignment. Additionally, all semesters of data collection with the assessment overlapped with the COVID-19 pandemic. Thus, many courses were taught at least partially remotely or online during all or part of those semesters. This had little impact on the administration of the assessment, as the intention for the U-STEP was always for it to be administered in an online format. However, the disruption caused by the pandemic likely had impacts on who was enrolled in physics courses as well as who completed the assessment during this time period. This, along with the hybrid or remote-nature of courses used for piloting, may have had impacts on the data collection that are difficult to quantify.

The basic scoring of the U-STEP was designed to value both consistency and correctness; thus, it is possible on some items for a student to answer the MC prompt incorrectly and still receive some credit if reasoning is consistent with their incorrect response. While the structure varies between items, typically the multiple-choice portion of the CMR item is worth 2 points and the reasoning portion is worth 3 points, and different items may have multiple reasoning followups. After summing the total points for the CMR item, the score is then normalized to 1 point per item resulting in a maximum total score on the U-STEP of 15 points. The full scoring structure for each item on the U-STEP can be accessed at Ref. [9]. In the next section we explore several options for utilizing or modifying the base U-STEP scoring so that it maps onto various Rasch models, both dichotomous and polytomous. All analyses described below were conducted in R using a variety of functions included in the mirt package.

IV. RESULTS & DISCUSSION

In this section, we will first explore options for adapting the base scoring of the U-STEP into a dichotomous scheme. Though polytomous Rasch models exist, they require much larger sample sizes than dichotomous Rasch analyses; smaller sample size restrictions are beneficial for upper-division courses, which have a relatively small pool of students to draw respondents from. Following this dichotomous analysis, we will also explore polytomous Rasch models.

A. Dichotomous Rasch analysis

We considered 2 possible methods for adapting the polytomous scoring of the U-STEP to a dichotomous scheme. The most obvious approach might be to require perfectly correct responses to receive credit for the item. This method would award 1 point for a students who select only the correct multiple-choice option and all of the correct/necessary reasoning element on the multiple-response portion of the CMR question with none of the incorrect/unnecessary options, and 0 points for any other response pattern. However, with CMR items, getting a perfectly correct response is more challenging than in the case of multiple-choice items since so many permutations of responses are possible. Indeed, in the case

of one of the U-STEP items, no student in our sample gave a perfectly correct response; two additional items had only 5-15 students give perfectly correct responses. Such a small number of correct responses has important implications for the potential discrimination of the items and thus increases the likelihood of model misfit due to the assumption of equal discrimination. Additionally, item statistics from Classical Test Theory (e.g., item difficulty and discrimination) were very low, with only a handful of items passing the traditional thresholds for these parameters [7]. Because of this, we do not further pursue this dichotomization method for the U-STEP. We suspect other CMR assessments may run into similar difficulties with respect to this scheme.

The second method for dichotomization that we explore is to set a particular threshold on the score a student must get on the item to be considered "correct." Here, we test three different thresholds - 40%, 50%, and 60% - where any student with a score above the threshold is given 1 point on that item and any student below is given 0 points. A 40% threshold implicitly values correctness over reasoning because for most items, selecting the correct response to the multiple-choice portion is enough to get you 40% of the available point for that item. A 60% threshold, on the other hand, requires at least some correct reasoning for the majority of items, and thus implicitly places more emphasis on correct justifications in addition to correct answers. Correlations between the full polytomous scoring and the 40%, 50%, and 60% thresholds are 0.96, 0.97, and 0.96 respectively.

To quantify overall model fit we used an M_2 statistical test. This test was proposed by Maydeu-Olivares and Joe [16] and provides a measure of model fit for the overall assessment. Acceptable model fit is critical to ensuring the reliability of model outputs, such as item difficulty and student ability. Large M_2 values correspond to lower p-values and more model misfit. Additionally, root mean square error of approximation (RMSEA) and comparative fit index (CFI) provide another measure of overall fit. Conservative discussions suggest an RMSEA value less than 0.05 indicates a close fit [17, 18] and CIF values greater than 0.95 indicate relatively good model fit [19].

