Article

LANGUAGE TESTING

Language Testing
2017, Vol. 34(3) 357–381

DOI: 10.1177/0265532216646605
journals.sagepub.com/home/ltj
SAGE

Evaluating different standard-setting methods in an ESL placement testing context

Sun-Young Shin and Ryan Lidster
Indiana University Bloomington, USA

Abstract

In language programs, it is crucial to place incoming students into appropriate levels to ensure that course curriculum and materials are well targeted to their learning needs. Deciding how and where to set cutscores on placement tests is thus of central importance to programs, but previous studies in educational measurement disagree as to which standard-setting method (or methods) should be employed in different contexts. Furthermore, the results of different standard-setting methods rarely converge on a single set of cutscores, and standard-setting procedures within language program placement testing contexts specifically have been relatively understudied. This study aims to compare and evaluate three different standard-setting procedures—the Bookmark method (a test-centered approach), the Borderline group method (an examinee-centered approach), and cluster analysis (a statistical approach)—and to discuss the ways in which they do and do not provide valid and reliable information regarding placement cut-offs for an intensive English program at a large Midwestern university in the USA. As predicted, the cutscores derived from the different methods did not converge on a single solution, necessitating a means of judging between divergent results. We discuss methods of evaluating cutscores, explicate the advantages and limitations associated with each standard-setting method, recommend against using statistical approaches for most English for academic purposes (EAP) placement contexts, and demonstrate how specific psychometric qualities of the exam can affect the results obtained using those methods. Recommendations for standard setting, exam development, and cutscore use are discussed.

Keywords

Bookmark method, Borderline method, cluster analysis, placement test, standard setting

In second/foreign language programs with multiple levels, appropriate placement is crucial to ensure that instruction is well suited to students' abilities and learning needs (Green, 2012). If placement procedures are unreliable, two types of misclassification

Corresponding author:
Sun-Young Shin, Indiana University, Bloomington, 1021 E. Third St., Memorial Hall 314, IN 47405, USA.
Email: shin36@indiana.edu

errors may increase: (1) "false positive" errors occur when students are placed into courses above their actual levels; and (2) "false negative" errors occur when students are placed into inappropriately low levels (Bachman, 2004). Both misplacement errors can lead to negative consequences such as inefficiency of teaching and learning and dissatisfaction among both teachers and students. Therefore, considerable effort is needed to minimize the likelihood that placement procedures result in these errors.

Placement procedures can take a number of different forms such as interviews, essays, multiple-choice placement tests, or a combination of methods. The means of evaluating and ensuring the reliability and validity of those procedures depend on their individual characteristics. In order to determine the levels into which students are placed, raters compare test-takers' performances to descriptions of *performance standards* – sets of characteristics that distinguish between categories of performance. With many assessments such as interviews and essays, students' behaviors are compared to explicit performance descriptors, often with fine-grained detail. For assessments that yield scores on their own scale such as multiple-choice placement tests, however, test scores must first be linked to performance standards by determining which scores correspond to category boundaries called cutscores: a process called *standard setting* (Cizek, 1996). The goal of standard setting is thus to operationalize performance standards with respect to the test's score scale, and this process has important consequences for the validity of test use (Cizek & Bunch, 2007). For example, if the cutscore for Level *x* is 80/100, a student who obtained 81 would be placed into that course, whereas a student scoring 79 would not. If the Level *x* cutscore were changed to 75 while others remained the same, however, then both students would be placed together, even though nothing about the nature of the test or either student's performance had changed. Therefore, crucially, even a test that produces perfectly consistent scores which perfectly reflect the "true" distribution of the construct in the test-taking population may still result in large numbers of false positive and/or false negative errors depending on how the cutscores are set.

It is thus important to evaluate the appropriateness of the standard-setting process. For this purpose, three types of validity evidence often serve as major elements for standard-setting evaluation: procedural, internal, and external validity evidence (Hambleton, Pitoniak, & Copella, 2012). Procedural evidence refers to the degree to which the performance standards and resulting cutscores are set by knowledgeable participants in a rigorous and systematic manner. The primary source of information about the procedural evidence includes detailed documentation of the process of selecting and training participants, defining performance standards, and setting the associated cutscores. Internal evidence for validity refers to the degree of consistency in the estimates of student and item characteristics relative to cutscores within each method; for example, a study using expert judges has stronger internal evidence for validity when there is more agreement among the judges on item difficulty levels or student ability levels. External evidence refers to the degree of correspondence between the results of different standard-setting methods as well as the degree to which those results align with other information about students' knowledge and skills.

In the psychometric literature (Cizek, 1996; Jaeger, 1989; Kane, 1994, 1998a), standard-setting methods are traditionally divided into two different approaches. With *test-centered methods* such as the Angoff method, Jaeger method, and Bookmark method,

experts analyze the test's *items* in order to judge the probable levels of performance of a minimally proficient student. In *examinee-centered methods* such as the Contrasting groups method and Borderline group method, experts identify *students* that exemplify the performance standards and then calculate cutoffs based on those students' performances on the test. An advantage of test-centered methods is that they result in a concrete description of what a minimally proficient student should be able to do on specific test forms, but these methods rely in part on a potentially tenuous assumption that judges can accurately determine the skill set and trait levels necessary to answer test items. In practice, student performance and test-taking strategies can differ radically from expert expectations (Alderson, 1993; Alderson & Kremmel, 2013; Mehrens, 1995). For examinee-centered methods, real students are used as prototypical examples of those who meet performance standards, but the risks, especially in programs with heterogeneous student populations, are that a variety of students with different test score profiles may all be considered minimally proficient for different reasons, or that students would exhibit some prototypical traits of the level while being idiosyncratically lower or higher in other areas. Furthermore, different judges may identify exemplar students differently (Kane, 1998b).

Both approaches thus have different strengths and weaknesses, which have led some researchers to call for using both together (Green, Trimble, & Lewis, 2003). However, perhaps unsurprisingly given the different focuses of test-centered and examinee-centered methods, the results of different standard-setting procedures usually differ, often by large margins (Hsieh, 2013a; Jaeger, 1989; Shin, 2004), and scholars do not agree on how to choose among the different standard-setting methods in educational assessment (Zieky, Perie, & Livingston, 2008). Indeed, although many different standard-setting methods have been compared and evaluated in a number of previous studies in the field of educational measurement (Zieky, 2001), these studies did not find evidence of convergence between resulting cutscores, nor did they provide support for the superiority of a single method (Alsmadi, 2007; Kane, 1994, 2001; Livingstone & Zieky, 1989; Näsström & Nyström, 2008). In part, this lack of agreement may stem from the nature of the task; the decision of what constitutes "high" or "low," "Level 2" or "Level 3," and so on, is not so much an empirical question as it is a policy decision based on the interests and desired outcomes of a program and its stakeholders. As Kane (1998a) noted in his seminal article on the subject:

> Neither the performance standards nor their associated cutscores exist 'out there' to be found; they must be created. As a result, there is no 'real' or 'true' value of the cutscore, no gold standard against which the results of a particular standard-setting study can be compared to determine its accuracy. (p. 137)

It is important to emphasize, however, that empirical support for standard-setting procedures remains central, especially in light of the increasing focus on program accountability within and outside academia. Camara (2013) reviews several examples of how empirical data in the form of post-test remediation rates, external test scores, and other methods have been used to inform judgments of both examinees in examinee-centered methods and items in test-centered methods. Recently, however, efforts have been made

to reduce the subjectivity of standard-setting methods yet further via a third approach, *statistical methods*, in which cutscores are set solely on the basis of test-internal statistical properties, usually by determining which cutscores, given the dependability of scores and their distribution, would result in the most homogeneous groups with the least likelihood of classification error. One such method that has been applied in standard-setting studies is cluster analysis (Sireci, 2001; Sireci, Robin, & Patelis, 1999). In cluster analysis, cutscores are computed such that students are categorized into maximally homogeneous groups called "clusters." While cluster analysis differs radically from other approaches in its independence from performance standards, researchers have provided procedural and internal evidence for validity and even its correspondence with cutscores derived from other methods (Shin, 2004; Sireci, 2001), suggesting the potential for greater application in language program contexts.

