Issues Related to Common-Item Location in Test Design

Programa Estándares e Investigación Educativa
USAID Guatemala

In IRT scoring and equating, the probability of answering an item correctly is assumed to be a function of the examinee's ability and fixed item characteristics. Other characteristics of examinees or item position are assumed to have no effect. In some cases, a violation of this assumption has had significant effects.[1] In an important case of NAEP reading in 1986, significant changes in item position that changed the percent of students reaching certain items resulted in meaningful score differences.

When using IRT, changes in item parameters due to changes in item location may be very small for individual student scores; however, as Mislevy (1990)[1] argued, changes that are very small for individual students might add up to important differences when evaluating trends in achievement over time. In the case of 1986 NAEP reading test, the form changes were substantial, including changing the context in which reading passages were presented in association with writing and other subjects (a big change actually).

The *Standards for Educational and Psychological Testing*[2] address the concern regarding the IRT assumption of item invariance across forms and changes in item location on multiple forms. The *Standards* require that the potential for changes in item performance be investigated. This is the excerpt from the *Standards*:

> Standard 4.15
>
> When additional test forms are created by taking a subset of the items in an existing test form or by rearranging its items and there is sound reason to believe that scores on these forms may be influenced by item context effects, evidence should be provided that there is no undue distortion of norms for the different versions or of score linkages between them.
>
> *Comment*: Some tests and test batteries are published in both a full-length version and a survey or short version. In other cases, multiple versions of a single test form may be created by rearranging its items. It should not be assumed that performance data derived from the administration of items as part of the initial version can be used to approximate norms or construct conversion tables for alternative intact tests. Due caution is required in cases where context effects are likely, including speeded tests, long tests where fatigue may be a factor, and so on. In many cases, adequate psychometric data may only be obtainable from independent administrations of the alternate forms.

[1] Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10-16.

[2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

The psychometric advice is: do not change item locations on multiple test forms and keep anchor items used for equating in the same position. Having said that, small changes in item location are often required to provide greater test security – these purposes are not based on psychometric principles but on practical constraints or testing policy decisions. Based on past empirical research, there is enough concern regarding the context position and location of items to be very cautious about changing item location more than one page (that is, placing an item from page 1 to page 3 or further). Most measurement specialists recognize the practical constraints on testing programs and acknowledge the need to change item locations. Slight changes in location (for example, alternating odd/even items) do not present significant concern to psychometricians. Such practices should be evaluated after tests have been administered to assess the impact of this practice on item performance.

In the context of standards-based tests where there are significant opportunity-to-learn concerns and where a large number of students do not complete the test or tend to skip items, changing the location of items beyond one page may create complications in equating and maintaining item performance consistency and score consistency over time.

Programa Estándares e Investigación Educativa
USAID Guatemala

IRT models allow us to estimate person ability (theta) in a way that is item free (conditioned on item response) and not dependent on the specific items on a test and to estimate item parameters that are person free (conditioned on ability) and not dependent on the specific persons that responded to the item. In addition, IRT provides a test model that allows us to match test items to ability levels (item difficulty and person ability are on the same scale).[3]

In the 1PL (1-parameter logistic) model, data are transformed into measures that are interval level through a probabilistic model for distributions of item responses, probabilities of correctly responding to items are converted to logits, allowing us to compare item difficulties and person abilities across tests and over time. One primary benefit of the 1PL model is its simplicity and ease of explaining to the public. But more technically, this simplicity is associated with an important fact: Examinees with the same raw score will have the same ability (theta) value. This suggests that each item is valued similarly.

The alternative, 2PL or 3PL models, will allow for "pattern" scoring, which suggests that not only the number correct is important for estimating a score, but to which items the examinee responds correctly will determine the score. This is because some items do not provide as much information for the estimation of some levels of ability. This treats items differently in terms of their influence in estimating a score.

For individual examinees, this may have a positive effect, while for others, it will have a negative effect. For example, in a few areas, Mayan numbers are covered well and students learn and practice the use of this number system. In many regions and schools, students do not spend much time learning or practicing the Mayan number system. Item parameters will reflect this by resulting in high levels of difficulty with low levels of discrimination. These items will not discriminate between high ability and low ability students because performance is not a function of ability but is a function of opportunity to learn. Now, ability scores (person-theta values) will not be a function of the items covering Mayan numbers because the low discrimination will be used to weight these items less than the other items. Students who get all of the Mayan-number questions correct will not get as much credit toward their ability score as students who may get other questions correct and all of the Mayan-number items incorrect.