Table II presents fit statistics for each of these three thresholds. All three thresholds result in contradictory conclusions as to model fit with all indicating misfit according to the M_2 statistic but reasonable fit according to RMSEA and CIF. Item fit statistics identified 3 misfitting items for the 40% threshold and 2 for each of the 50% and 60% thresholds ($S-\chi^2$, $p<0.05$). For the 40% threshold, removal of the 3 misfitting items resulted in no further misfitting items, but did not meaningfully change overall fit statistics. For the 50% threshold, removal of the 2 misfitting items resulted in an improvement in overall fit statistics (though M_2 still detected overall misfit), but also resulted in another item with significant misfit. Further removal of this item resulted in no significant item or overall misfit suggesting that the 50% threshold matched the Rasch model upon the removal of the 3 items with significant model misfit. For the 60% threshold, removal of the 2 misfitting items resulted in acceptable overall model fit.

TABLE II. Overall assessment fit statistics using three thresholds for dichotomizing data. Significant misfit is indicated by p<0.05. Root mean square error of approximation (RMSEA) less than 0.05 indicates a relatively good model fit; comparative fit index (CFI) values greater than 0.95 indicate good model fit [20].

	M_2	p	RMSEA	CFI
40% Threshold	155	<0.001	0.033	0.968
50% Threshold	159	<0.001	0.035	0.964
60% Threshold	149	0.003	0.031	0.971

B. Polytomous Rasch analysis

Another approach to analysing data from CMR items with Rasch analysis is to simply utilize one of the polytomous Rasch models. Here, we explore the Partial Credit Model (PCM), which as its name implies was born from traditional IRT to account for *partial credit*. The PCM in particular (as opposed to the *generalized* partial credit model) is a variant of the Rasch model. While the use of a polytomous model like the PCM has the benefit of preserving the full scoring scheme of the CMR format (or something close), analysis with polytomous models typically require larger datasets to achieve reliable parameter estimates and fit statistics. For example, literature suggests that use of the PCM model requires very large sample sizes (e.g., close to 1,000 respondents) [21]. The dataset for the U-STEP is not even half this size, suggesting that this approach is not appropriate for the U-STEP. However, with this caveat, in this section, we will still present analysis of the U-STEP data set using the PCM in order to demonstrate the process, highlight some of the challenges, and explore potential options to overcome them.

The PCM was developed by Masters in 1982 [22]. For the PCM, one assumes the maximum possible score can be achieved by going through steps of other score "categories" defined by the amount of partial credit. In the example provided in his 1982 paper, there is a math problem with 3 parts and thus four possible scores achievable: 0, 1, 2, and 3. These are the score categories, and from them, there are 3 steps that can be taken to achieve the maximum score: going from score $0 \rightarrow 1$, from score $1 \rightarrow 2$, and from score $2 \rightarrow 3$. For each step in this progression, there is a difficulty associated with the step, which is not necessarily the same for each step. For example, going from $0 \rightarrow 1$ may be easier than going from $2 \rightarrow 3$, resulting in different difficulty values. Essentially, the PCM treats each "step" as a dichotomous item, which can either be correct and progress a score to the next step (e.g., $1 \rightarrow 2$), or incorrect and leave a score at the latter step (e.g., remain at score of 1). Each of these steps is then modeled using the Rasch model (i.e., Eqn. 1), producing a difficulty value for each step. It is worth noting that the PCM may operate on the assumption that scores are ordinal, and thus build

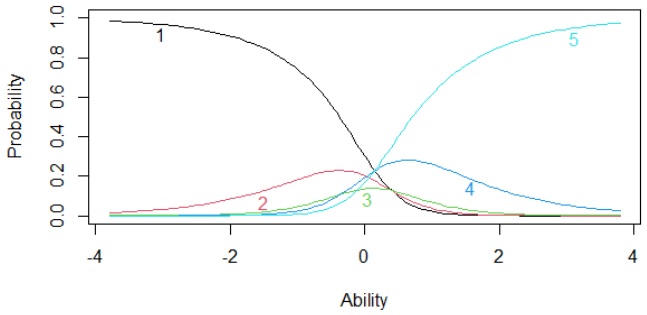

FIG. 1. PCM item response category characteristic curves for an example U-STEP item. Each curve is labeled with its associated score, ranging from 1 to 5, and its shape is described by Eqn. 1.

on each other. However, this is not the case for all U-STEP items, especially those composed of multiple CMR pairs.

