USING ITEM MAPPING TO EVALUATE ALIGNMENT BETWEEN CURRICULUM AND ASSESSMENT

USANDO O MAPEAMENTO DE ITENS PARA AVALIAR O ALINHAMENTO ENTRE O CURRÍCULO E A AVALIAÇÃO

USO DE LA ASIGNACIÓN DE ELEMENTOS PARA EVALUAR LA ALINEACIÓN ENTRE EL PLAN DE ESTUDIOS Y LA EVALUACIÓN

Leah T. Kaira

Stephen G. Sireci

ABSTRACT

In educational testing, it is critical that the content of a test is aligned with the curriculum the test is designed to measure. Most methods for evaluating test-curriculum alignment rely on the subjective judgment of content made by experts who focus on how well the items on a test match curricular objectives. However, it is also important to ensure educational test items align with their expected levels of difficulty, which is much harder for experts to judge. In this study, test-curriculum alignment was evaluated by assessing the degree to which observed item difficulty aligned with intended item difficulty as determined by the test specifications. Using student response data for the Massachusetts Adult Proficiency Test (MAPT) for math, Item Response Theory (IRT) was used to locate items on the proficiency scale using two criterion response probability (RP) values. Item mapping results were compared to the item writers' classifications of the items, and degree of agreement between the two sets of data were statistically compared. In general, higher alignment was observed using RP50 than RP67, and for items assessing lower cognitive levels. Subject matter experts concluded cognitive demand, item clarity, and language complexity were viable reasons for misalignment.

Keywords: alignment; item mapping; Item Response Theory; response probability; validity.

RESUMO

Na avaliação educacional, é importante que o conteúdo do teste esteja alinhado com o currículo que ele pretende avaliar. A maioria dos métodos para avaliar o alinhamento entre o teste e o currículo tem como base o julgamento subjetivo do conteúdo feito por especialistas, que avaliam o quanto os itens do teste correspondem aos objetivos propostos no currículo. Não obstante, também é relevante garantir que os itens estejam alinhados com o nível de dificuldade esperado para eles, o que é mais difícil para os especialistas julgarem. Nesse estudo, o alinhamento entre o teste e o currículo foi verificado por meio da avaliação do grau com que o nível de dificuldade observado está de acordo com o esperado, conforme as especificações do teste. Para tanto, foram utilizadas as respostas dos estudantes ao Teste de Proficiência em Matemática para Adultos de Massachusetts (MAPT). A Teoria de Resposta ao Item (TRI) foi utilizada para localizar os itens na escala de proficiência usando os valores de dois critérios de probabilidade de resposta (RP). Os resultados do mapeamento dos itens foram comparados com a classificação feita pelos elaboradores e o grau de concordância entre os dois conjuntos de dados foram comparados estatisticamente. Em geral, um maior alinhamento foi observado usando RP50 do que RP67, e para itens que avaliavam níveis cognitivos mais baixos. Os especialistas concluíram que a demanda cognitiva, a clareza do item e a complexidade da linguagem foram as razões mais prováveis para o desalinhamento.

Palavras-chave: alinhamento; mapeamento de itens; Teoria de Resposta ao Item; probabilidade de resposta; validade.

RESUMEN

En la evaluación educativa es importante que el contenido de la prueba esté alineado con el currículo que se pretende evaluar. La mayoría de los métodos para evaluar la conformidad entre la prueba y el currículo tiene como base el juicio subjetivo del contenido hecho por especialistas, que evalúan cuánto corresponden los puntos de la prueba con los objetivos propuestos en el currículo. No obstante, también es relevante garantizar que los puntos estén alineados con el nivel de dificultad esperado para ellos, lo que para los especialistas es más difícil juzgar. En este estudio, la correspondencia entre la prueba y el currículo fue verificado por medio de la evaluación del grado con que el nivel de dificultad observado está de acuerdo con lo esperado, según las especificaciones de la prueba. Para esto, fueron utilizadas las respuestas de los estudiantes al Test de Competencia en Matemática para Adultos de Massachusetts (TCMA). Se utilizó la teoría de respuesta al ítem

(TRI) para ubicar los ítems en la escala de competencia usando los valores de los criterios de probabilidad de respuesta (PR). Los resultados del mapeo de los ítems fueron comparados con la clasificación hecha por los elaboradores, y el grado de concordancia entre los dos conjuntos de datos, se compararon estadísticamente. En general, una mayor alineación se observó al usar RP50, del que RP67, y para puntos que evaluaban niveles cognitivos más bajos. Los especialistas concluyeron que la demanda cognitiva, la claridad del punto y la complejidad del lenguaje fueron las razones más probables para la falta de alineación.

Palabras clave: alineación; mapeo de ítems; Teoría de Respuesta al Ítem; probabilidad de respuesta; valores.

Introduction

In educational testing, accurate evaluation of student learning can be achieved only if there is agreement among the curriculum, what the students learn, and what appears on the assessment. Similarly, assessment results are useful for accountability purposes if the assessment mirrors the curriculum. One strategy for evaluating the match between a curriculum and the assessment designed to measure it is carrying out alignment studies. Bhola et al. (2003, p. 21) define alignment as "the degree of agreement between a state's content standards for a specific subject and the assessment(s) used to measure student achievement of these standards".