[3] Hambleton, R.K. and Jones, R.W. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

In the context of an assessment that is new with the understanding that some of the content is not being taught, or at least not taught consistently across the country, the differential valuation of items will create interpretation problems over time – particularly as the curriculum that is actually taught catches up with the curriculum based on the standards. As the curriculum improves and students have more opportunity to learn the curriculum as specified in the standards, the weights associated with different items will change.

The change in item parameters over time introduces additional problems for equating – the methods of putting tests on the same scale over time to improve comparability of scores over time. If items for topics which students do not have consistent opportunities to learn are used to anchor the scale through equating, the equating function will become more distorted over time. The anchoring of item parameters fixes the parameters over time (treats them as though they are constant and unchanging), when in fact the item parameters may be changing because of changes in opportunity to learn (for example, the items may become more discriminating as the content is being taught more consistently).

Equating assumes that the same content is being covered from year to year, and although it is not part of the assumptions for estimation, differential opportunity to learn has implications for the validity of equating.

The primary purpose of the national assessment program in Guatemala at this time is for system reform and monitoring standards-based achievement of groups (schools, departments). Within this goal are research interests in group differences (region or location of school and subgroups of students). Research on the use of 1PL and 3PL models for estimating group differences indicates that there are no consistent differences between the two models, even when the 3PL model fits the data much better. "These findings [and those of others] suggest that educators doing substantive research (addressing questions other than the properties of the test or the scores of particular students) would not gain much accuracy by choosing one model of the other to score the test".[4]

To be fair, 2PL and 3PL models provide person ability scores that are more precise and more sensitive to variation in ability. Also, more parameters almost always fit the data better than one parameter (we can describe the more completely with more parameters), which also gives information about the quality of person scores. But when 2PL or 3PL models are used, and the weighting of items (based on content) changes from one year to the next, the inference about changes in achievement are less meaningful – Is there real change in achievement due to a change in ability or due to a change in how items are weighted? If items are becoming more discriminating because of improved opportunity to learn, this affects item parameters, person scores, and our interpretation of results.

[4] DeMars, C. (2001). Group differences based on IRT scores: Does the model matter? *Educational and Psychological Measurement, 61*(1), 60-70.

Developing an Equating Plan

Programa Estándares e Investigación Educativa
USAID Guatemala

The program has selected the Rasch model for horizontal equating of forms within a year and across years. The equating plan is horizontal because it does not include equating across levels. Equating will be based on a common-item equating. In the future (3 or 4 years into the evaluation program), it is possible that this equating plan can be evaluated by using a small number of items from the first year to check the integrity of the equation over time.

A strong equating includes concurrent equating of multiple forms in the first year to define the initial scale as broadly and deeply as possible by including all items in the initial scaling – the scale to which all future forms will be equated. To assess the viability of concurrent equating of multiple forms in the first year and for the use of anchor items across time, the following standard procedures will be followed.

1. *Cross-plot item difficulties of common items from each form*:
    a. Forma A on the X-Axis
    b. Forma B on the Y-Axis

    In this graphic cross-plot (typically done is SPSS or Excel), the slope should be near 1.0. If the slope is near 1.0, then the intercept with the X-axis is the equating constant to make the appropriate score adjustment across forms. To complete this score adjustment:

    Test B measure in Test A frame of reference = Test B measure + X-Axis intercept
    [this is an approximation].

    This is the typical process for equating using common-items where the equating adjustment is done by hand (adding the equating constant to each measure from Test B).

    If the slope is NOT near 1.0, drop common items that are not on the common item line. If by dropping an item with very different p-values on two forms, then the slope approaches 1.0, this item should no longer be considered a common item for the purpose of equating. The item should be treated as a different item and freely calibrated.

2. *Equating Methods*

    In Winsteps, the procedures for equating scores (measures) for individuals can be done automatically through using scaling constants – rather than doing this by hand outside of Winsteps procedures. There are three decision rules that can be used. The first two are options when doing horizontal equating over time, where Test B is being equating to the score scale from Test A (from the previous year) – or can be considered as a generic process for common-item equating using anchors (where anchors come from Test A).

a. If the common item-difficulty slope is far from 1.0, adjust Test B Command File by adding the following command statements:
USCALE = value of 1/slope = SD_A/SD_B
UMEAN = value of X-Intercept = $Mean_A$ - $Mean_B$

b. If the slope is near 1.0, and Form A is the reference form, then the IFILE (item parameter file) from a previous calibration can be used as an IAFILE (item anchor file) in the new calibration analysis. The IFILE needs to be edited to include only the anchor items and so that the items are in the correct order based on the location of anchor items in the new forms. One way to edit the IFILE to tell Winsteps to ignore non-anchor items is to place a semicolon (;) before the item to be ignored.