Standard setting in language program placement

There have been relatively few standard-setting studies in language-testing contexts. Several studies have, for example, linked standardized test scores to levels of the Common European Framework of Reference (CEFR) (Bechger, Kuijper, & Maris, 2009; Papageorgiou, 2010), determined IELTS® test cutscores for nursing program admission (O'Neill, Buckendahl, Plake, & Taylor, 2007), and set cutscores to match performance descriptors of the sixth-grade national English curriculum in Taiwan (Hsieh, 2013a, 2013b). Despite the burgeoning need for guidance in establishing placement cutscores in English for academic purposes (EAP) or other language programs, there has, to our knowledge, been only one standard-setting study conducted in a language placement testing context: Shin (2004) used three approaches – the Angoff method, the Borderline group method, and cluster analysis – to determine cutoffs for a newly developed web-based English placement exam used in an advanced academic ESL program. Shin's work contributes greatly by providing procedural and internal evidence to evaluate the different standard-setting methods. However, since students were not actually placed into the program based on test results, there was a lack of independent external evidence such as existing classification data to evaluate their success.

In addition to the need for more empirical data, theoretical work investigating methods for establishing the validity of standard-setting procedures is needed for language program placement. As mentioned previously, all standard-setting studies begin by agreeing upon a set of performance standards (Tannenbaum & Cho, 2014), and in language program contexts, the performance standards are usually operationalized as the student learning outcomes (SLOs) specified by the curriculum. However, SLOs specify *outcomes* of instruction, and not minimum requirements for *incoming* students, and so SLOs may not be appropriate as performance indicators for placement testing. It is not necessarily the case that "readiness for instruction" for a particular level will equate to mastery of the previous level's outcomes. In EAP contexts in particular, outcomes commonly address academic behaviors that incoming students are not expected to have mastered prior to enrollment, and which are not directly tied to general language proficiency (e.g. writing a multi-page paper with instructor feedback). Setting cutscores based on language program curricula thus requires making abstract inferences between language

proficiency and readiness for instruction on outcomes that are contextually embedded within academia, and these inferences have not been explored in detail.

Context and motivation of the study: Intensive English program placement

The current study investigated placement test cutscores for an intensive English program (IEP) housed in a large Midwestern university. The IEP has seven levels of instruction and enrolls approximately 150–225 students at a time, ranging from near absolute beginners to high-intermediate language users (TOEFL ITP® scores of approximately 550[1]). There are six instructional sessions annually, each lasting seven weeks. Students take four hours of core classes every weekday (Reading & Writing [2 hours], Communication [1 hour], and Grammar [1 hour]), and may, depending on the level, take a fifth class either in Extensive Reading or an elective course, for a total of 20-25 hours of instruction per week.

Students are initially placed into a level using a 3.5-hour placement test consisting of four sections. The Writing section consists of a 45-minute essay on an independent prompt (e.g. "Discuss whether you would like to live in a big city or in a rural area. Explain your reasons for your choice, giving specific details"). The prompt is considered "independent" because it does not require comprehension or interpretation of material other than the prompt to compose a response. Essays are scored holistically by two experienced teachers on a scale of 1–7, indicating the placement level, with a third rater in the case of disagreement. A previous internal program review found that a third rater was needed in fewer than one-third of essays not used for calibration, and that over 99% of disagreements involved discrepancies between adjacent levels, indicating relatively high inter-rater agreement. After the Writing section, the norm-referenced Listening, Grammar, and Reading sections consist of 36, 60, and 40 discrete, dichotomously scored multiple-choice items, respectively. The Listening and Reading sections include items testing comprehension of main ideas and details of aural and written texts of various lengths and difficulty levels, whereas the Grammar section consists of fill-in-the-blank items requiring knowledge of a large variety of grammatical structures. The Listening, Reading, and Grammar sections each have seven extant cutscores that were set many years before either author of the present study came to the institution. The methods and data used to determine these cutscores are no longer available, but they correspond to the minimum scores required to enter Levels 2–7 or be endorsed for program exit (students obtaining scores lower than the Level 2 cutscore are scored as Level 1). The average level of the scores for the Writing, Listening, Reading, and Grammar sections for each student is used for initial program placement. Owing to scheduling constraints, students take all courses in the same level; students may not, for example, take Level 3 Communication and Level 4 Grammar concurrently.

Since seven weeks' worth of intensive instruction is determined by placement test scores, and since in-class homogeneity is strongly desired for instructional outcomes, the placement exam can be seen as a medium-stakes test. To account for the possibility of misplacement, teachers may recommend reassigning students to a new level during the first three days of instruction if they believe it necessary. In practice, such re-placement

had historically occurred for less than 2% of incoming students. However, that figure probably underestimates the number of misplaced students since teachers have expressed reluctance to re-assign students for a variety of reasons. For example, especially for less experienced teachers, it is difficult to acquire enough information from formative assessments to make level change recommendations in the first few days of classes while students are still familiarizing themselves with each other, the teacher, and program expectations. Additionally, owing to its face-threatening nature, re-placing students into a level lower than that originally assigned is exceedingly rare. Finally, before the current study began, program administrators believed that the re-placement rate had risen somewhat as a result of a large influx of students from different geographic regions with different typical skill profiles compared to previous years.

Immediately prior to the current study, the IEP conducted a systematic review of the curriculum, simplifying and, in some cases, reassigning the student learning outcomes (SLOs) for each level and skill. SLOs explicate what students are expected to know and be able to do at the end of instruction in order to merit a passing grade and advancement to the next level. In that sense, SLOs define performance standards for each level of the curriculum, and given the extensive changes to them, it was necessary to review the cutscores that operationalize those performance standards, a process last undertaken more than 15 years prior. The motivation for our standard-setting study was thus threefold: large demographic changes in our program, changes to our curriculum, and the length of time since the last review of placement procedures.

In deciding on standard-setting methods, several options were considered. Large language programs such as ours have multiple levels, greatly increasing the already high cognitive load for judges participating in the standard setting, especially for test-centered methods such as the Angoff method. In Shin's (2004) study, for example, expert judges participating in the Angoff method portion of the study estimated the probability that students who minimally meet performance standards would respond correctly to individual items. Our placement test, however, has 136 total items, and seven cutscores each. Using the Angoff method would therefore involve making 952 response probability (RP) estimations, which is highly impractical, especially given the attested difficulty of estimating each individual RP (Lewis, Mitzel, & Green, 1999; Mehrens, 1995). There have been simplifications to the Angoff method such as the Yes/No Angoff, in which experts determine whether a borderline proficient student would be expected to respond correctly to an item more often than a certain RP or not (Hsieh, 2013b). However, in our context, even making that judgment presented a very daunting task to raters, and therefore there was a strong desire to simplify the judgments by making use of the accumulated item-level data and also to examine whether the psychometric properties of the test made it suitable for continued use, so the Bookmark method, explained below, was employed.