Alignment is closely related to the interpretations made from test scores. According to the Standards for Educational and Psychological Testing (the Standards) (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION [AERA], AMERICAN PSYCHOLOGICAL ASSOCIATION [APA], & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION [NCME], 2014, p. 11), validity is "the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests". Validation is therefore a process of collecting evidence to support the type of inferences that are drawn from test scores. Results of an alignment study can thus be used as validity evidence to support the interpretation of test scores.

The goal of alignment is to establish the degree of match between test content and subject area content as specified in curriculum standards. It

is important to emphasize the expression 'degree of agreement' because, as La Marca et al. (2000, p. 18) noted, "It is improbable that a single assessment instrument will provide the breadth of coverage necessary for an aligned system".

Breadth and coverage of curriculum standards necessitates test items that vary in difficulty. Some alignment methods, such as the Achieve Method (ACHIEVE, 2001), go beyond test-curriculum alignment by also evaluating "challenge," that is the degree to which the observed difficulty levels of items match their expected difficulty levels. In evaluating the "level of challenge" of items on a test, reviewers determine whether the sets of assessment items span an appropriate range of difficulty for students in the target grade level.

Current alignment methods

There has been an increase in research aimed at developing methodology for assessing alignment. According to Bhola et al. (2003), alignment methods can be categorized as having a low, moderate or high complexity based on level of focus, that is, the number of dimensions considered. For instance, a low complexity alignment study would only focus on the match between content of the items and the standards, while a high complexity study would also consider other dimensions such as the match in depth of content and the match between the levels of emphasis placed on a particular content area in the curriculum and in the assessment. One implication of this categorization is that different alignment studies may come up with different results depending on the levels of focus employed. As such, results from alignment studies of the same assessment, but employing different levels of focus, cannot be meaningfully compared.

Almost all alignment methods use subject matter experts (SME) to ensure they clearly understand the standards, the alignment criteria, and the scales being used to judge alignment. While expert judgments are essential in various steps in educational assessment, it is well known that despite training, humans make errors of unknown magnitude in their judgment. For example, Bhola et al. (2003) noted that SMEs may be overly generous in the number of matches they envision. Apart from the financial resources and the time required to convene SMEs, having SMEs review each item and make judgments over multiple criteria can also be cognitively challenging.

Another problem with current alignment methods is a lack of consensus regarding what constitutes sufficient alignment. Ananda (2003b, p. 20) noted one reason for lack of consensus is "...when articulating expectations for what students should learn (what they should know and be able to do), it is common for states to have different levels of statements, ranging from more global statements ...to narrower more targeted statements clustered under the broader statement". Thus, choice of alignment method is partly dictated by the breadth of statements describing what students should learn. This outcome could pose problems in evaluating improvements in the assessment as measured by student achievement.

Some alignment methods, such as the Achieve (2001) method, try to evaluate the appropriateness of the range of difficulty of the items on an assessment and the grade level of the students the assessment is intended for. In this process, it is assumed that after some training the SMEs have a common understanding of the range of abilities of the students in the target grade, and that they can accurately judge the difficulty of the item for a target group. However, research has shown that it is difficult for SMEs to make accurate judgments about the difficulty of items (Impara & Plake, 1998; Plake et al., 2000; Ryan, 1968; Shepard, 1994; Plake & Impara, 2001).

A good example of this difficulty is the 1990 National Assessment of Educational Progress (NAEP) math standard setting study in which great variability was observed among SMEs in making item judgments, despite training. The United States General Accounting Office (1993) claimed that the instruction given to the SMEs during training was not sufficient to bring the SMEs to a common understanding of what students at different achievement levels should know and be able to do. As a result, each SME formulated their own definition of what a basic, proficient or advanced student can do, resulting in large variability of judgments among the SMEs. The consequence of this variability was cut scores that were largely disputed and viewed as not representative of the knowledge and skills of the students assessed.

With respect to current alignment studies, it seems that evaluating the alignment of test items to their intended difficulty levels is important, but that alignment studies that rely on SMEs' subjective judgments are not going to be effective. That is, a mismatch between the SMEs' understanding of the range of student abilities at the target grade, and what the students

can actually do, could lead to item difficulty alignment results that are erroneous and misleading.

Thus, it seems reasonable to consider other approaches for evaluating item difficulty alignment than methods that rely on subjective judgment. In particular, methods are needed that (a) account for student's actual performance on items, (b) reduce reliance on subjective human judgment, and (c) apply consistent criteria for evaluating alignment.