Despite this possible concern, exploratory PCM analyses were conducted for the 15 U-STEP items using the ltm package in R. This analysis yielded complicated results in many instances, due to the many number of score categories within different items. For example, item 11 on the U-STEP, which was composed of 3 CMR pairs and allowed for half-integer points for some reasoning selections, had 27 possible score categories. Thus, this item had 26 difficulty values, one for each score step. PCM analysis of the raw U-STEP scores resulted in fit statistics that were consistent with acceptable overall model fit; however, given the small dataset, these statistics should not be interpreted as robust.

To provide an example of the output of PCM analysis, Fig. 1 provides the item characteristic curves for one item on the U-STEP. This particular example had 5 steps due to its 6 possible score categories: 0, 1, 2, 3, 4, and 5. Each curve represents the probability of a student with a particular ability parameter receiving the associated score of 1, 2, 3, 4, or 5. Notably, the curve for achieving a score of 5 closely matches the expected pattern described by Eqn. 1, with higher-ability students have a higher probability of receiving the highest possible score. Similarly, the curve for a score of 1 takes a similar shape with opposite directionality; this indicates that lower-ability students have a higher probability of only making the lowest-level step (0→1). The curves representing scores of 2, 3, and 4 have peaks that gradually shift towards higher abilities, indicating that students need higher ability levels in order to achieve the progressively increasing scores, suggesting some conceptual alignment of responses to the model.

One of the primary challenges to using PCM for CMR items is, thus, concerns about having too many score categories to either provide meaningful item parameters or ability estimates. One possible strategy to address this is to reduce the number of score categories by binning students raw CMR scores into smaller steps. To investigate the impact of reducing the number of score categories for PCM analysis, we revised the item scoring such that possible scores for each item could take on only one of three possible values: 0, 1, and 2. The traditional (5pt) CMR items were scored such that 1 point was awarded for a correct MC selection and 1 point was awarded for a minimum of half-correct reasoning and 2 points for fully correct reasoning. Scoring for paired CMR items was more nuanced and varied by item but still produced at most 3 score categories. Analysis of model fit showed statistically significant model misfit (χ^2, $p = 0.1$). Again, the validity of these statistics for the U-STEP in particular are questionable due to the smaller N of the current U-STEP dataset.

V. CONCLUSIONS & LIMITATIONS

In this paper, we explored methods for utilizing Rasch analysis to analyze assessment using the CMR format, which feature polytomous item scores. We explored two different methods for structuring the analysis - dichotmization of the scores and use of polytomous models - using example data from an upper-division thermal physics assessment.

Our results suggest that dichotomization using a threshold value may be a promising approach for producing valid ability and item difficulty parameters using Rasch analysis. For the U-STEP, in particular, the 60% threshold appears the most robust, with this threshold requiring the fewest dropped items to achieve model fit. Dichotomization based on the full polytomous scoring has the advantage of requiring smaller data sets to produce valid parameters and fit statistics, and therefor would be useful for assessments targeting smaller student populations. Polytomous Rasch models may also be a promising approach for assessments that have datasets of sufficient size. However, depending on the fine grained nature of the assessment's CMR scoring scheme, it may be necessary to revisit the scoring scheme or bin students scores in order to maximize the number of students in each scoring category and produce more actionable student ability estimates.

This work has important limitations. The U-STEP, was not designed with IRT analysis in mind; thus, there is no *a priori* reason to assume students' responses should match the Rasch model. Additionally, the smaller data set of responses to the U-STEP, while large considering it targets upper-division physics population, is still small relative to what is necessary for polytomous Rasch models. Thus, the findings reported here should be interpreted as exploratory and not as evidence that any particular model is or is not valid for analysis of data from CMR assessments. This analysis serves as a foundation for assessment developers interested in leveraging the affordances of both IRT and the CMR format.