The current study in part methodologically replicates and expands on Shin's (2004) study, using the Borderline group method as an examinee-centered approach and cluster analysis as a statistical approach, but we employ a different test-centered method, the Bookmark method, and compare the results of each method to extant cutoffs already used on enrolled students. Kane (1994, 1998a) argued that standard-setting methods should be evaluated in terms of the degree of appropriateness and defensibility of the

process in a specific context, and that procedural, internal, and external evidence should be provided to defend the choice of specific standard-setting method. We thus sought such evidence for our language testing context, and organized our study around two research questions:

1. Do the cutscores derived from different standard-setting methods correspond to each other and/or extant cutscores in an ESL placement testing context?
2. In the event that cutscores differ, to what extent does each standard setting method provide evidence of procedural, internal, and external validity evidence that would support the use of particular cutscores over others?

Method

Placement test data

We collected IEP placement exam scores from 538 entering students (228F, 310M) over the course of a two-year period. The students came from a mix of different language backgrounds, with the largest six groups speaking Korean (n = 188), Arabic (n = 130), various languages of China (n = 124), Japanese (n = 32), Spanish (n = 17), and Portuguese (n = 13), while others spoke Kazakh, Turkish, Russian, Hungarian, Uyghur, Thai, Mongolian, Vietnamese, and Bambara. As mentioned earlier, the exam consists of four sections, but the Writing section is graded holistically and averaged with the level scores of the remaining three sections for final placement decisions, so it was not part of the standard-setting studies. Nevertheless, students' Writing scores were collected for the last session of the study (covering 35 students) to compare the effects of cutscore changes on final placement. Students' scores on each item of the remaining three sections – Listening (36 items, α = .79), Grammar (60 items, α = .93), and Reading (40 items, α = .89) – were used for the Bookmark method and cluster analysis. Overall scores (out of 136, α = .96) ranged from 21 to 129, and at least five students were placed into each classification level ("Level 1" through "Program Exit"), both overall and within each individual skill. More detailed descriptive statistics are given in Table 1.

Bookmark method

We chose the Bookmark method as this study's test-centered approach; a discussion of the rationale behind that decision follows. Until the mid-1990s, the Angoff method, in

Table 1. Scores for the placement test (N = *538* for all sections).

	Listening	Grammar	Reading	*Total*
k	36	60	40	136
Score range	5–32	7–60	3–40	21–129
Mean (*SD*)	17.36 (5.85)	33.82 (12.52)	24.62 (7.99)	75.81 (24.34)
Cronbach's α	.79	.93	.89	.96

which judges examine the test's items and estimate the probability that hypothetical marginally proficient examinees would answer each item correctly, was the most popular test-centered approach to standard setting. From 1995 to 2000, however, a confluence of events changed that situation. As mentioned above, it was found in several studies that judges, especially those without psychometric training, had difficulty estimating correct response probabilities (RPs), and were frequently inaccurate in their estimations of both absolute and relative levels of item difficulty (Lewis, Mitzel, & Green, 1996; Mehrens, 1995). Second, the number of states with codified performance standards for English/Language Arts increased from 20 to 49 (Council of Chief State School Officers, 2000), and there was a concurrent push for increased empirical support for standard-setting methods.

The Bookmark method was developed in an attempt to address at least four specific perceived needs in test-centered standard setting: (1) simplification of the cognitive tasks required of expert judges; (2) more empirical data to support expert judgments; (3) accommodation of multiple cutscores on a single test; and (4) accommodation of multiple item types (e.g. selected and constructed response) (Karantonis & Sireci, 2006; Peterson, Schulz, & Engelhard, 2011). The key feature of the Bookmark method is that, rather than having experts estimate the relative difficulty level of each item, item difficulty is calculated with Item Response Theory (IRT) techniques, based on how actual students have performed. The items are then arranged in order, one per page, in an ordered item booklet (OIB), and judges then read through the OIB one page at a time, placing a "bookmark" at the point where a minimally proficient student at a given level would no longer be expected to answer correctly at the chosen RP. Since IRT estimation requires large sample sizes, the Bookmark method is not always feasible in language program contexts, especially for newly developed tests, but in our particular case, there was sufficient performance data to create the OIB. Still, there are several possible ways to do this, so it was necessary to determine the specific set-up best suited to our context.

First, an RP value must be chosen to represent "mastery." The ability level (θ) required to reach that RP is then used to order the items in the OIB. Setting the RP value has historically been the most contentious aspect of the Bookmark method, with some researchers using .50 and others using .67 or even .80 to represent mastery (see Peterson, Schulz, & Engelhard, 2011 for a review). This choice is crucial because, if item discrimination parameters are allowed to vary, differences in RP could result in different rank orders, and even when discrimination parameters are held constant as in the Rasch model, for example, Wyse (2011) showed how the choice of RP value can still affect the eventual cutscores generated. In a language placement context, an RP of .50 is not easily interpreted as "mastery," but requiring .80 would reduce item information available at that item, and is difficult for judges to interpret (Karantonis & Sireci, 2006). In addition, Reckase (2006a, 2006b) used an RP of .50 in several simulation studies and found that the Bookmark method may result in systematically underestimated cutscores. The value .67 matches well with notions of mastery in an instructional context, Huynh (2006) argues that .67 maximizes item information for correct responses, and judges in previous studies found it relatively easy to evaluate whether a marginally proficient student could correctly respond to a given item 67% of the time (Lewis et al., 1999), so we used an RP value of .67 to order items.

The second step for Bookmark method studies concerns the choice of IRT model and establishment of model fit. As mentioned before, the choice of model – for example, between a Rasch or 1PL as used by Wang (2003), or a 3PL as recommended by the Bookmark method's original designers (Lewis, Mitzel, & Green, 1996) – can influence both the order of items in the OIB and subsequent cutscores (Beretvas, 2004). Our position is that the IRT model used should be the one that best fits test performance data. While advocates of the Rasch model may state that items that do not fit Rasch assumptions should be modified or discarded in favor of ones that do, such concerns are relevant primarily during the stage of test *development*. In a standard-setting context such as ours where the test is already in use, measurement of the item characteristics as they exist is paramount.

We used BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) to estimate item parameters and test model fit. We first examined dimensionality of the test sections in order to ensure the appropriateness of using a unidimensional IRT model. A factor analysis provided evidence that each section of the test was probably unidimensional (the first eigenvalues were more than 10× the value of the second for each, and the elbow of the scree plots bent at the second eigenvalues). Next, 1PL and 2PL models were fit to each of the three sections individually. A negative 2log likelihood ratio test (Hambleton, Swaminathan, & Rogers, 1991) showed that the 2PL models fit statistically significantly better than the 1PL estimations for all three sections. Subsequently, the 3PL model was tested, but unfortunately, stable estimates were not obtained owing to our limited sample size and low discrimination parameters for several items. For the 2PL model, 135/136 items showed non-significant chi-square values for tests of misfit, indicating that, for all but one item, the 2PL parameter estimations satisfactorily fit the data. These parameter estimates were converted into item locations with RP = 0.67 using the formula given by Beretvas (2004, p. 39) as follows:

$$\theta_{2PL} = \frac{1}{D\hat{a}} ln\left(\frac{RP}{1-RP}\right) + \hat{b}$$

In this equation, D is a scaling constant (1.7), $\hat{a}$ is the estimated discrimination parameter, and $\hat{b}$ is the estimated difficulty parameter. Item difficulty ranged widely from $-3 < \hat{b} < 5$. As for the misfitting item, Item 60 from the Grammar section was by far the most difficult item on the test, answered correctly by fewer than 12% of the students in our sample. A simple point-biserial correlation with total scores on the Grammar section was −0.27, indicating exceedingly poor discrimination. Since this item would have been the last item in the Grammar OIB regardless of the IRT model used, the item was placed in that position in the OIB, and the 2PL model was run again with that item excluded, resulting in negligible changes in $\hat{b}$ estimates and no changes in rank order of item difficulty.

