Validation Framework for Multiple-Choice Conceptual Physics Examinations

John Stewart

September 2023

1 Introduction

2 Identifiers and Versioning

All identifiers will be tagged at their end with the initials of the person who introduced the ID; for example, KD1-3-JS. In general, revisions that make changes but leave the construct measured (the knowledge model) the same should be left with the same general number KD1-3, but new versions can be added, KD1-3-V2JS. Revisions that change the construct measured should get a new core number. With this scheme, the sequence of versions cannot be inferred by the version number. As such, it is important that a new version be marked with the "revises" latex macro telling which version it is revised from. The old version should be given a status "HasNewVersion".

3 General Validation Plan

An item or instrument moves through the validation process by passing a number of tests. Each test can change the status of the item.

3.1 Validation Steps

New The items begins the validation process.

FV1 - Face Validity 1 The development team examines the new item and suggests revisions.

QL1 - Qualitative Testing 1 Five student think-aloud interviews to determine general item soundness. Item is revised based on interview. After this point, item revision is tracked.

QL2 - Qualitative Testing 2 Twenty student group session to further check stem validity and to identify additional distractors. Items revised.

QL3 - Qualitative Testing 3 Five student think-aloud interviews to determine if changes produced sound item. Items revised if issues are identified.

FV2 - Face Validity 2 The development team examines the revised item to determine whether it represents a reasonable item.

QT1 - Quantitative Testing 1 Item is tested at scale at one or more institutions. Item statistics are used to identify problematic items.

QL4 - Qualitative Testing 4 20 think aloud interview to verify qualitative validity of item. After this point the item is presumptively qualitatively valid.

QL5 - Qualitative Testing 4 10 think aloud interviews to verify qualitative validity of item at a second institution.

QT2 - Quantitative Testing 2 Item is tested at scale at one or more institutions. Item statistics are re-examined to verify items are valid.

Valid The items has completed validation process.

QT3 - Quantitative Testing 3 Using the data collected in QT2, a fairness analysis is performed.

Fair The item has completed fairness analysis. The item's state is the last stage of validation that the item has passed.

3.2 Item Status

New The item is at the beginning of the validation process and has not undergone testing.

Revised The item has been tested and revised to such an extent that it must be fully retested.

Inactive The item performed weakly on some tests and is not being actively validated. The issues were not great enough to force revision or removal from the item pool.

HasNewVersion The item has been tested and revised produces a new version. This version is no longer under consideration.

Active The item is actively being validated.

RequiresRevision Item failed to pass a certain level of testing but may still be valuable and needs to be revised.

NotUnderConsideration The item has not failed any tests but is currently not being tested.

Discontinued The item is no longer under consideration and is unlikely to be considered in the future.

Validated Item has passed all first level validation tests.

Fair Item has passed all first level validation tests.

4 Qualitative Item Validation

4.1 Student Response Process Data Collection

The Three-Step Test Interview (TSTI) protocol developed by Hak, van der Veer, and Jansen [**?**] was adapted for our qualitative validation process. This approach to cognitive interviewing combines a think-aloud protocol with retrospective probes, to first observe the student response behavior with as little interference as feasible and then fill in gaps in observations and elicit opinions from the respondent.

Step One

In a one-on-one interview setting, each student was asked to read an inventory item aloud and share whatever they were thinking, with explicit instruction from the researcher, "We'd like for any thought that comes into your head to come out of your mouth." The items were presented on a iPad that recorded audio and any written student work. The only additional prompts during the first stage of the TSTI were reminders to "Please keep talking," when a student ceased thinking aloud or "Please speak up," when a student was inaudible.

Step Two

After selecting an answer choice, every student was asked "You said (answer choice). How sure are you of that?" and "Was there anything about this problem that you think your classmates would find confusing?". These probes were intentionally open-ended to encourage students to provide whatever descriptive feedback they believe is relevant rather than constraining them to Likert scale answers.

In some cases, students who chose answers corresponding to salient distracting features of items decided to change their answers after being asked about their degree of certainty. Students frequently described the mindware they believed was necessary to answer an item, sometimes enumerating the specific content knowledge they worried they lacked as a reason for their own uncertainty or as a potential reason classmates might select a specific incorrect response. In addition to feedback on potentially confusing wording or figure features, students also described common misconceptions that could lead classmates to incorrect answers or what they believed their own pre-instruction answers would have been.