An item mapping approach to evaluate item difficulty alignment

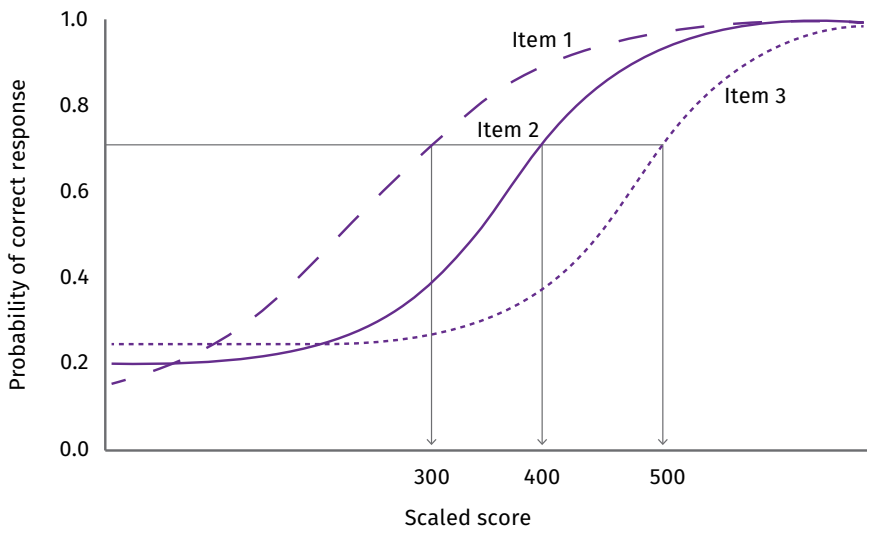
One method that could be used to evaluate the alignment of the intended and the observed difficulty of an item is item mapping. Item mapping has been widely used in educational assessment in areas of standard setting (e.g., Wang, 2003), scale anchoring (e.g., Gomez et al., 2006), and score reporting (e.g., Hambleton, 1997; Kirsch et al., 1993). Despite the various applications, the ultimate purpose of item mapping is to identify and describe what students at a specific level of achievement know and are able to do.

For the purposes of this study, item mapping is simply defined as the process of locating items along the test score scale. The idea behind item mapping is that given their characteristics, items could be systematically located on the test score scale based on some criteria. In most cases, the criterion used is the likelihood that examinees of a specified proficiency level have a high probability of success on the item.

The most popular approach for mapping items is the use of Item Response Theory (IRT). In IRT models, student achievement levels and item difficulties are on the same scale. Thus, given an examinee's proficiency, items the examinee would most likely answer correctly can be identified. The phrase 'most likely answer correctly' is usually defined by the probability that the examinee gives a correct answer to an item. This probability is referred to as the response probability (RP) criterion. In IRT models, each item is represented by an item characteristic curve (ICC), which gives the probability of correctly answering an item for a given proficiency level. Figure 1 shows ICCs for three dichotomously scored items. Item 3 has the lowest probability an examinee would give a correct response throughout most of the score scale. This implies that item 3 is more difficult compared to items 1 and 2.

Using a response probability of 70% (i.e., RP70), items 1, 2 and 3 would be mapped to scale scores of 300, 400, and 500 respectively. This means for example, that students with a scale score of 300 could be expected to correctly answer item 1 about 70% of the time. Similarly, students with scaled scores of 400 and 500 would be expected to correctly answer items 2 and 3, respectively, about 70% of the time.

Figure 1 Item characteristic curves for 3 hypothetical items



Application of item mapping to alignment

Item mapping could be used in an alignment study by locating items at specific points on the test score scale to help describe what students at that proficiency level can do. In cases where curriculum standards span several grade levels, and a vertical scale exists across those grade levels, the degree to which the items written for curriculum at higher grade levels are more difficult than items at lower grade levels can be evaluated. A system of tests that are aligned with respect to item difficulty will have items located along the IRT scale at locations that are implied by the test specifications. If items are located higher or lower with respect to their difficulties, some form of misalignment is present, and the source of the misalignment should be investigated. In this study, we show how item mapping can be used to

identify items that are misaligned with respect to their difficulty, and we use SMEs to evaluate reasons why the items are misaligned.

Purpose of current study

The purposes of our study are to investigate the utility of item mapping for evaluating the alignment between intended item difficulty (in terms of the grade span in which items are located) and actual item difficulty, and to discover reasons why misalignment in item difficulty may occur. The specific questions are:

- Can item mapping be used to enhance the evaluation of curriculum-assessment alignment from the perspective of item difficulty?
- Do response probability values have an impact on item difficulty alignment?
- If misalignment is observed, what are the likely causes?

Method

Empirical data were used to illustrate use of item mapping in assessing alignment among curriculum and assessment. The analyses were applied to data from a large-scale assessment in adult education.