Once the OIBs were constructed, we recruited six experienced teachers from our IEP to make two teams of judges.[2] Four teachers (1M, 3F), two in each team, held the position of "Master Teacher," meaning that they had taught in the program for at least 10 years and held various program responsibilities. The remaining teachers in each team, both female, had taught in our program for two years and six years, respectively, but had

extensive outside experience as well. These teachers were recruited in the hope of diversifying the perspective of the expert panel. All teachers had an MA in TESOL or Applied Linguistics, and all were involved in regular professional development work in the program and through regional TESOL organizations.

In the first stage, each team met separately. The teachers were instructed to read individually through each OIB in order, without going back, and mark on a worksheet the first item where a student "ready to be placed into Level 1" could no longer be expected to answer correctly two-thirds of the time with a bookmark. This bookmark effectively represented the lower cut-off for Level 2. Teachers were given a short form of the curriculum, and were instructed to refer to it by a researcher present in the room in order to ensure that teachers were basing decisions on curricular performance descriptors. The teachers then continued through the rest of the OIB to mark the remaining bookmarks for each of the three sections, resulting in 21 total decisions (three tests × Levels 2–7 and program exit). After a short break, the three teachers then discussed their bookmarks together, and negotiated team bookmarks. During this time, in the event of disagreement, the researcher asked raters to justify their decisions in terms of the performance indicators. Their discussions during this second stage were recorded in order to collect procedural evidence for validity. The third stage occurred several days later in order to encourage reevaluation of existing bookmarks and reduce the threat to face in cases of disagreement, since our program's teachers work together commonly unlike many Bookmark method studies in general education. During the third stage, the two teams came together, reread the OIBs and curricular documents, and were given the teams' anonymized bookmarks and instructed to make final bookmark decisions.

Once bookmarks had been obtained, an ability level (θ) corresponding to a .67 RP for the bookmark item was calculated using the item location formula above. The cutscores were determined by calculating the maximally likely test score for a borderline student with that ability level. That is to say, for all test section items j, the cutscores ($\widehat{X}$) were calculated as the summed item probabilities for a student with θ at the borderline, as follows:

$$\widehat{X} = \sum_{j=1}^{J} \frac{e^{Da_j\left(\theta_{borderline} - b_j\right)}}{1 + e^{Da_j\left(\theta_{borderline} - b_j\right)}}$$

where e is the base of the natural logarithm (≈ 2.718), a and b are the item discrimination and difficulty parameters, respectively, and D is the scaling constant (1.7). For the Reading test, however, even the very first item was considered too difficult for Level 1 students. Setting θ to the value corresponding to a 0.67 RP on the first item risked overestimating the cutscore owing to the possibility that a hypothetical easier item not included on the current test form also would be marked as a bookmark. Therefore, after consulting with the teachers, θ for entry into Level 2 was set to a 0.50 RP on the first item for the Reading section.

Borderline group method

In the Borderline group method, the current study's examinee-centered approach, experts identify "borderline" students who are either just barely proficient for the level in which

they are enrolled, or just barely not proficient enough to enroll in the next level. The median score of borderline students is then used as the cutscore for each level (Shin, 2004). Among examinee-centered approaches, the Borderline group method is often used in contradistinction with the Contrasting groups method, where instead of borderline students, "prototypical" students who clearly meet the performance standards for each level are identified by teachers or other external means. The means of deriving a cutscore from the distributions of "prototypical" student scores is the matter of some debate (Berk, 1986; Kane, 1998a; Livingston & Zieky, 1989), but the goal is to calculate a cutscore that maximizes differences between the distributions. The Contrasting groups method was, in fact, the method chosen in the mid-1990s to determine our program's extant cutscores, using pilot results from the population of students in our IEP at that time as source data, although no record of the exact method used to calculate cutscores was available at the time of the study.

In previous studies, the Borderline group method has been more commonly employed than the Contrasting groups method, with researchers citing its conceptual simplicity on the part of judges (Jaeger, 1989; Hambleton & Pitoniak, 2006). However, a potential drawback of both methods is that teachers commonly identify only a small number of borderline or prototypical students (Zieky, Perie, & Livingston, 2008), while examinee-centered methods typically require large numbers of students before cutscore estimates stabilize (Tannenbaum & Cho, 2014). Researchers have also found that there is often large variation in the test scores of borderline students, and even larger variability and overlap in distributions for prototypical students (e.g. Berk, 1986), calling into question the stability of cutscores. For the Borderline group method, the degree of (in-)stability can be estimated as the standard deviation of test scores from the borderline groups compared to the standard deviation of the total group (Zieky, Perie, & Livingston, 2008), but there are no simple fixes for estimates that are found to be unstable; rather, researchers would need to collect more borderline student data, and there is the risk that estimates will not stabilize even with larger *n* (e.g. McLaughlin, 1993). It has been suggested that such inconsistency results potentially from teachers' lack of familiarity with their students' skills or from the possibility that teachers would judge students using criteria unrelated to performance descriptors (Livingstone & Zieky, 1989).

In our study, throughout the period of placement test data collection ($N=538$), within the first two weeks of each session, a total of 21 IEP instructors were given a modified roster asking them to classify their students as either "borderline up" or "borderline down," or to state that they had no borderline students. Teachers were instructed to identify borderline students based on their performance relative to the curricular SLOs.

Cluster analysis

Finally, in our statistical method, using SPSS 20 (2011), hierarchical cluster analysis (HCA) using Ward's (1963) method was first conducted to obtain the optimal number of clusters based on Z-scores from the Listening, Grammar, and Reading sections separately. A six-cluster solution was selected as the optimal number of clusters for each section respectively, since a large change was observed between the sixth and fifth cluster solutions in the agglomeration coefficients from each section. The Q-cluster solution, where Q stands for the optimal number of clusters derived from HCA, provides an

estimate of the minimum (Q−2) and maximum (Q+2) number of homogeneous clusters (Sireci, Robin, & Patelis, 1999). Thus, 4–8 cluster solutions were obtained from our data.

Next, we ran a *k*-means cluster analysis, again using SPSS 20 (2011), to form a fixed number of clusters. The HCA results indicated that the number of clusters should be between four and eight, and since our placement test divided students into eight groups, the number of clusters was fixed at eight. The eight-cluster solutions were applied to Listening, Grammar, and Reading test scores, and one-way analyses of variance (ANOVAs) were conducted to verify statistically significant differences in test scores among clusters, because clusters are sometimes formed when no true clusters exist in the data (Sireci, 2001). The ANOVAs showed that scores from the clusters formed from each subsection are overall significantly different (Listening: F (7,530) = 389.67, $p < .001$, $\eta^2 = .84$; Grammar: F (7,530) = 497.63, $p < .001$, $\eta^2 = .87$; Reading: F (7,530) = 471.93, $p < .001$, $\eta^2 = .86$). However, subsequent Tukey post-hoc comparison results revealed that not all pairwise comparisons were significantly different in Listening and Reading test scores. One clustering from the Listening section and two clusterings from the Reading section separated students whose test scores were not significantly different. Thus, a seven-cluster solution for Listening and five-cluster solution for Reading were applied by using the same cutscore for both Levels 3 and 4 in Listening and Reading and Levels 5, 6, and 7 for Reading. In effect, this meant that Level 3 on both tests and Levels 5 and 6 on Reading were removed as possible scores in this method.