Step Three

In a standard TSTI, probes intended to clarify and complete the data from the think-aloud follow immediately rather than after opinion questions, but early students indicated that they became less confident in their answers after being asked by a researcher to elaborate so the TSTI protocol was modified by switching the order of canonical steps two and three. The amount of detail students provided while thinking aloud varied dramatically, and the retrospective prompts ”Can you tell me more about how you chose (answer choice) in this problem?” and ”Can you tell me more about how you made sense of the figure in this problem?” were used to elicit more detail from students whose reasoning process in answering a given item was unclear or, in some cases, completely unstated. Students who paused for lengthy periods, expressed discomfort or confusion, or reevaluated their choice of answers after making an initial selection were asked to describe their thoughts.

The researcher then selected from a list of constructed probes as appropriate to clarify potential issues with comprehension or lack of appropriate mindware available for retrieval to answer (e.g. ”What does the word (term) mean to you as used in the problem?”) and communication (e.g. ”Was there an answer you wanted to give that was not available in the choices shown?”).

4.2 Item Revision Using Student Response Process Data

Student think-alouds are compared with an expert reasoning model. Responses that are judged to be consistent with the expert reasoning model (or, in some cases, to provide an alternative reasoning path consistent with physics) and that result in the student selecting the correct answer are tentatively categorized as true positives. Responses that are judged to indicate incorrect reasoning and/or a lack of requisite physics knowledge that results in selecting an incorrect answer are tentatively categorized as true negatives.

More interesting are tentative false positives and false negatives, in which students respectively either demonstrate the declarative and procedural reasoning steps expected to answer correctly yet select an incorrect answer choice or are able to select the correct answer choice via incorrect reasoning. Analysis of these responses informed revision of items. In some cases, item features were found to be unexpectedly salient distractions, and we modified items in an attempt to reduce salience.

Item wording and figures were revised based on student feedback about confusing features in Step Two, and an item style guide with best practices for writing items that avoid common sources of (non-physics-content related) confusion is under construction. Some items have been revised as a result of a single distractor corresponding to multiple incorrect reasoning paths. Many other items had additional distractors added to correspond to student reasoning paths we had not anticipated during initial item development.

4.3 Reporting on Qualitative Validity

For items that complete the qualitative and quantitative validation process and are deemed valid in the tested contexts, we plan to provide a catalog of the correct and incorrect reasoning paths observed in cognitive interviews and the resulting answer choices with sample student quotes from each path, along with data on the prevalence of false negative and false positive responses processes. Multiple raters will participate in the qualitative coding of student responses to ensure adequate interrater reliability, which has not been feasible during the item revision stage due to the need for rapid turnaround times. Revision history and a summary of rationales for changes made to prior item versions will be available.

5 Quantitative Item Validation

Initial quantitative item validation began with Classical Test Theory and correlational statistics. Due to the nature of continual item development, values for test statistics were chosen to flag items that are performing abnormally and should be considered for revisions or removed from the testing pool. Although the term cutoff value is used, then realize that some items may remain in the testing pool with minor or no alterations due to other arguments for their inclusion. The *Item Difficulty* quantifies how difficult an item is. For dichotomously scored multiple choice questions, it is simply the proportion of students who got the item correct.

$$p_i = \frac{\text{number of participants with a score of 1 on Item i}}{\text{total number of participants}} \tag{1}$$

This leads to the counterintuitive fact that a higher item difficulty index means the item is easier, since more people got it correct. Item difficulty cutoffs were chosen to be .2 and .8 for items that are abnormally difficult and easy, respectively.

Item Discrimination refers to how well an item discriminates between those who perform well on the instrument and those who do not. It is calculated by splitting the sample into an upper and lower group then subtracting the proportion of students in the lower group who got the item correct from the proportion of students in the upper group who got the item correct, $D_i = p_{i_{upper}} - p_{i_{lower}}$. The upper and lower groups are formed by taking the top and bottom X percentile of total scores on the instrument, where X = 50^{th}, X = 27^{th}, and X = 25^{th} have all been commonly used. At large enough sample sizes, they will not differ significantly from one another, but the top and bottom 25^{th} percentiles were used for the development of this instrument. The common item discrimination cutoffs are .2 for elimination, or total revision, and .3 for marginal revision [Introduction to Classical & Modern Test Theory (Crocker & Algina)]. The problem with using a standard cutoff for item discrimination is that the item discrimination is dependent on item difficulty. A cutoff that accounts for this dependency is necessary, since it may be desirable by item developers to have items that are easier in order to demonstrate mastery of the basics, or conversely, a more difficult item to capture mastery of advanced reasoning.