ACKNOWLEDGMENTS

This work was supported by funding from the Center for STEM Learning and the Department of Physics at University of Colorado Boudler, and the National Science Foundation DUE Grant No. 2013332.

[1] Adrian Madsen, Sarah B McKagan, and Eleanor C Sayre, "Resource letter rbai-1: research-based assessment instruments in physics and astronomy," American Journal of Physics **85**, 245–264 (2017).

[2] https://www.physport.org/assessments/, (2015).

[3] Marcos D. Caballero, Leanne Doughty, Anna M. Turnbull, Rachel E. Pepper, and Steven J. Pollock, "Assessing learning outcomes in middle-division classical mechanics: The colorado classical mechanics and math methods instrument," Phys. Rev. Phys. Educ. Res. **13**, 010118 (2017).

[4] Homeyra R. Sadaghiani and Steven J. Pollock, "Quantum mechanics concept assessment: Development and validation study," Phys. Rev. ST Phys. Educ. Res. **11**, 010110 (2015).

[5] Suzanne White Brahmia, Alexis Olsho, Trevor I. Smith, Andrew Boudreaux, Philip Eaton, and Charlotte Zimmerman, "Physics inventory of quantitative literacy: A tool for assessing mathematical reasoning in introductory physics," Phys. Rev. Phys. Educ. Res. **17**, 020129 (2021).

[6] Bethany R. Wilcox and Steven J. Pollock, "Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics," Phys. Rev. ST Phys. Educ. Res **10**, 020124 (2014).

[7] Paula Engelhardt, "An introduction to classical test theory as applied to conceptual multiple-choice tests," in *Getting Started in PER*, Vol. 2 (2009).

[8] Frances M Yang *et al.*, "Item response theory for measurement validity," Shanghai Archives of Psychiatry **26**, 171 (2014).

[9] Katherine Rainey, *Upper-Division Thermal Physics Assessment Development and the Impacts of Race & Gender on STEM Participation*, Dissertation, University of Colorado Boulder (2021).

[10] Georg Rasch, *Probabilistic models for some intelligence and attainment tests.* (ERIC, 1993).

[11] Michael L Nering and Remo Ostini, *Handbook of polytomous item response theory models* (Taylor & Francis, 2011).

[12] Michael R Harwell and Janine E Janosky, "An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in bilog," Applied psychological measurement **15**, 279–291 (1991).

[13] Katherine D. Rainey, Michael Vignal, and Bethany R. Wilcox, "Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage," Phys. Rev. Phys. Educ. Res. **16**, 020113 (2020).

[14] Katherine D. Rainey, Michael Vignal, and Bethany R. Wilcox, "Validation of a coupled, multiple response assessment for upper-division thermal physics," Under Review: Phys. Rev. Phys. Educ. Res..

[15] National Science Foundation, "Women, minorities, and persons with disabilities in science and engineering," (2017).

[16] Albert Maydeu-Olivares and Harry Joe, "Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework," Journal of the American Statistical Association **100**, 1009–1020 (2005).

[17] Michael W Browne, Robert Cudeck, Kenneth A Bollen, and J Scott Long, "Testing structural equation models," (1993).

[18] Karl G Jöreskog and Dag Sörbom, *LISREL 8: Structural equation modeling with the SIMPLIS command language* (Scientific Software International, 1993).

[19] Li-tze Hu and Peter M Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," Structural equation modeling: a multidisciplinary journal **6**, 1–55 (1999).

[20] Yan Xia and Yanyun Yang, "Rmsea, cfi, and tli in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods," Behavior research methods **51**, 409–428 (2019).

[21] Tanja Kutscher, Michael Eid, and Claudia Crayen, "Sample size requirements for applying mixed polytomous item response models: results of a monte carlo simulation study," Frontiers in psychology **10**, 2494 (2019).

[22] Geoff N Masters, "Partial credit model," in *Handbook of Item Response Theory, Volume One* (Chapman and Hall/CRC, 2016) pp. 137–154.