Description of test

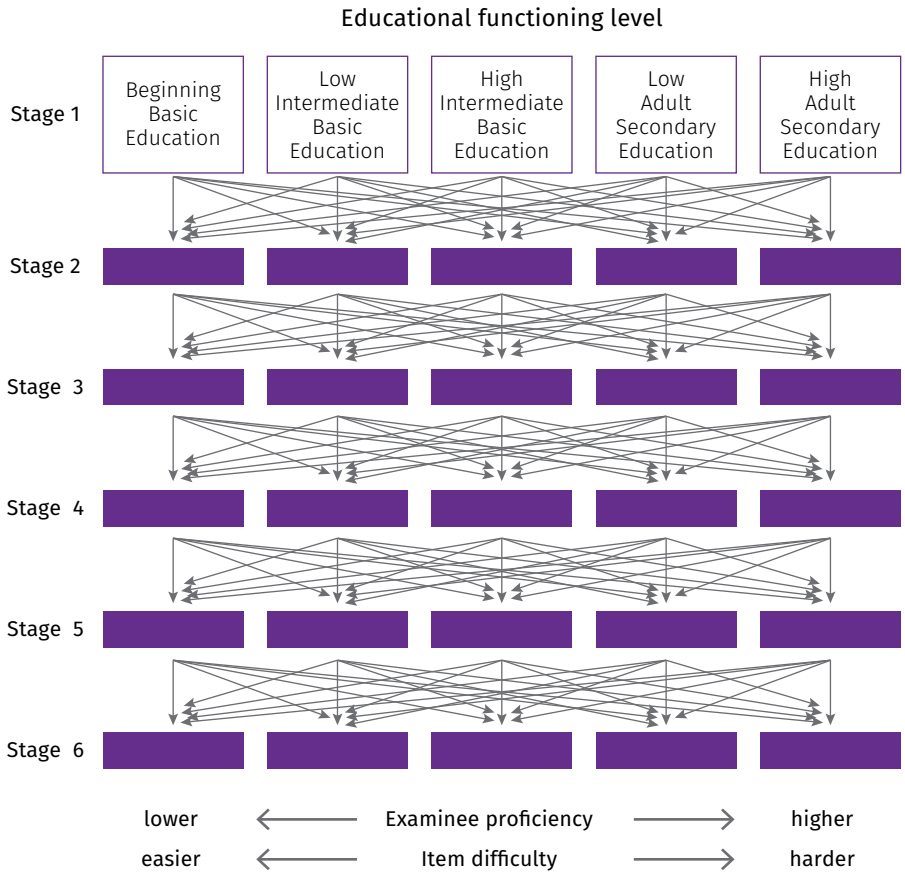
The Massachusetts Adult Proficiency Test (MAPT) for mathematics and numeracy is a computerized-adaptive multistage test (MST) designed to measure the mathematics achievement of adult education students in Massachusetts, USA. The MAPT for math is used by the State of Massachusetts to fulfill the Federal accountability requirements in adult education known as the National Reporting System (WASHINGTON, 2006). The NRS stipulates 6 achievement levels called educational functioning levels (EFLs), which are similar to grade levels in elementary through secondary school (i.e., each EFL spans about two grade levels). The MAPT measures all but the lowest EFL, and students taking the MAPT for the first time begin the test at an EFL designated by their teacher. Thus, there are 5 starting points for the MAPT. However, there are not separate tests for each EFL. Rather, all EFLs are calibrated along a common score scale and examinees may be routed to any EFL as they take the test, depending on how well they perform on the items administered at each stage.

The MST design for the MAPT involves six-stages, as illustrated in figure 2. The design includes two parallel panels, each consisting of 30 sets of items called modules. A panel is a collection of modules that defines all potential paths examinees may be routed to when taking the test (SIRECI et al., 2008). In MST, panels are analogous to alternate forms as defined in linear testing. The arrows in figure 2 show some (but not all) potential paths to which examinees may be routed. The first time a student takes the MAPT s/he is randomly assigned to one of the two panels. The other panel is used for a second test administration. A total of 40 scored items are administered to each student across the six stages. Students take 15 items during the first stage and 5 items in each of the subsequent stages. Proficiency estimates at each stage are used to determine the set of items (i.e., module) the examinee will take during the next stage. All items are dichotomously scored multiple-choice items with four answer choices.

The content of the MAPT for math is specified using two dimensions — one for test content and one for cognitive level. Four content areas are measured — Geometry and Measurement; Patterns, Functions and Algebra; Statistics and Probability; and Number Sense (hereafter referred to as Geometry; Patterns; Statistics; and Number Sense, respectively). The distribution of the items is 84, 68, 93, and 116 across the four content areas, respectively. With respect to cognitive level, three levels are specified — Knowledge and Comprehension; Application; and Analysis, Synthesis, and Evaluation. There were 114 items assessing Knowledge and Comprehension, 175 items assessing Application, and 73 items assessing Analysis, Synthesis and Evaluation. For convenience, the three cognitive skill areas will be referred to as Comprehension, Application, and Evaluation, respectively.

Each panel of the MAPT is designed to assess students' proficiency in math at five different EFLs: Beginning Basic, Low Intermediate, High Intermediate, Low Adult Secondary, and High Adult Secondary. There are separate test specifications for each EFL corresponding to the specific curricula for the EFL as described in the Massachusetts Adult Basic Education Curriculum Frameworks for Mathematics and Numeracy (SIRECI et al., 2008).

Figure 2 MST structure for the MAPT for math



Item mapping data

Response data for both panels for the 2009 administrations of the MAPT for math were used. About 7,361 examinees' responses to 362 math items were analyzed. The 3-parameter logistic IRT (3PL) model was used to estimate item parameters from examinee responses. Parameter estimation was done using BILOG-MG (ZIMOWSKI et al., 1996).

The items were also coded by the test developers with respect to content attributes including EFL, content strand (Geometry and Measurement; Patterns, Functions and Algebra; Statistics and Probability; and Number

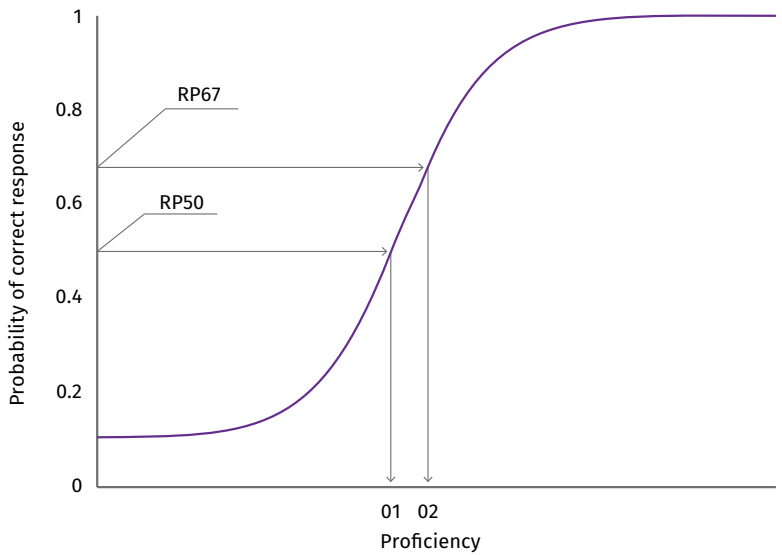
Sense), and cognitive skill (Knowledge and Comprehension; Application; and Analysis, Synthesis and Evaluation). The EFL designations for each item were originally made by the item writers, and subsequently confirmed by an independent group of SMEs. There were 100 Beginning Basic, 97 Low Intermediate, 94 High Intermediate, and 71 Low Adult Secondary items.