Utilizing *Maximum Item Discrimination*(MID) allows for one to overcome this dependency, since it is calculated using the maximum possible difference between the high and low groups. Using the top and bottom 25^{th} percentile of total scores, a table similar to the one produced by [Basic Measurement and Evaluation of Science Instruciton (Rodney L. Doran)] can be calculated which reflects the maximum item discrimination.

Item Difficulty	0	.05	.1	.15	.20	.25	.30	.35	...	.65	.7	.75	.8	.85	.9	.95	1
p_{upper}	0	.2	.4	.6	.8	1	1	1	...	1	1	1	1	1	1	1	1
p_{lower}	0	0	0	0	0	0	0	0	...	0	0	0	.2	.4	.6	.8	1
MID	0	.2	.4	.6	.8	1	1	1	...	1	1	1	.8	.6	.4	.2	0
.3 MID Cutoff	0	.06	.12	.18	.24	.3	.3	.3	...	.3	.3	.3	.24	.18	.12	.06	0

Table 1: MID For Top 25^{th} Percentile & Bottom 25^{th} Percentile of Test Scores

p_{upper} and p_{lower} are obtained by considering the maximum possible discriminatory power for an item with the set difficulty. Consider an item with a difficulty of .1, then the best discrimination would occur when only those in the top 10% get it correct. Those top 10% would make up .4 of the top 25%, which is the upper group, and $p_{upper} = .4$. This procedure could be used to calculate MID for any desirable discrimination percentiles. Multiplying the MID by .3 was used to graph a line that would be used as the cutoff for MID.

There exist several ways to measure item performance within the overall exam, both using internal consistency and correlational coefficients. The Kuder Richardson 20(KR-20) coefficient is the dichotomously scored form of Cronbach's α

$$KR_{20} = \frac{k}{k-1}(1 - \frac{\Sigma_{allitems} p_i(1-p_i)}{\sigma_T^2}) \tag{2}$$

where k is the total number of items, p_i is the item difficulty, and σ_T^2 is the total test variance. A useful measure of an item performing inconsistently with an instrument is if the KR-20 with the item removed is higher than the original instrument KR-20, so the original instrument KR-20 value acts as a cutoff value for the KR-20 with the item removed.

The *Point Biserial Correlation Coefficient* is used to find the correlation between a dichotomous variable and a continuous variable, item score and test score in this case. The test score can be treated as continuous for sufficiently long tests, say 20 questions or more [Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment (Ding, Chabay, Sherwood, Beichner)]. As long as the test is of sufficient length, ≈ 25 items, then no correction for the item score being included in the test score is necessary. The point biserial correlation coefficient formula is

$$\rho_{i_{pbs}} = \frac{\mu_1 - \mu_T}{\sigma_T}\sqrt{\frac{p_i}{1-p_i}} \tag{3}$$

		Item j	
		0	1
Item i	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Table 2: Frequency Table For Dichotomously Scored Items

where μ_1 is the mean test score of those who got item i correct, μ_T is the mean test score of the total sample, σ_T is the standard deviation of the total group, and p_i is the item difficulty. A widely adopted cutoff for an item point biserial correlation coefficient is .2 [Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment (Ding, Chabay, Sherwood, Beichner)]. If the number of items is small to where the item score contributing to the total score is a problem, then the corrected by

$$\rho_{i_{T-i}} = \frac{\rho_{i_{pbs}}\sigma_T - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_T^2 - 2\rho_{i_{pbs}}\sigma_T\sigma_i}} \tag{4}$$

where σ_i is the item standard deviation.

If it can be reasonably argued that the latent variable underlying item performance is normally distributed, then the *Biserial Correlation Coefficient* can be used to find the correlation between this latent variable and a continuous variable such as test score. The formula for the biserial correlation coefficient is

$$\rho_{i_{bis}} = \frac{\mu_1 - \mu_T}{\sigma_T}(p_i/Y) \tag{5}$$

where Y is the ordinate of the standard normal curve at the z-score associated with the item difficulty, p_i. It should be noted that the biserial correlation will always be at least one-fifth greater than the point biserial correlation [Introduction to Classical & Modern Test Theory (Crocker & Algina)]. Since .2 was chosen to be the point biserial correlation cutoff value, then the biserial correlation cutoff value was chosen to be one-fifth greater than .2, or .24.