An alternative is to add the IAFILE command directly in the command file for Form B, as in the following example:

```
…
IAFILE = *
2  -1.234
5   0.048
16 1.328
17 -0.978
…
*
&END
```

Consider equating Form B (new form) to Form A (old form). The first number is the item sequence number from Form B (it is the item location of the common item on Form B, not the item number from Form A), the second number is the theta-value for the item based on Form A calibration (the anchor measure-values from Form A IFILE).

c. As an alternative, if the slope is near 1.0, and Form A and Form B are equally weighted (one is not considered the reference form), then it is possible to employ concurrent equating with MFORMS command to concurrently calibrate all items clearly identifying common items across forms. MFORMS must be used to restructure the response data so that unique items and common anchor items are correctly named.

Decision Summary

In summary, there are two methods of equating in the base year (the year the scale is developed).

- One is to identify a reference form and equate all other forms to the reference form.
  - This identifies one single form as the core form and all other forms are placed on the scale of the core form.
  - This allows direct equating to a single form for future forms in future years.
- The second is to do concurrent calibration.
  - This implies that all unique items across each form are important for establishing the scale. This may produce a scale that more accurately reflects the construct, if the construct is better defined with all of the forms instead of a single core form.
  - In future years, all forms must then be scaled to the reference year using anchors for common item equating.

Programa Estándares e Investigación Educativa
USAID Guatemala

Theta (θ) values are not practical since they typically range from -3.0 to +3.0. These scores can be rescaled using a linear function to reset the mean and standard deviation.

The location of the scale should be considered carefully because it could be set to serve a specific function; for example, the scale could be centered at the cut score for "satisfactory".

The following formula rescales scores from Y on the X score scale:

$$\text{Scale-Score} = \frac{S_X}{S_Y}\left(Y - \bar{Y}\right) + \bar{X}$$

This formula can be used in a more general manner, to center the scale on a cut-score based on the θ value associated with the item sequence number selected by the standard setting panel (as an advisory group).

For example, if the "Satisfactory" cut-score was associated with a raw score of 26, and the 26th item in the item order table was associated with a θ = .5533, we could center the scale on .5533.

To continue the example, if we choose the scale-score = 500 to signify "Satisfactory", with a scaled-score SD = 50, we can use these values to convert θ to the scale-score.

Desired Scale-Score Mean = 500, SD = 50
Theta associated with Cut-Score = .5533, observed SD(θ) = .6907

$$\text{Scale-Score} = \frac{50}{.6907}(\theta - .5533) + 500$$

```
COMPUTE sscore = (50/.6907)*(measure-.5533) + 500.
EXECUTE .
```

This converts θ values for individuals (in SPSS this has the label "measure") onto the scale-score scale where 500 is the "Satisfactory" level and the SD = 50.

To convert the Standard Error of the Scale-Score:

$$\text{SE(SS)} = \frac{SD(SS)}{SD(\theta)} SE(\theta)$$

```
COMPUTE ss_se = (50/.6907)*se.
EXECUTE .
```

SE(θ) is reported in the PFILE associated with each θ value for each student.

Programa Estándares e Investigación Educativa
USAID Guatemala

Standard setting panels typically participate in the standard setting process by recommending cut scores. These recommendations results from their participation in one of a number of standard setting procedures. Whatever standard setting procedure is used, the result is a recommendation made to education decision makers who are responsible for the final performance standards.[5]

When the final standards are set, the input of the panels should be held in high regard to the extent that the evaluation of the process supports their final recommendations. In any case, the recommended cut scores should be reviewed for potential adjustments. This brief provides conditions and considerations that may result in justification for adjusting the recommendations of standard setting panels to define the final cut scores.

There is no psychometric standard that defines the conditions under which cut scores as recommended from standard setting panels should be adjusted or that limit the degree to which cut scores may be adjusted. In all cases, performance standards are judgment-based and the final decision is always a policy-based decision, but should be informed through a process supported by psychometrics.