Item mapping method

We used a model-based item mapping method to identify items that mapped to a particular EFL. The steps were:

- Obtain parameter estimates for each item using the 3PL IRT model. We used the parameters for all items that were operational in 2009, with the exception of items in the High Adult Secondary EFL.
- Given the item parameter estimates, calculate the theta (θ) value required for an examinee to have some specified probability of correct response for each item. Figure 3 illustrates this estimation. The task is to find θ_1 and θ_2 for each item for which examinees have a probability of .50 (RP50) and .67 (RP67), respectively, of success on the item.
- Determine the EFL to which each item mapped. We used theta values obtained in step 2, and the cut scores for each EFL.

Figure 3 An item characteristic curve illustrating the Model Based Item Mapping Method



Response probability values

RP50 and RP67 were used to determine the EFL to which each item mapped and to assess the impact of RP value on the alignment results. RP50 and RP67 were chosen because these are the most common RP values in literature (KARANTONIS & SIRECI, 2006). Use of these RP values allowed for comparison of results of this study with findings of similar studies reported in literature. Second, because a goal was to illustrate how item mapping could be applied to evaluation of curriculum-assessment alignment, an operational definition of what students "can do" was needed. Based on literature, there seems to be a consensus that for tests that do not have very high stakes for individuals, RP values higher than 67 may be too high (U.S. GENERAL ACCOUNTING OFFICE, 1993), and certainly RP values lower than 50 cannot be used to claim students have mastered the concepts tested by an item.

An item was considered to map to a particular EFL if the probability of success on the item was .50 (for the RP50 condition) or .67 (for the RP67 condition) for examinees whose proficiency estimates (θ) were within the specified EFL. Each item was considered to map to the lowest level where examinees had a probability of providing a correct response at the RP or higher. After items were mapped to the various EFLs, results were compared to the test developer's classifications of the items. An item was considered to match or align to the intended EFL if the item mapping results agreed with the test developer's classification. A situation where an item is mapped to an EFL other than intended was considered a mismatch and misaligned.

Reasons for curriculum-assessment misalignment

Seven teachers were convened for a one-day meeting to look at 20 items that did not map as intended to find potential reasons to explain the misalignment. Stratified random sampling was used to selected the misaligned items to ensure that items from all EFLs are represented. The teachers came from all geographical locations across Massachusetts and were drawn from current ABE teachers. Seventy-one percent of the teachers were female and the rest were males. All teachers were Caucasian with teaching experience ranging from 3.5 to 32 years. The meeting began with self-introductions of the participants followed by training that the facilitator conducted. The training sessions began with communicating the goal of the meeting, which was to review items that mapped to higher or lower EFLs than the test

developers had intended and suggest reasons for the misalignment. The teachers were then given a set of 6 items, which were used as practice items. The teachers looked at the items, the objective and level it was intended for and tried to find reasons why the item did not map to the intended level (i.e., why it mapped to an easier or more difficult location than it was written to). The teachers first looked at the practice set of items individually, which was followed by a group discussion.

After training, the teachers were split into two groups: one group analyzed the items that were misaligned using the RP50 criterion, followed by items that were misaligned using the RP67 criterion; the other group followed the opposite order. The items were presented in two booklets with one booklet presenting items that misaligned at RP50 first and RP67 last, while the second booklet had the opposite ordering. Each teacher was presented with an item review sheet on which they recorded their reviews. Group discussions of some of the items followed individual review of the items. A questionnaire was administered to evaluate the item review process. This questionnaire contained 5 Likert-type and 2 open response items. The Likert-type questions were rated on a 5-point scale from strongly agree to strongly disagree. In general, the survey sought teachers' views on aspects of the meeting such as adequacy of time for item review, adequacy of training and clarity of the item review task. The open-ended questions asked the teachers about some factors that they used in coming up with possible reasons for the observed misalignment and suggestions for the future.

Data analyses

Results were analyzed to assess the degree of agreement between item mapping results and intended EFLs for each item. Comparisons across the intended and item-mapped classifications were made at the item, content strand, and cognitive skill levels. The comparisons involved examining the agreement between the item mapping results and the intended classifications for each RP value (R50 and RP67). Chi-square tests and correlations were used to assess the degree of alignment. Content analysis (Gall et al., 1996) was used to analyze written accounts provided by teachers.