Two item statistics of interest were taken from the item-item phi correlation coefficient. The *Phi Correlation Coefficient* between two items is calculated using Table 2 with the formula

$$\rho_{ij} = \frac{n_{00}n_{11} - n_{01}n_{10}}{\sqrt{(n_{00}+n_{01})(n_{00}+n_{10})(n_{01}+n_{11})(n_{10}+n_{11})}} \tag{6}$$

where a,b,c,d are the number of respondents who scored 0 or 1 on the two items. After the phi correlation was calculated between every pair of items, then the average and maximum phi for an item were collected. The cutoff values for average phi correlation and maximum phi correlation were chosen to be .075 and .275, respectively.

Similar to how the biserial correlation assumes a normally distributed latent variable underpinning dichotomous item performance, the *Tetrachoric Correlation* is used between two dichotomous variables with underlying normally distributed latent variables. Correct item response then relies on passing a fixed, unknown threshold τ_i [Computational Aspects of Psychometric Methods (Martinkova & Hladka)]. Using Table 2 the marginal empirical probabilities of obtaining a score of 1 on item i and item j are, $\pi_{1\bullet} = \frac{n_{10}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$ and $\pi_{\bullet 1} = \frac{n_{01}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$, respectively. The thresholds can be found by enumerating the marginal empirical probabilities as

$$\tau_i = \Phi^{-1}(1 - \pi_{1\bullet}) \text{ and } \tau_j = \Phi^{-1}(1 - \pi_{\bullet 1}) \tag{7}$$

where Φ^{-1} is the quantile function of the standard normal distribution. **There are currently no cutoff values chosen for item tetrachoric thresholds.** The tetrachoric correlation, $\rho_{ij_{tetra}}$, is then found by using numerical optimizing algorithms to solve with respect to $\rho_{ij_{tetra}}$ the equation

$$\pi_{11} = \int_{\tau_i}^{\infty} \int_{\tau_j}^{\infty} \phi(y_i, y_j, \rho_{ij_{tetra}}) \mathrm{d}y_j \mathrm{d}y_i \tag{8}$$

where ϕ is a density of the bivariate standard normal distribution and $\pi_{11} = \frac{n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$ is a joint empirical probability. It should be noted that the tetrachoric correlation is a simplified case of the polychoric correlation for two dichotomous items. The tetrachoric correlation can be approximated by

$$\rho_{ij_{tetra}} \approx cos\left(\frac{\pi}{1 + \sqrt{\frac{\pi_{00}\pi_{11}}{\pi_{10}\pi_{01}}}}\right) \tag{9}$$

where $\pi_{00}, \pi_{11}, \pi_{01}$, and π_{10} are the empirical joint probabilities for scores on item i and item j. After the tetrachoric correlation was calculated between every pair of items, then the average and maximum tetrachoric correlations were collected. **There are currently no chosen cutoff values for average or maximum tetrachoric correlations.**

One problem with using the phi correlation, or tetrachoric correlation, to determine item performance is that item-item correlations may be due in large part to item-test correlations. Thus, a partial correlation accounting between items accounting for the test score may be more appropriate. The *Partial Correlation* between item i and item j accounting for the total test score T can be found using equations (3) and (6) such that

$$\rho_{ij \cdot T} = \frac{\rho_{ij} - \rho_{i_{pbs}}\rho_{j_{pbs}}}{\sqrt{1-\rho_{i_{pbs}}^2}\sqrt{1-\rho_{j_{pbs}}^2}} \tag{10}$$

where ρ_{ij} is the phi correlation between items i and j, $\rho_{i_{pbs}}$ is the point biserial correlation between item i and the test score, and $\rho_{j_{pbs}}$ is the point biserial correlation between item j and the test score. Note that partial correlations will be also calculated using the tetrachoric and biserial correlations instead of the phi and point biserial correlations, respectively. Using partial correlations between items as edges in a network graph can provide a useful interpretation for how items interact with one another, **but there are currently no chosen cutoff values for average or maximum partial correlations**.

6 Fairness Analysis

Fairness analysis will be conducted with Differential Item Functioning (DIF)