1. Secure the best data from the standard setting panel.
   Ratings of an individual panel member may be considered invalid and removed prior to summarizing the results for the panel for a number of reasons, including if the panelist:
   a. had a difficult time understanding the process;
   b. made comments that were not consistent with the process or purpose of standard setting;
   c. clearly was out of line with the other panelists or trying to make a point or argument when setting a cut score.

2. Consider the possibility of graduated standards.
   For new testing programs with high standards, one option is to begin with lower standards so that schools have time to adjust to the new standards and develop the curriculum expertise needed to meet the demands of the new tests. This may be done in a number of ways, including two common methods[6]:
   a. Subtract 1 or 2 times the standard error of measurement;
   b. Subtract the error related to interjudge variability.

[5] Cizek, G.J. (2006). Standard Setting. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 225-258). Mahwah, NJ: Lawrence Erlbaum.
[6] Hambleton, R.K., & Pitoniak, M.J. (2006). *Setting performance standards*. In R.L. Brennan (Ed.), Educational Measurement (4th ed.) (pp. 433-470). Washington DC: American Council on Education and Praeger Publishers.

3. Consider the consistency and coherence of standards set across grade levels.[7] There are many conditions that should be considered in making decisions about this process[8]:
   a. The purpose of the test
   b. The potential effect of changing passing rates over time
   c. Needs of the country
   d. Socially and politically acceptable passing rates/failure rates
   e. Adverse impact on subgroups (would changing standards widen performance gaps among subgroups of students)
   f. The role of motivation and effort of students (addressing issues related to the accuracy of scores because students do or do not take the tests seriously)

This last form of making adjustments to standards recommended by standard setting panels is largely a policy-based decision, but should be informed based on at least two considerations, including content differences across grades and performance expectations across grades. If the standards as recommended appear consistent with the purpose of the test, the content of the test and professional judgment about the coherence between instructional practice, curriculum standards, and opportunity to learn, then differences in passing rates across grades are not of concern. When the differences in passing rates are quite different and no argument can be made that justifies such differences based on the curriculum standards, test content, and expected passing rates, then smoothing out such differences should be considered.

There may be purely political reasons for making passing rates consistent across grades. If content standards and learning expectations, opportunity to learn, and educational resources are fairly consistent across grades, then we should expect passing rates to be similar. Differences in passing rates might otherwise suggest differences in the demands of the tests as set by the standards – not expecting the same level of performance across grades.

There are no commonly agreed upon methods for making such smoothing decisions. Statistical models for smoothing results are not defensible – there are no common items across grades, content shifts are significant and meaningful, and the population of students changes dramatically from primary to secondary educational programs. Most adjustments in practice tend to be based on moving the cut scores in terms of standard errors of measurement – toward a common outcome (passing rate). The process used and direction of the adjustment should be consistent with considerations of the purpose and content of the test.

If the intent is to set high performance standards, then those grades with higher passing rates might be adjusted downward to be consistent with the grades that have lower passing rates (as recommended by the standard setting panels). If the role of the test is to function as a basic skills test (minimum competency), then the grades with low passing rates might be adjusted upward.

---

[7] Crane, E.W., & Winter, P.C. (2006). *Setting coherent performance standards.* Washington DC: The Council of Chief State School Officers.

[8] Hambleton, R.K., & Pitoniak, M.J. (2006). *Setting performance standards*. In R.L. Brennan (Ed.), Educational Measurement (4th ed.) (pp. 433-470). Washington DC: American Council on Education and Praeger Publishers.

Many entities setting performance standards across grades will convene a panel of participants from grade-specific standard setting panels to meet in a joint cross-grade panel to review the results of the individual panels. They will consider process, purpose, content, and all of the other conditions described above. They can provide additional advice about the potential direction of adjustments.

The alternative is to convene a meeting of individuals representing organizations that have a stake in the decision (individuals from schools, teacher organizations, and government agencies including the ministry of education) and ask for additional input regarding their best professional opinion on what percent of students currently are proficient given the curriculum standards. This will guide the decision as to whether it is necessary to adjust cut scores to smooth out variation in passing rates across grades and inform the decision about the direction of that smoothing – to smooth upward to raise the lower passing rates or to smooth downward to lower the higher passing rates.

There is no reason, all else considered, for all passing rates to be exactly the same across all grades. However, differences should be explainable. Even with adjustments to cut scores, resulting passing rates do not need to be equivalent for psychometric reasons. That is to say, any adjustment to cut scores to smooth passing rates across grades does not have to result in equal passing rates across grades.

Whatever process is used to set the final cut scores, the *Standards*[9] require this process to be explained and documented.

[9] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.