Results

Overall item mapping results

A review of all the misaligned math items at RP50 revealed that in general, misaligned items were slightly more discriminating and harder than the aligned items. The average discrimination and difficulty parameter estimates were 1.49 and 0.68 respectively for misaligned items versus 1.33 and -0.23 respectively for the aligned items. The average pseudo-guessing parameter estimate was 0.2 for both groups of items. This observation may imply that both the a- and b-parameters had an impact on alignment results. The misaligned items selected for review had similar average discrimination parameter estimates to all misaligned items (1.51 vs. 1.49). However, the reviewed items were much harder ($\bar{b}$ =1.15 vs. 0.68) and had slightly lower average pseudo-guessing parameter estimates (0.17 vs. 0.20).

Table 1 presents the overall classifications of the math items based on RP50 and RP67. It can be seen that exact agreement (proportion of items mapping to intended level) for the beginning basic level was 34% using the RP50 criterion, and only 17% using the RP67 criterion. The "correct" mapping locations for the other EFLs tended to be lower. In all but one case – Low Adult Secondary (the highest level of items evaluated in the set) – the RP50 criterion located more items in the intended EFL than RP67. The majority of items were mapped to higher difficulty levels across all EFLs for both RP50 and RP67.

Table 1 Math overall item mapping results for RP50 and RP67

Intended MAPT level	% items mapped to level based on RP									
	BB		LI		HI		LAS		HAS	
	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67
BB	**34.0**	**17.0**	37.0	33.0	24.0	31.0	5.0	17.0	0.0	2.0
LI	5.2	1.0	**28.9**	**9.3**	40.2	40.2	21.6	35.1	4.1	14.4
HI	0.0	0.0	11.7	3.2	**28.7**	**16.0**	33.0	28.7	26.6	52.1
LAS	0.0	0.0	2.8	0.0	16.7	2.8	**18.1**	**20.8**	62.5	76.4
Total	10.7	5.0	21.5	24.0	28.1	24.0	19.3	25.6	20.4	33.1

Notes: Shading indicates items mapped into intended levels. BB: Beginning Basic; LI: Low Intermediate; HI: High Intermediate; LAS: Low Adult Secondary; HAS: High Adult Secondary.

In looking at the RP50 results across all EFLs, the overall exact agreement between test developers' classification and IRT based item mapping at RP50 was 28.1%. This means only 28.1% of the items were mapped to the same level as intended. Combining exact and adjacent (items mapping to one EFL lower or higher than intended) agreement as a measure of overall agreement between test developers and item mapping classifications, overall agreement at RP50 was 77.5%. The highest adjacent agreement (84.8%) was obtained at the LAS level. The Spearman correlation between the RP50 classifications and intended classifications was 0.69, which is considered moderate based on Cohen's (1988) r2 criteria (r2 = .48). The chi-square test for these results was 234.66 (df = 15, p < .001) implying statistically significant differences exist between the PR50 item mapping results and the test developer's classifications of the items.

For RP67, the overall exact agreement between item mapping results and test developers classification was 15.4%, which is just over half the level of exact agreement for RP50. More items mapped to the LAS level or higher, relative to RP50. The highest exact agreement for RP67 was 20.8% at the Low Adult Secondary level. Only 36.6% of the items mapped to one EFL lower or higher based on the intended classification at RP67 compared to 47.1% for RP50. Overall agreement for RP67 was 59.5%. The highest adjacent agreement between the intended and IRT classification was 100% for the LAS level. The Spearman correlation between the RP67 classifications and the intended ELFs was 0.71 (r2 = 0.50), which was slightly higher than the correlation observed for RP50. Similarly, the chi-square results were statistically significant (χ^2_{15} =256. p < 0.001).

In summary, more congruence between the item mapping results and the classifications intended by the test developers was obtained at RP50. For both RP values, larger proportions of items map to one EFL higher than the EFL for which the item is intended. This may suggest that the items are generally harder than the test developers had anticipated.

Qualitative results: reasons for misalignment

Six broad categories pertaining to characteristics of items were derived from the reasons provided by the teachers during the study. The categories were: item difficulty, cognitive demand of the item, language level of the item compared to language level of the students, the type of math

concept being assessed, clarity of the item, and technical issues related to the item.

It was observed that the math concept being assessed in the item was a factor contributing to item difficulty misalignment in 13 items. The teachers noted that some mathematical concepts such as order of operations, calculating the mean in reverse order, finding the inverse, and math tasks involving mathematical symbols like greater than or less than, were generally harder for students. The teachers confirmed there were differences between the item developers' classifications and item mapping results due to some characteristics of the items that made them easier than intended. These characteristics included distractors that could be easily eliminated and familiarity of the scenario presented in the item.

The teachers identified cognitive demand of the item as a factor contributing to misalignment in 12 items. Most (9) items in this category asked students to derive and integrate new information into subsequent steps. The other items required students to extrapolate or perform multiple steps to arrive at the correct response.

With respect to why items may have been more difficult than expected, complexity of the language used in an item compared to reading level of the student was one factor that teachers suggested as contributing to misalignment in 11 items. Teachers noted that some items contained words that were hard for students at some EFLs and hence the poor performance on those items. For example, one teacher pointed out that reading and interpreting true/false statements was generally challenging for Beginning Basic students for whom English was a second language. Teachers noted that vocabulary such as doubling every minute, consistent, mean, inequality, average, perimeter, more than half, three times more, twice as often, and equivalent were hard for students to comprehend especially at the lower EFLs. The teachers also noted that some items contained long and complex sentences that required more sophisticated reading skills that students for whom the item was intended did not possess.

