Assessing the assessment: Mutual information between response choices and factor scores

Cole Walsh and N.G. Holmes
Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY 14853, cjw295@cornell.edu
(Dated: September 30, 2019)

Validated formative assessment tools provide a reliable way to compare student learning across variables such as pedagogy and curricula, or demographics. Such assessments typically employ a closed-response format developed from student responses to open-response questions and interviews with students and experts. The validity and reliability of these assessments is usually evaluated using statistical tools such as classical test theory or item response theory. The suitability of individual questions on an assessment can be examined using either of these methods, but so far little attention has been given to evaluating the suitability of individual response choices available in each question. Here, we use mutual information, a tool rarely used in PER, to quantitatively evaluate the utility of response choices in an assessment. We use the Physics Lab Inventory of Critical thinking (PLIC) as an example to illustrate how assessment developers can use this method for their own assessments. The PLIC was designed to measure three latent constructs and we confirm this structure through a factor analysis. We calculate factor scores that represent performance on each of the proposed constructs and evaluate the suitability of individual response choices in terms of how much information they provide about these factor scores.

2019 PERC Proceedings edited by Cao, Wolf, and Bennett; Peer-reviewed, doi.org/10.1119/perc.2019.pr.Walsh_C

I. INTRODUCTION

Closed-response formative assessments are typically developed from interviews with students and responses to open-response questions [1]. The validity and reliability of these assessments are usually evaluated using interviews with students and experts, and statistical tools such as classical test theory (CTT) or item response theory (IRT) [2]. These statistical tools evaluate the validity of an instrument, in the case of CTT, or individual questions, in the case of IRT; they do not generally evaluate the suitability of individual response choices. Generalizations of IRT do exist for examining the quality of response choices [3, 4], however one needs to make a number of assumptions about the nature of the questions and the data, which may or may not be valid for all assessments.

As PER trends toward new and innovative assessments [5–7], researchers are seeking new validation methods. Brewe *et al.*, for example, developed a novel method for examining response choices by employing network analysis to find communities of non-normative responses [8]. We build on this work by introducing a method to evaluate response choices that does not require any of the assumptions of IRT including any presupposed relationship between item discrimination, item difficulty, and latent student ability.

We use a concept from information theory, mutual information [9], to examine how much information response choices provide about the underlying factors of an assessment. In the rest of the paper, we introduce mutual information and describe how it can be used to evaluate response choices using the Physics Lab Inventory of Critical thinking (PLIC) [6] as an example. The PLIC uses a *multiple response* format, but the method described here can be used with any assessment. Here, we use mutual information in a way that is mainly directed towards assessment developers. However, mutual information is a tool that PER could use in myriad ways such as to answer questions like "how much information is gained about students' exam grades by knowing how many lectures they attended?" The secondary results of this paper describe a factor analysis of the PLIC that provides further validity evidence that the PLIC is measuring three distinct constructs of critical thinking in physics labs.

II. METHODS

A. Data Sources

The PLIC is a 10-question, closed-response assessment that probes students' critical thinking skills in the context of a physics lab. Respondents are presented with case studies of two groups completing an experiment to test the relationship between the period of oscillation of a mass on a spring and the spring constant, k, and mass, m. The first hypothetical group (Group 1) takes repeated measurements of the period of two different masses and finds the k-values for the two masses with uncertainties. The second hypothetical group (Group 2) takes two measurements of the period for many different masses and plots T^2 vs m expecting a straight line through the origin. When there is a disagreement between their prediction and experimental results, Group 2 adds an intercept to the model, which improves the quality of the fit, but raises questions about the assumptions of the model.

PLIC questions are presented in Table I. Questions Q1B, Q1D, and Q1E concern Group 1, Q2B, Q2D, Q2E, Q3B, Q3D, and Q3E concern Group 2, and Q4B asks respondents to compare the two groups. The assessment uses a *multiple response* format where students may select up to three response choices for each question from a pool of 6-17 and many response choices are repeated across similar questions. Questions are scored on an approximately continuous scale and bounded between 0 and 1. For a more detailed description of the PLIC including the development and validation process, scoring scheme, and example questions with response choices see [6].

We use data collected since March 2018 when data collection began with the most recent version of the PLIC. This dataset contains 7525 responses from students enrolled in 91 courses: 59 introductory (10 algebra-based and 49 calculus-based) and 32 beyond-first-year courses. There are also 39 institutions represented in this dataset: 2 two-year colleges, 15 four-year colleges, 2 master's granting institutions, and 19 Ph.D granting institutions.