Once students were grouped into homogeneous subsets, logistic regression was used to identify the test score that best segregated students from different clusters, following Sireci, Robin, and Patelis (1999). Livingston and Zieky (1989) first used this method to compute cutoffs from Contrasting group method data, and Sireci, Robin, and Patelis (1999) applied logistic regression to identify the cutscore associated with a .50 probability of a student being classified into different clusters using the following probability function:

$$p = \frac{1}{1 + e^{-(a+bx)}}$$

where e is the base of the natural logarithm (≈ 2.718), and the a and b parameters are the intercept and slope obtained through logistic regression. Cutscores were then derived by setting $p = .50$ and solving for x for each level and test section.

Results

The results of three different standard-setting procedures – the Bookmark method (a "test-centered" approach), the Borderline group method (an "examinee-centered" approach), and cluster analysis (a statistical approach) are presented. The cutscores obtained from each method were then compared to the extant cutscores used for placement in the program.

Bookmark method

During the bookmark placement process, both teams expressed considerable difficulty with the task, not because of features of the Bookmark method itself so much as features

of this particular test and the OIBs. In particular, experts felt that there were very few low-level items. Indeed, as mentioned, there were no Level 1 Reading items and only one Level 1 Grammar item on the entire test, and over half of the Reading test used items that were past the bookmark for Level 7 entrance.

In addition, the order of items in the OIB often clashed strongly with expectations, prompting the experts to adopt new strategies. For example, some experts made "tentative" bookmarks, and read ahead to see whether the next few items were also noticeably higher before deciding on where to place bookmarks. Experts did so despite explicit (repeated) instructions not to work backwards and to base their decisions solely on the item presented on each page. One set of consecutive items from the Grammar OIB exemplifies the trouble in the task. According to the results of IRT estimation, Items G20, G21, and G22 gradually increase in difficulty level from one to the next, but experts were unable to come up with a plausible explanation for this occurrence.

Item G20. If there _____ so much traffic yesterday, I would have arrived on time.

a. has been
b. hasn't been
c. had been
d. hadn't been

Item G21. Eric studied Japanese _____ four years.

a. since
b. before
c. for
d. until

Item G22. The car _____ was very luxurious.

a. what she wanted
b. what she wanted it
c. which she wanted it
d. which she wanted

It was suggested at first that Item G21 might be more difficult than Item 20 because prepositions are difficult in general, but this would not explain why both were easier than Item G22. One suggested explanation for Item G22's difficulty was that the L1 for Arabic speakers, who make up a large percentage of the program's population, has obligatory resumptive pronouns, possibly leading them to choose the incorrect options B and C more frequently, and inflating the difficulty level overall. Thus, there was the possibility for differential item functioning (DIF) effects. Perhaps because of these complex results, the experts focused heavily on other features of the items or the test task, such as the length of the prompt or low frequency vocabulary words as potential sources of unexpected difficulty or ease. Several experts admitted to adjusting bookmarks from their initial impression in order to have any items in a level at all, and five out of the six

experts decided to give "alternate" ratings (e.g. "could be Level 3 or 4") during the individual stage, again despite explicit instructions not to do so. Similar to the experts in Hein and Skaggs's (2009) study, expert strategies shifted between different rounds of bookmark setting.

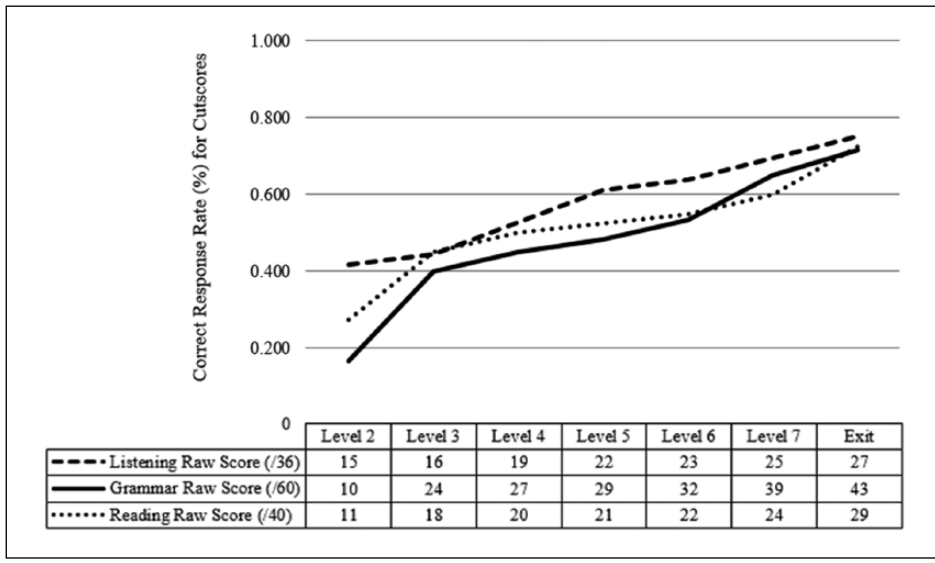
In terms of the reliability of ratings, methods for calculating agreement rates have not been standardized in the literature, but absolute agreement between the three experts in terms of the location of bookmarks was low (ranging from 10.0% absolute agreement for Team A on Reading to 52.7% absolute agreement on Listening). When examining the actual bookmark items, however, it becomes clear that absolute agreement underrepresents the similarity in judgment. For example, on Team B, all three raters listed the same six items as bookmarks on the Listening test, but one of the raters was one level lower on all but one of them, greatly decreasing absolute agreement. Similar patterns were seen for Team A's bookmarks, indicating that experts were more easily able to identify "jumps" in difficulty level, but did not always agree on the absolute level of those jumps.

In such cases of disagreement, there was an overwhelming tendency to defer to the most experienced or highest ranking member of the team. Indeed, all 21 initial bookmarks initially suggested by Team B's most senior member were used as the final bookmarks for that team, and such was also the case for 16/21 bookmarks for the senior member of Team A. In terms of the discussion, though, there were no instances of appeals to authority, nor any implicit or explicit commands. Instead, this tendency also may be the result of two other factors: in over 90% of the cases of disagreement resolution, teams resolved in favor of setting the higher level bookmark earlier in the OIB (which happened also to be the bookmark set by the senior members), and it also happened to be the case that 39/42 total bookmark decisions resolved in favor of the 2/3 majority of the team. With our limited sample, distinguishing among these three potential contributing factors in Bookmark method judge behavior is not possible.

The final cutscores derived from the Bookmark method are presented in Figure 1. Of concern, the resulting cutscores for several adjacent levels are very close to each other, sometimes separated only by a single item. Additionally, more than a quarter of the total test score range lies above the cutoff for program exit, which is interpreted by the program as indicating that the students do not need instruction in our IEP.