Eleven items were noted to exhibit some technical problems or ambiguities leading to students' poor performance. For example, in one item the stem

did not explicitly state that students needed to provide their answer in different units of measurement than the units in the stem. In another question, students were presented with a scenario where a fence needed to be put around a circular pond. However, the question did not specify that the fence also needed to be circular. Teachers cited lack of clarity of the item as a reason contributing to misalignment for 10 items. For instance, one teacher noted that in one item, students needed to reformulate the question to be able to answer it because the question was unclear. For one question, teachers noted the question was framed in such a way that it led examinees to carry out a wrong mathematical operation. Teachers also noted that presenting items in long sentences increased the likelihood of reducing the clarity of the item making the item become harder than intended. Similarly, teachers stated that some items contained information that was not necessary for students to respond to them and that may have led to confusion among some students.

A summary of the reasons teachers gave to explain why items were misaligned in terms of difficulty is presented in table 2. In addition to providing comments on individual items, teachers were also asked about the factors they considered in reviewing the item to generate possible reasons for misalignment. They cited language complexity, appropriateness of content for level of examinee, editorial errors in the question, and the number of steps required to solve the question. The teachers also mentioned the cognitive skill the item requires, the ability of examinees the item is intended to evaluate, and also the vocabulary used in the item as some of the factors they took into consideration.

Table 2 Summary of reasons for misalignment

Reason	Number of Items
Math concept assessed	13
Item difficulty	12
Cognitive demand	12
Language level	11
Technical issues with item	11
Item clarity	10

Discussion

This study was designed to illustrate how student responses to test items could be used to inform curriculum-assessment alignment. Item mapping based on IRT was applied to an adult basic education math assessment to illustrate the process. IRT was used to map the items in terms of their difficulty and the results were compared to test developers' classification of the items to evaluate the degree of agreement. SMEs (teachers) were used to help explain why some items were misaligned with respect to item difficulty.

The results of the present study indicate that some significant differences occurred between test developers' and item mapping classifications of the items. More items mapped to lower EFLs (that is High Intermediate or lower) at RP50 while more items mapped to higher EFLs (LAS or higher) at RP67. These results were expected because most of the items used in this study had c-parameter values that were less than 0.35. As such, the theta value at which students have a 50% chance of providing a correct response to an item (that is RP50) will always be less than the b-value. The only exception to this is when the c-parameter is equal to zero. On the other hand, the theta value at which students have a 67% chance of providing a correct response to an item will always be higher than the b-value. However, the assumption being made here is that the test developers took difficulty and discrimination of the item into consideration in classifying the items. The other assumption is that the test developers' estimation of the difficulty of the items for a particular group of learners was accurate. These assumptions are discussed later.

Results also showed that in general, greater alignment between test developers' and item mapping results was obtained at RP50 compared to RP67. These results are similar to results obtained by Kolstad et al. (1998). In their study aimed at evaluating the impact of RP value on selection of exemplar items for describing what students at a particular proficiency level could do, the authors found the greatest agreement between the percentage of items mapped along the proficiency scale and percentage of scores for examinees along the proficiency scale at RP50.

We used the degree of agreement between test developers' classifications of the items and item mapping results as a measure of a new kind of

alignment—alignment of item difficulty. This alignment evaluation is similar to the "level of challenge" criterion in the Achieve (2001) model, which uses subjective judgment. Alignment of item difficulty is important whenever a testing program uses a vertical scale across different grade levels. As the results of this study show, items do not always map to their intended levels. Such misalignment can only be discovered by analyzing students' responses to them.

Comparing agreement between test developers' classifications of the items and location of the items on the proficiency scale assumes that some common notion of difficulty was used in the two classifications. It is hoped that test developers consider not only the match between the item content and the level of the curriculum at which the content is taught, but also the relative difficulty of the item. As such, trustworthiness of test developer's ratings of the items for the intended group hinges upon their ability to accurately judge or estimate difficulty of the item for the target group.

Based on these results, it appears reasonable to state that test developers were not completely accurate in their estimates of the difficulty of the items. Results of this study closely match results obtained by Zwick et al. (2001), who investigated alternative item mapping methods for the NAEP. Zwick et.al asked SMEs to list the five easiest and the five hardest items from a test without ordering the items by difficulty within each set. They found that the SMEs difficulty rankings matched very closely to student's performance. Specifically, a Spearman correlation between the SMEs rankings and the proportion of 8th graders answering an item correctly was 0.65. Based on this correlation, Zwick et al. (2001, p. 22) concluded that the SMEs "rankings were substantially in line with the actual difficulty of the items" . Similar conclusions could be drawn about the math results obtained in the current study. Spearman correlations between test developers' and item mapping results were about 0.7 for both RP50 and RP67.

We also found that at RP50 greater exact agreement between test developers and item mapping results was obtained at the lower EFLs (i.e., High Intermediate level or lower) while the least was obtained at higher EFLs. Considering RP67, greater agreement between the two classifications was obtained at the higher EFLs compared to lower EFLs. These results imply that most items intended for lower EFLs mapped to low EFLs while those intended for higher EFLs did map to high EFLs. This finding also provides

some evidence that test developers made reasonably accurate judgments about items intended for lower EFLs and those intended for higher EFLs.