B. Confirmatory Factor Analysis

We designed the PLIC to evaluate three critical thinking constructs that are important in physics experimentation: evaluate models, evaluate methods, and suggest follow-ups. Each of these constructs was predicted to be represented by at least three questions (see Table I). We identify this factor structure since we later evaluate the information that a response choice provides about a latent construct rather than the whole test or each question, which are more difficult to interpret.

We perform a confirmatory factor analysis (CFA) to evaluate this proposed factor structure using Maximum-likelihood (ML) estimation to extract the variances from the data. We use a model chi-square test to evaluate the null hypothesis that there is no difference between the hypothesized model and the observed relationships within the data. It has been argued elsewhere that this is an unrealistic hypothesis [10] since the chi-square test is sensitive to sample size; with larger statistical power the null hypothesis will be rejected even if the difference between the model and the data is small. In response, as recommended in [11], we also examine our model using the comparative fit index (CFI; relative fit index, good model ≥ 0.90), the standardized root-mean-square residual (SRMR; absolute fit index, good model < 0.08), and the root-mean-square error of approximation (RMSEA; parsimony-adjusted fit index, good fit < 0.08).

Factor scores (i.e., a score representing a student's ability in a latent construct as measured by the PLIC) are estimated using Thurstone's regression method [12] and discretized into equal frequency bins, where the number of bins is chosen according to Scott's criterion [13]. This

TABLE I: Multiple response questions from the PLIC with their predicted factor structures.

Predicted Factor	Question Code	Question Text
Evaluating Models	Q1B	What features were most important in comparing the two k-values?
	Q2B	What features were most important in comparing the fit to the data?
	Q3B	What features were most important in comparing the fit to the data? (after changing intercept)
	Q3D	Which items reflect your reasoning for determining which fit you think Group 2 should use?
Evaluating Methods	Q1D	What features of Group 1's method were most important for evaluating the method?
	Q2D	What features of Group 2's method were most important for evaluating the method?
	Q4B	What features were most important for comparing the two groups?
Suggesting Follow-ups	Q1E	What do you think Group 1 should do next?
	Q2E	What do you think Group 2 should do next?
	Q3E	What do you think Group 2 should do next? (after changing intercept)

discretization step is necessary as we use the form of mutual information for discrete random variables.

C. Mutual Information

The mutual information, $I(X;Y)$, between two random variables, X and Y, indicates how much information is gained about one variable by observing the other and is measured in dimensionless units (i.e., bits). $I(X;Y)$ can be interpreted as the reduction in entropy (uncertainty) in X after observing Y. If someone knew the distribution of X and wanted to guess the value of a particular x chosen at random from X by dividing the probability distribution in half with each guess: "is x greater than x_0?", then $I(X;Y)$ is the expected reduction in the number of yes/no guesses required after observing Y. Using the question posed in Sec. I as an example, suppose we knew the distribution of exam grades (variable X) for a class and wanted to guess what grade a particular student received (x). Without any further information, this may take many guesses, but if we also knew how many lectures the student attended (variable Y), then we would consider a smaller subset of possible exam grades, reducing the number of guesses required to guess the student's grade. We can also think of mutual information as a more general form of correlation; when two variables are independent both the mutual information and correlation coefficient between the variables is zero. The advantage of mutual information is that it measures general dependence and can be used to gauge the strength of relationships even when the relationship is non-linear [14].

We can evaluate response choices from the PLIC in terms of how much information they provide about the underlying factors (i.e., how much information does knowing whether or not a student selected a particular response choice provide about their score on a given factor?). The information gained about factor F by observing response choice R is:

$$I(F;R) = \sum_{f\in\mathcal{F}} \sum_{r\in\{0,1\}} p(r,f) \log_2 \frac{p(r,f)}{p(r)p(f)}, \qquad (1)$$

where $\mathcal{F}$ is the set of possible factor scores for factor F and r is a binary variable that is equal to 1, if the response choice R is selected, or 0, if it is not selected. $p(r)$ is the marginal probability of observing a value r for response choice R, $p(f)$ is the marginal probability of observing a score of f on factor F, and $p(r,f)$ is the joint probability of observing values r and f concurrently. These probabilities are calculated empirically from the dataset. In general, the upper bound on mutual information depends on the the number of possible values that the two variables can take on. In our case, where one of the variables is binary, the maximum mutual information between the two variables (i.e., a response choice and factor) is 1 bit.