Borderline group method

In total, 14 among 21 instructors identified 65 borderline students from their classes, the distribution of which is shown in Table 2. Of note, most teachers in most sessions either did not identify any borderline students, or those students were not newly enrolled and thus were not included in the study since their enrollment at the time was no longer dependent on their initial placement exam scores. No students were identified as borderline above Level 7 (i.e. "Program exit") at all. For all other levels, the median test scores of the identified borderline students were used as the cutscores for each level. In the case of half points, the effective cutscore was the next number rounded up; for example, the cutscore of 15.5 indicates that a student scoring 16 or higher would be scored Level 4 in Listening. As shown in Figure 2, the cutscores are fairly dissimilar to those of the



	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Exit
Listening Raw Score (/36)	15	16	19	22	23	25	27
Grammar Raw Score (/60)	10	24	27	29	32	39	43
Reading Raw Score (/40)	11	18	20	21	22	24	29

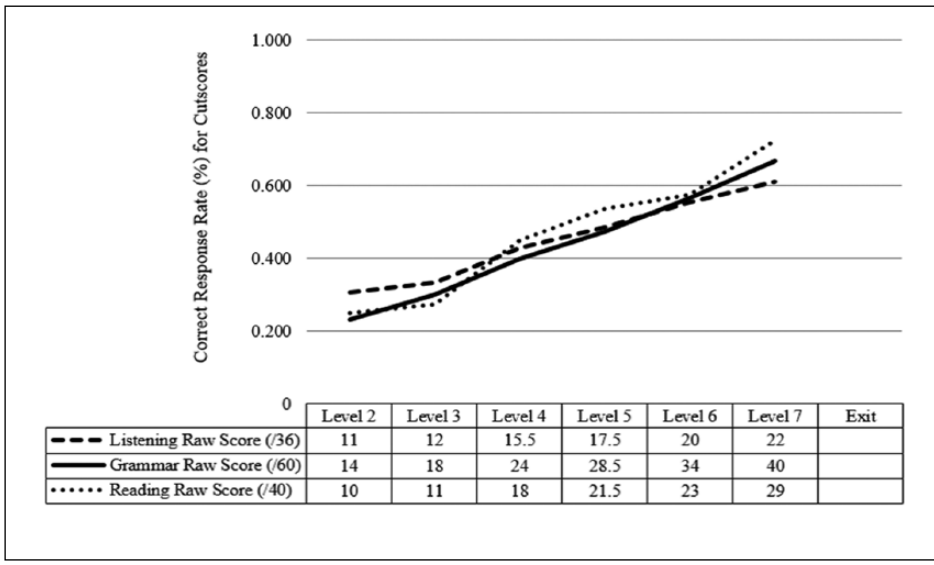
Figure 1. Cutscores derived from the Bookmark method.

Table 2. Standard deviations (*SD*s) of borderline groups around the cutscores from the borderline method at each level.

Level	Listening cutscore (*SD*)	Grammar cutscore (*SD*)	Reading cutscore (*SD*)	*n* of Borderline students
2	11 (1.52)	14 (2.65)	10 (1.92)	5
3	12 (2.75)	18 (5.93)	11 (2.39)	8
4	15.5 (4.37)	24 (4.65)	18 (4.13)	16
5	17.5 (4.28)	28.5 (4.65)	21.5 (3.07)	10
6	20 (4.33)	34 (5.72)	23 (2.83)	17
7	22 (4.60)	40 (7.45)	29 (7.12)	9
Exit		No cutscores available		0
Overall *SD*	(5.85)	(12.54)	(7.99)	

Bookmark method for Listening and Reading, and more widely distributed for Grammar. One similarity between the results, however, is that most of the high range of scores lies above the highest cutoff available, providing some convergent evidence that the difficulty levels of much of the test lies above the range of our incoming student population and does little work to separate students by level.

In addition, the standard deviations of the scores of borderline students were calculated as a measure of the stability of the cutscores. As shown in Table 2, many of the distributions of scores overlap considerably, and this remains an issue even (indeed, especially) for Levels 4–6 where there were larger numbers of students identified.



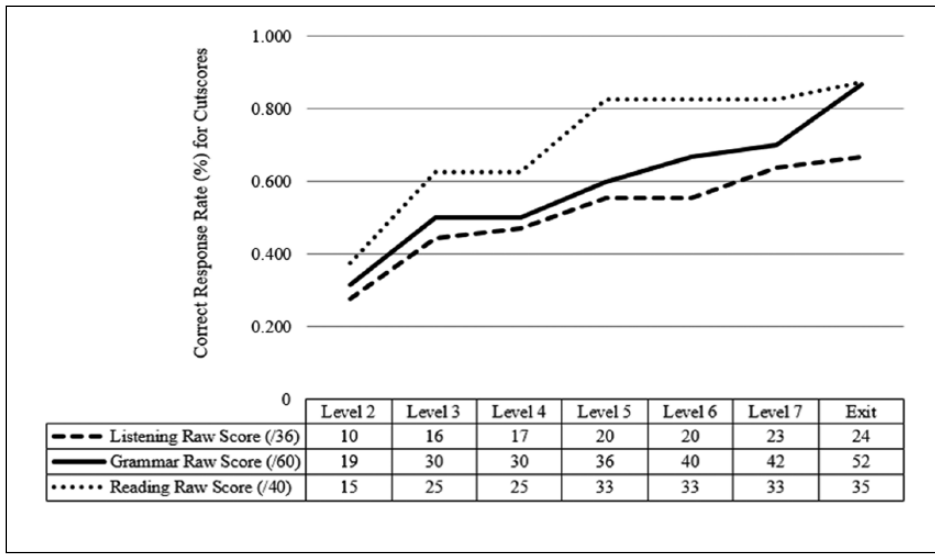
	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Exit
Listening Raw Score (/36)	11	12	15.5	17.5	20	22	
Grammar Raw Score (/60)	14	18	24	28.5	34	40	
Reading Raw Score (/40)	10	11	18	21.5	23	29	

Figure 2. Cutscores derived from the Borderline method.

When responding to the study, teachers indicated difficulty in separating between students who were "borderline" from those who were simply "misplaced." This was especially difficult for new teachers in the program whose experience with adjacent levels was very limited or null. Teachers stated that they were confident about the rank order of their students within the class, but it was much more difficult to tell whether their entire class was high or low relative to the performance standards, an observation also made by Wall, Clapham, and Alderson (1994). Another difficulty expressed by teachers was that students who had haphazard attendance or homework submission rates were harder to assess or sometimes unassessable. Finally, the choice of performance descriptors was an issue. Many teachers repeated our observation that the curriculum specifies instructional outcomes, but that judgments about incoming students' placement are not based on those outcomes. Rather, whether or not a student needs to be re-placed is a question of whether or not the student is ready to receive instruction on the outcomes given a level of general proficiency, and that is left to teachers' largely subjective judgments.

Cluster analysis

The cutscores resulting from the cluster analysis are listed in Figure 3. Notably, the cluster analysis places cutoffs such that students must attain a higher percentage of correct responses on the Reading test compared to the other two sections, with more lenient RP requirements for Listening. In contrast, the Bookmark method had found much the opposite pattern of results, with percentage requirements highest for Listening at nearly all levels. Similar to the findings of the other methods, cluster analysis had difficulty in separating between upper-intermediate students, but unlike with the other methods, in



	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Exit
Listening Raw Score (/36)	10	16	17	20	20	23	24
Grammar Raw Score (/60)	19	30	30	36	40	42	52
Reading Raw Score (/40)	15	25	25	33	33	33	35

Figure 3. Cutscores derived from cluster analysis.

the cluster analysis, Level 5 and 6 Listening, and Levels 4–6 Reading were not statistically significantly different from each other at all. In sum, there was not sufficient spread between levels of incoming students – at least, as indicated by test scores on the placement exam – in order to derive the same number of significantly different clusters as levels in our program.