With respect to the SME teachers' reviews of the misaligned items, in general, items demanding higher levels of thinking were perceived to be more difficult. Analysis of the items that teachers identified as cognitively more demanding showed that they were those that the item writers classified as measuring evaluation and synthesis skills, which provides some validity evidence for the MAPT. Teachers identified difficult vocabulary, use of long sentences and excess verbiage as some factors contributing to misalignment. It was interesting to note that lack of student exposure to content was not a factor that contributed to low performance of examinees. It was the level of cognitive thinking the content in the item demanded that mattered most.

Test developers could improve alignment between intended and actual item difficulty by ensuring that language in the item matches language level of the students. This does not only improve the clarity of the item and student understanding but also eliminates construct irrelevant variance that could interfere with student performance. Another strategy would be to match cognitive demands of the item to cognitive capability of examinees. This would reduce the frustration and stress that might affect student performance in an item. Alignment can also be improved by ensuring that items are free from error. Items should be stated in simple language, and accompanying visuals should be well drawn and well labeled where appropriate. It is also important to ensure that the distractors are plausible, that is, they cannot be easily eliminated by less knowledgeable examinees or they do not offer clues to the correct response.

Implications of results

This study illustrated that the utility of alignment study results could be greatly enhanced if students' actual performance on the assessment can be taken into consideration. This would provide information on the strengths and weaknesses of the students and also inform teachers which areas of the curriculum need extra emphasis. Given that many large-scale assessment systems in K-12 education use assessments that are vertically equated across grades, the degree to which intended and actual item difficulty aligns is an important validity issue. The results of this study will help facilitate intended and actual alignment of item difficulty, and provide a method for helping evaluate it.

References

Achieve, Inc. *Measuring up* – a commissioned report on education assessments for Minnesota. Washington, D.C.: Author, 2001.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association, 2014.

Ananda, S. *Rethinking issues of alignment under No Child Left Behind*. San Francisco: WestEd, 2003a.

Ananda, S. Achieving alignment. *Leadership*, v. 33, n. 1. p. 18-21, 2003b.

Bhola, D. S. et al. Aligning tests with content standards: methods and issues. *Educational Measurement*: issues and practice, Washington, v. 22. p. 21-29, 2003.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Earlbaum, 1988.

Gall et al. *Educational Research:* An Introduction. 6th ed. New York: Longman Publishers, 1996.

Gomez, P. G. et al. Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, v. 24, n. 3, p. 417-444, 2007.

Hambleton, R. K. Enhancing the validity of NAEP achievement level score reporting. In: PROCEEDINGS OF ACHIEVEMENT LEVELS WORKSHOP. Washington, D.C.: National Governing Board, 1997.

Impara, J. C. & Plake, B. Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, v. 35, n. 1, p. 69-81, 1998.

Karantonis, A.; Sireci, S. G. The bookmark standard setting method: A literature review. *Educational Measurement*: issues and practice, v. 25, n. 1. p. 4-12, 2006.

Kirsch. I. et al. *Adult literacy in America*: A first look at the findings of the National Adult Literacy Survey. Washington, D.C.: National Center for Education Statistics; U.S. Department of Education, 1993.

Kolstad, A. et al. The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard? Washington, D.C.: American Institutes for Research, 1998.

La Marca, P. M. et al. *State Standards and State Assessment Systems*: A guide to alignment. Washington, D.C.: Council of Chief State Officers, 2000.

Plake, B. S. et al. Consistency of Angoff based predictions of item performance: Evidence of technical quality of results from the Angoff standards setting method. *Journal of Educational Measurement,* v. 37, n. 4. p. 347-355, 2000.

Plake, B. S.; Impara, J. C. *Ability of panelists to estimate item performance for a target group of candidates*: An issue in judgmental standard setting. Educational Assessment, v. 7, n. 2. p. 87-97, 2001.

Porter, A. C. Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, v. 31, n. 7. p. 3-14, 2002.

Ryan, J. J. Teacher judgment of test item properties. *Journal of Educational Measurement*, v. 5, n. 4. p. 301-306, 1968.

Shephard, L. A. Implications for standard setting of the NAE evaluation of NAEP achievement levels. In: JOINT CONFERENCE ON STANDARD SETING FOR LARGE SCALE ASSESSMENT. Washington, D.C.: National Assessment Governing Board, National Center for Educational Statistics, 1994.

Sireci, S. G. et al. *Massachusetts Adult Proficiency Tests technical manual*. Research Report n. 677, v. 2. Amherst, MA: University of Massachusetts, Center for Educational Assessment, 2008.

U.S. General Accounting Office. *Educational achievement standards*: NAGB's approach yields misleading interpretations. Report n. GAO-PEMD-93-12. Washington, D.C.: Author, 1993.

Wang, N. Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement*, v. 40 n. 3. p. 231-253, 2003.

Washington, D.C. U.S. Department of Education. Division of Adult Education and Literacy. Office of Vocational and Adult Education. *Implementation guidelines*: measures and methods for the national reporting system in adult education. Washington, D.C., 2006, July. Available at <http://www.nrsweb.org/foundations/implementation_guidelines.aspx>.

Zimowski, M. F. et al. BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items. Chicago: Scientific Software International, Inc., 1996.

Zwick, R. et al. An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement*: issues and practice, v. 20, n. 2. p. 15-25, 2001.

Leah T. Kaira
Doutora em Educação pela Universidade de Massachusetts Amherst, EUA
leahkaira@gmail.com

Stephen G. Sireci
Ph.D. em Psicometria pela Fordham University
Professor da Universidade de Massachusetts Amherst, EUA
sireci@acad.umass.edu