In Sec. III B, we calculate the mutual information between response choices and their associated factors. Response choices that provide the most information are typically picked frequently and are either very expert-like or very novice-like. Expert-like and novice-like response choices have the greatest impact on an individual's score, so it is not surprising that they provide more information about students' scores. We draw attention to these response choices as a validity check of our method, but focus on response choices that provide little information from multiple questions and are prime candidates to be removed from future versions of the PLIC. We also note interesting patterns that arose during this analysis that may hint toward more complex relationships in the data.

III. RESULTS AND DISCUSSION

A. Confirmatory Factor Analysis

The CFA (see Fig. 1) demonstrates that the PLIC, as predicted, measures three distinct constructs for critical thinking in physics labs: evaluate models, evaluate methods, and suggest follow-ups. A chi-square goodness-of-fit test indicated that the expected results from the specified model were statistically different from those observed in the data ($\chi^2 = 385.3$, $df = 32$, $p < 0.001$). This disagreement is not surprising given the large sample size used ($N = 7525$) and may be due to only minor model misspecification. To check whether this was the case, we used additional model fit indices that are less sensitive to sample size: SRMR was 0.031, RMSEA was 0.038, and CFI was 0.904, which meets the criteria described in Sec. II B. These fit indices indicated that the specified model was supported by the data, though we note that other studies [11] have used more strict cutoffs of 0.95 for CFI, so there may be minor model misspecification that we explore in more detail now.

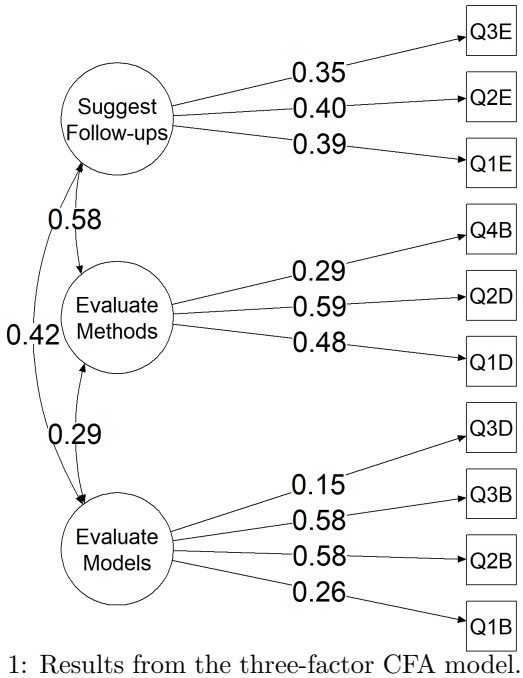
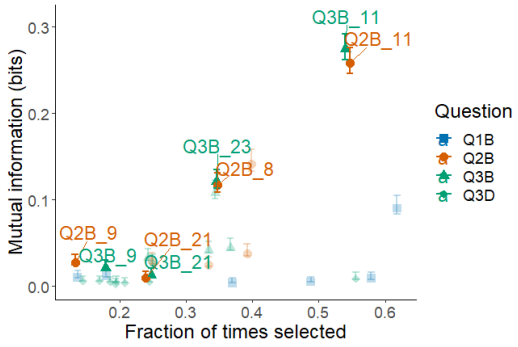


FIG. 1: Results from the three-factor CFA model. PLIC questions are represented by squares and factors are represented by circles. Double-headed arrows represent correlations between factors; single-headed arrows represent standardized factor loadings. $p < 0.001$ for all estimates.

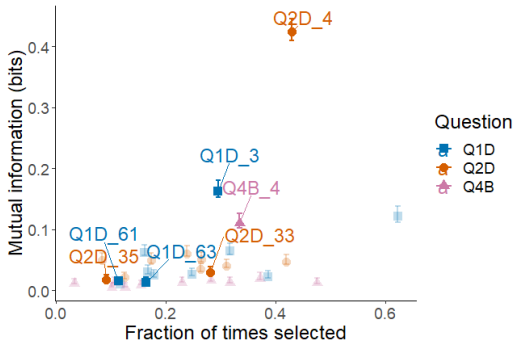
Questions Q1B, Q3D, and Q4B all have factor loadings less than 0.3, likely because these questions are slightly different than the other questions that are part of their predicted factors (see Table I). However, these three questions are generally more strongly correlated with the other questions in their predicted factors than to the other questions on the assessment. For instance, the Pearson's r correlation coefficient for Q4B with Q1D and Q2D is 0.114 and 0.201, respectively. Only one other question, Q1E, has a correlation stronger than $r = 0.039$ with Q4B. We argue then that, though theoretically more distant from the other questions, Q1B, Q3D, and Q4B should be included in the model since removing them would reduce the number of questions and these questions represent slightly different pieces of their respective factors. Future work may explore alternative factor structures. We do not, however, include Q1B, Q3D, and Q4B in our interpretation of response choices with low mutual information because these questions correlate less strongly with their respective factors and the response choices available for these questions are typically unique.