Overall results

Table 3 demonstrates how the results of all three methods differ significantly from each other and from the extant cutscores, the latter of which are much more uniformly distributed throughout the test score ranges. This replicates previous findings that showed divergent results for different methods. Table 4 quantifies the average divergence from the extant cutoffs for each of the methods, showing that, overall, the Borderline method achieved results closer to the extant scores for the Listening test, and the Bookmark method achieved closer results for the Grammar and Reading tests. That is not to imply that the Bookmark method's cutscores or the extant cutscores are somehow more "accurate," as true, unbiased cutscores cannot be known, but rather simply to show that adjusting cutscores to those of the Borderline method or cluster analysis results would constitute an overall larger change.

In order to gauge the impact of these potential cutscore modifications, Writing exam scores also were collected for the final session of the study. For that session, 35 students in total were placed according to the extant cutscores. The students received Writing test scores ranging from Level 1 to Level 6, and placement test scores ranging from 2 to 113 out of a maximum 136 (the student scoring 2 points was an absolute beginner who

Table 3. Combined cutscore results.

Section	Level	Bookmark cutscore	Borderline cutscore	Cluster analysis cutscore	Extant exam cutscore
Listening	2	15	11	10	7
	3	16	12	16	11
	4	19	15.5	17	15
	5	22	17.5	20	18
	6	23	20		21
	7	25	22	23	25
	Exit	27	n/a	24	30
Grammar	2	10	14	19	13
	3	24	18	30	21
	4	27	24		25
	5	29	28.5	36	29
	6	32	34	40	38
	7	39	40	42	46
	Exit	43	n/a	52	52
Reading	2	11	10	15	11
	3	18	11	25	15
	4	20	18		19
	5	21	21.5	33	22
	6	22	23		25
	7	24	29		29
	Exit	29	n/a	35	35

Table 4. Average distance of cutscores from extant cutscores by method.

Section	Bookmark	Borderline	Cluster
Listening	3.83	0.17	1.50
Grammar	−1.83	−2.25	4.17
Reading	−0.83	−1.42	7.17

stopped taking the exam early into the Listening section). We calculated the levels into which those 35 students would have been placed if, instead of the extant cutscores, the cutscores derived from the three methods were used; this meant averaging their level assignments on the Listening, Grammar, Reading, and Writing sections. The results show that between 14 and 15 students would have been placed differently depending on the method, as summarized in Table 5. In contrast, only two of those students did, in fact, get re-placed levels after teacher recommendations. Both of those misplaced students would have been placed in their eventual level had our program used the Bookmark method cutscores, one of them using the Borderline cutscores, but neither using the cluster analysis scores, in part because the students were placed into Levels 3 and 4 where clusters are indistinct for some skills, and in part because both students were re-placed

Table 5. Comparison of placement decisions of standard-setting methods with extant decisions.

	Bookmark	Borderline	Cluster
Students who would be placed higher	9 (25.7%)	14 (40.0%)	3 (8.6%)
Students who would be placed lower	5 (14.3%)	1 (2.9%)	11 (31.4%)
Average difference in level assigned	0.176	0.382	−0.265

Table 6. Average $\Phi(\lambda)$ for cutscores by method.

Section	Bookmark	Borderline	Cluster
Listening	0.830	0.825	0.829
Grammar	0.942	0.948	0.943
Reading	0.902	0.923	0.924

into higher levels. As can be seen in Table 5, the cluster analysis cutscores tended to place people lower than the extant cutscores. No other students were re-placed, indicating low agreement between the three methods' proposed changes and the teachers' expressed need to alter placement decisions.

There also was a need to evaluate the dependability of those cutscores. As a quantitative measure of dependability, $\Phi(\lambda)$ (Brown, 2013) was calculated for each of the 21 cutoffs. $\Phi(\lambda)$ ranges from 0 to 1, with higher numbers indicating greater dependability. However, in part because of the relatively large number of items, $\Phi(\lambda)$ was high and overall similar for cutscores obtained through all three methods, as summarized in Table 6. Dependability, at least as measured by $\Phi(\lambda)$, was insufficient to discriminate between methods.

Discussion

The answer to our first research question is straightforward: as predicted, the three standard-setting methods we employed did not converge either to a single set of cutscores, nor to the extant cutscores. Using the cutscores derived by the methods in this study would have resulted in different initial placement for about 40% of students matriculating in the final session of the study's duration, despite the fact that average absolute differences between extant cutscores and method-derived cutscores were, for the most part, small. This underscores the sensitivity of placement decisions to cutscores and the importance of rigorous standard-setting methods, but on its own, the fact that scores differ does not help in deciding which scores to use.

While the extant cutoffs have historically resulted in few re-placements, Wall, Clapham, and Alderson (1994) give two general cautions against interpreting re-placement as evidence for or against external validity beyond those stated for our specific program above. First, they state that teachers strive to make students feel that they belong in the classroom, regardless of their initial levels relative to performance descriptors. Second, teachers typically only have extended contact with their own students, and thus

could not know whether other students, with different test scores and levels of language ability and test scores, also should be placed in the same class. The current study found ample support for these cautions, but we would like to emphasize the nature of performance descriptors for placement, as well. Namely, the curriculum does not specify the general proficiency levels expected for initial entrance, but only which abilities students are expected to acquire by program exit. Thus, although the Bookmark method and Borderline method cutscores had better correspondence with re-placement decisions than the cluster analysis results, internal and procedural evidence for validity seem more valuable for method comparison. In order to examine external evidence for validity further, future studies could, for example, collect data from a range of students on other performance assessments and evaluate those with respect to performance descriptors.

In terms of procedural and internal evidence for validity, the least promising method from our study was cluster analysis. While cluster analysis undeniably increases the replicability of cutscores, its results cannot be linked directly to performance descriptors even in principle, which is crucial in any standard-setting study linked to curricular goals. Although Shin (2004) found that the results of this method corresponded well with cutscores set using other methods in his program, our study presents evidence that cluster analysis results do not always correspond well, and can differ dramatically, including failing to find as many clusters as there are program levels. In our study, this probably resulted from a combination of two factors: the lack of discriminatory power of the instrument, and the relative homogeneity of the incoming testing population. It is to be expected that nearly no new incoming students have ability levels bordering on being able to exit the program, for example, but since cluster analysis depends exclusively on the distribution of incoming student scores, estimating a cutscore for program exit is not possible.

It also bears mentioning that in a language program placement context such as ours, small numbers of significantly different clusters are not unexpected. Our program, like many, does not have proficiency requirements for application; accordingly, incoming students' proficiency backgrounds are highly diverse and would, over time, trend towards a continuous, random distribution that may not have distinct groupings. Increasing the discriminatory power of the test, while clearly needed, would not on its own lead to more or more distinct clusters if the distribution of student ability levels did not also subdivide further; meaningful clusters must exist in order to be found. Further, even when distinct clusters exist, it is entirely possible that curricular performance descriptors would not correspond to those boundaries. What exactly separates one cluster from another is not linked to performance characteristics. Finally, repeating a cluster analysis after each incoming group of students was added to the scores could result in changes in cutscores, meaning that their level in any one session would be essentially arbitrary. Therefore, cluster analysis (and possibly any statistical approach) seems more useful for deriving cutscores in community-based language programs or other language program contexts where the curriculum is revised constantly based on the needs of incoming students, and where in-class homogeneity is the primary concern. In programs with pre-existing curricula like ours, a valid statistical approach would have to incorporate some form of evidence of external correspondence to performance standards. Again, the use of concurrent performance assessments presents one possibility for future research here.