B. Mutual information between response choices and factor scores

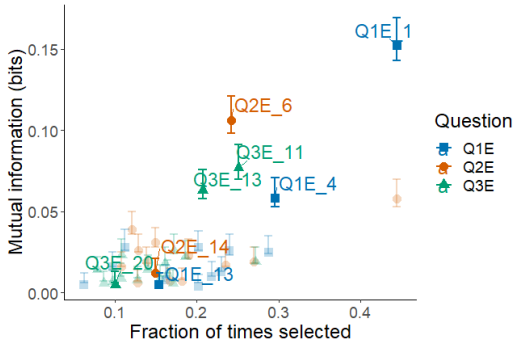
We now examine the mutual information between response choices and factor scores using only response choices associated with the factor (see Fig. 1). In Fig. 2, mutual information between a response choice and factor is plotted on the y-axis as a function of the fraction of respondents that selected the response choice. We chose to plot the fraction of respondents that selected a response



(a) *Evaluating models* factor.



(b) *Evaluating methods* factor.



(c) *Suggest follow-ups* factor.

FIG. 2: Mutual information between response choices and factor scores. Only response choices from questions belonging to the factor are shown. The fraction of times a response choice was selected is plotted on the x-axis. 95% confidence intervals were calculated via bootstrapping with 1000 replicates. Labelled response choices are discussed in the text.

choice on the x-axis because a common way (and the way we initially intended) to evaluate a response choice is to examine how frequently it is picked. Figure 2 illustrates that response choices that are not picked frequently can sometimes provide more information than those that are.

We examine the factors separately beginning with the *evaluate models* factor (see Fig. 2a). Q2B_11 and Q3B_11, corresponding to the response choice "how close the points are to the line compared to the uncertainties", are the most informative response choices about students' scores on this factor. These two response choices are also the most expert-like for questions Q2B and Q3B, respectively. Similarly, Q2B_8 and Q3B_23, corresponding to

the response choice "how close the points are to the line", are the most novice-like response choices available to Q2B and Q3B, and also provide a relatively large amount of information about students' scores on the factor. We identified two classes of response choices that provided very little information about students' performance on this factor: Q2B_9 and Q3B_9, corresponding to the response choice "the number of outliers", and Q2B_21 and Q3B_21, corresponding to the response choice "the number of points above the line compared to the number below the line". These response choices also represent relatively novice-like ideas (picked by less than 25% of experts), but knowing whether a student selected them or not does not provide much information about their performance on this factor. Future work should seek to debug potential issues with these response choices.

For the *evaluate methods* factor (see Fig. 2b), the most informative response choice for all three questions that make up this scale, including Q4B, is "the number of masses tested" (Q1D_3, Q2D_4, Q4B_4). This is also the most expert-like response choice available for each question. Q1D and Q2D share three uninformative response choices: "how they tested other possible variables", "how clear, organized, or detailed their lab notes are" (Q1D_61, Q2D_35), and"their analysis and calculations" (Q1D_63, Q2D_33). These are all novice-like response choices (picked by less than 25% of experts), except for Q2D_33, which was picked by 44% of experts. Whether a respondent selects Q2D_33 or not provides very little information about their score on the *evaluate methods* factor despite this being a more expert-like response. This result may point to an issue with the response choice or students' interpretation of it that warrants further consideration.

The questions contained in the *suggest follow-ups* factor (Q1E, Q2E, and Q3E; see Fig. 2c) have more response choices (≥ 13) available to select from than the other questions. Individual response choices, then, generally have lower mutual information with this factor than was the case for the other factors. There are several highly informative response choices from these three questions that share the theme of extending the investigation: "test more masses" (Q1E_1), "test other variables" (Q1E_4, Q2E_6, Q3E_13), and "design a new experiment to test the non-zero intercept" (Q3E_11). These response choices are among the most expert-like for each question (picked by greater than 35% of experts), but there are other response choices that were picked by at least as many experts (and worth just as many points) that did not provide the same level of information about student performance on the factor. It appears that students with the goal of extending the investigation may generally score higher on this factor than students who select other response choices, including other expert-like response choices. There are, similarly, several response choices that consistently provided little information. One of these is "increase the number of bounces per trial" (Q1E_13, Q2E_14, Q3E_20; labelled in Fig. 2c). Other response choices that were relatively uninformative for all three questions include: "use their k to predict and compare a new mass" (relatively expert-like for Group 1) and "include other measures of uncertainty (e.g. stopwatch or ruler precision)". "Test more masses" was also among the most uninformative response choices for Q2E and Q3E, but is the most informative response choice for Q1E (Q1E_1). This indicates that whether a student suggests that Group 2 "test more masses" does not provide much information about their ability to suggest follow-up steps as measured by the PLIC, but students who identify that Group 1 should test more masses generally perform better on this factor.

IV. LIMITATIONS AND FUTURE WORK

In this paper, we have presented a novel method for evaluating the utility of response choices in closed-response assessments using the mutual information between response choices and factor scores. We first performed a CFA that established the predicted factor structure of the PLIC. This result provides researchers and instructors with another lens to examine data collected using the instrument; scores can now be separated into a finer-grained and more interpretable form using the factors established here. One subtlety in our method was the discretization of factor scores. We chose to discretize factor scores based on the recommendation in [13], but different choices for the number of bins will change the results as presented. If more bins are used, there will be increased variance as bins will contain fewer points, and if fewer bins are used we will lose information [9].

We then examined the mutual information between response choices and their associated factors. We could use the results obtained here for the PLIC to remove or modify response choices that provide little information about a student's performance. This could be particularly useful for questions, such as those associated with the *suggest follow-ups* factor, that have many available response choices. It is possible that a response choice could provide information about other factors as well, especially given that there is correlation between factors. In the future, this method could be extended to examine how much information is provided by response choices about vectors of factor scores rather than single factor scores.

The main purpose of this paper was to illustrate a new method for assessment developers to use when critiquing response choices for closed-response assessments. Mutual information can, however, be employed in many PER contexts and so we encourage researchers to consider broader applications of this tool in future research.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1611482. We would like to acknowledge Veit Elser for helpful discussions about developing this method, and Anne Alesandrini, Elias Euler, Erin Scanlon and Rachel Scherr for useful feedback.

[1] Wendy K Adams and Carl E Wieman, "Development and validation of instruments to measure learning of expert-like thinking," Int. J. Sci. Educ. **33**, 1289–1312 (2011).

[2] Lin Ding and Robert Beichner, "Approaches to data analysis of multiple-choice questions," Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[3] R Darrell Bock, "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," Psychometrika **37**, 29–51 (1972).

[4] Fumiko Samejima, *A New Family of Models for the Multiple-Choice Item.*, Tech. Rep. No. 79-4 (The University of Tennessee, Knoxville Department of Psychology, 1979).

[5] Bethany R Wilcox and Steven J Pollock, "Validation and analysis of the coupled multiple response colorado upper-division electrostatics diagnostic," Phys. Rev. ST Phys. Educ. Res. **11**, 020130 (2015).

[6] Cole Walsh, Katherine N Quinn, C Wieman, and NG Holmes, "Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking," Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[7] James T Laverty and Marcos D Caballero, "Analysis of the most common concept inventories in physics: What are we assessing?" Phys. Rev. Phys. Educ. Res. **14**, 010123 (2018).

[8] Eric Brewe, Jesper Bruun, and Ian G Bearden, "Using module analysis for multiple choice responses: A new method applied to force concept inventory data," Phys. Rev. Phys. Educ. Res. **12**, 020131 (2016).

[9] Thomas M Cover and Joy A Thomas, *Elements of information theory* (John Wiley & Sons, New York, 2012).

[10] An Gie Yong and Sean Pearce, "A beginners guide to factor analysis: Focusing on exploratory factor analysis," Tutor Quant Methods Psychol **9**, 79–94 (2013).

[11] Li-tze Hu and Peter M Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," Struct Equ Modeling **6**, 1–55 (1999).

[12] L.L. Thurstone, *The Vectors of Mind* (University of Chicago Press, Chicago, 1935).

[13] David W Scott, *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons, New York, 2015).

[14] Wentian Li, "Mutual information functions versus correlation functions," J Stat Phys. **60**, 823–837 (1990).