As for the Borderline group method, it is in one sense surprising that the resulting cutscores differed so dramatically from extant cutscores since teachers only evaluated students who had just been placed in their classes using those same cutscores. The Borderline cutscores were, however, unstable, perhaps resulting from the method's exclusion of the vast majority of students from the data. The Contrasting groups method would not necessarily fare better, however, as it similarly discards borderline and misplaced students. These limitations on entry data seem artificial in light of existing quantitative methods that can accommodate full ranges of student classifications. Most prominently, Signal Detection Theory and response operator characteristic (ROC) analysis has been extensively used in psychological diagnosis for this purpose (Swets, 2014). In standard-setting contexts, ROC analysis would benefit by taking into account *both* prototypical students who were placed in levels as well as misplaced students, and multiple cutscores could be evaluated at the same time by examining the entire ROC curve and determining the point along it at which a balance (which the program would have to determine) between false negatives and false positives was achieved. If *criterion (c)* were set to 0, then the likelihood of false positive and false negative errors would be equal, but programs may wish, for example, to set the cutscore at a value higher than 0 in order to prevent more false positive errors, which are typically harder to correct.

Including more students in the analysis would be helpful, but it is not a panacea, either. In our study, larger (although admittedly still small) numbers of borderline students did not correspond to increased stability across levels, suggesting that some imprecision was owing to the measurement instrument itself, which we will return to shortly. Separate from imprecision, a further limitation of this approach is that teachers need to have extensive experience with both their students and the curricular levels in order to differentiate between borderline, well-placed, and misplaced students reliably, but they must also make decisions shortly after placement in order to establish a credible link to their performance levels at the time of the placement exam. This is much like asking teachers to turn left while turning right. By providing systematic formative assessment tools targeting performance descriptors, assessment specialists could reduce the difficulty and subjectivity of standard-setting studies using this approach.

Finally, the results of our Bookmark method study are not ideal in some respects – e.g. cutscores for adjacent levels were sometimes separated by single items. It is likely that some of the difficulties experienced in our study result from the limitations of the placement test itself. In particular, if discrimination parameters even for only some items are low, as was the case in our study, those weakly discriminating items may end up located at uninterpretable positions in the OIB, affecting judgments of all subsequent items. The utility of the Bookmark method's cutscores might be bounded by the discriminatory power and content validity of the test. We recall Reckase's (2006a, 2006b) criticism that the Bookmark method may systematically underestimate cutscores, but in this study the derived values were not lower than those of the other methods, and did not trend downwards during subsequent rounds of discussion. Rather, our group of raters tended to adjust bookmarks higher with each successive round of discussion. More research on the actual decision-making processes of raters during the Bookmark procedure is clearly needed.

There are several limitations in this work that will need to be addressed in future studies. In the borderline method, teachers expressed difficulty in identifying borderline students because some students rarely attend classes and submit homework. It would be desirable to exclude such ratings for which teachers are not confident in future studies. It should also be noted that the amalgamation coefficient and ANOVA results are relatively weak criteria for determining the number of clusters in an HCA. In future studies, other solutions such as the adjusted RAND statistic (Santos & Embrechts, 2009) should be tried and implemented to identify and support the best clustering solution.

Conclusion

Our examination of the Bookmark method, Borderline group method, and cluster analysis for standard setting in a language program placement context revealed strengths and weaknesses associated with each method. Our study revealed several insights: cluster analysis is likely to be unsuitable for placement tests in programs with fixed curricula; the Borderline group method would be improved by expanding the range of evaluation categories included in the data and by distributing systematic formative assessment tools linked to performance descriptors at the onset of instruction; and the Bookmark method's utility appears strongly dependent on at least a moderate degree of discriminatory power that is consistent across test items. In addition, both student-centered and examinee-centered methods would benefit from the explication of separate incoming student performance descriptors, which may not be equivalent to having mastered curricular SLOs at lower levels. As for our program's placement context specifically, we can make several concrete recommendations. First, the exam as it is lacks the discriminatory power necessary to make decisions about seven levels of placement, and should be revised. This echoes the fact that minimum levels of test reliability are required for the desired number of proficiency levels set by the cutoff scores (Kaftandjieva, 2004). The Writing section of the exam, however, is more clearly linked to performance indicators; indeed, "scores" on the Writing exam are given as program levels, which is why this section was not included in the standard-setting study itself. Because calibration procedures for Writing have a longer, more established tradition and involve active engagement with teachers who have knowledge of the curriculum as it evolves, our recommendation is to increase the weight assigned to the Writing section's scores when combining it with the other sections, at least until such time as a new set of items can be developed and validated. In addition, it is essential that our program incorporate formative assessment more systematically to increase the ease with which teachers could identify misplaced students and take corrective action as necessary. In order to facilitate that process, program-level descriptors of prototypical, borderline, and misplaced students would help new teachers to participate in that process and provide the learning environment we desire to provide for our students.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The "Test of English as a Foreign Language Institutional Testing Program" (TOEFL ITP) is a paper-based version of the TOEFL on the same score scale as the TOEFL pBT. It is administered by the IEP to all students at the end of every session to enable external comparison of student levels. According to ETS, a score of 550 can be taken to indicate a CEFR level of B2 (www.ets.org/toefl_itp/research).
2. Given the large number of adjunct instructors in our program and the additional burden of having all teachers participate in the Borderline group method study, this was the largest number of teachers we could recruit while maintaining a balance of experience and program familiarity. Ideally, Tannenbaum and Cho (2014) recommend at least 10 experts per method. Our experience would indicate that reaching consensus on bookmarks in teams of larger than four would become very cumbersome, however, so we urge caution in using much larger groups for the Bookmark method.

References

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46–57). Alexandria, VA: TESOL.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*, 535–556.

Alsmadi, A. A. (2007). A comparative study of two standard-setting techniques. *Social Behavior and Personality*, *35*, 479–486.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bechger, T. M., Kuijper, H., & Maris, G. (2009). Standard setting in relation to the common European framework of reference for languages: The case of the state examination of Dutch as a second language. *Language Assessment Quarterly*, *6*, 126–150.

Beretvas, N. S. (2004). Comparison of bookmark difficulty locations under different item response models. *Applied Psychological Measurement*, *28*, 25–47.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*, 137–172.

Brown, J. D. (2013). Cut scores on language tests. In C. A. Chappelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley-Blackwell.

Camera, W. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practices*, *32*, 16–27.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice*, *15*, 20–31.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Council of Chief State School Officers [CCSSO]. (2000). *Key state education policies on K-12 education: Time and attendance, graduation requirements, content standards, teacher licensure, school leader licensure, student assessment. Results from the 2000 CCSSO policies and practices survey*. Washington, DC: Council of Chief State School Officers.

Green, A. (2012). Placement testing. In C. Coombe, B. O'Sullivan, P. Davidson, & S. Stoynoff (Eds.), *The Cambridge guide to language assessment* (pp.164–170). Cambridge: Cambridge University Press.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice*, *22*, 22–32.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). New York: Routledge.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publications.

Hein, S. F., & Skaggs, G. E. (2009). A qualitative investigation of panelists' experiences of standard setting using two variations of the bookmark method. *Applied Measurement in Education*, *22*, 207–228.

Hsieh, M. (2013a). An application of Multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing*, *30*, 491–512.

Hsieh, M. (2013b). Comparing Yes/No Angoff and Bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly*, *10*, 331–350.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice*, *25*, 19–20.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Kaftandjieva, F. (2004). *Standard setting. Section B of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe. Retrieved from www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionB.pdf

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

Kane, M. (1998a). Choosing between examinee-centred and test-centred standard setting methods. *Educational Assessment*, *5*, 129–145.

Kane, M. (1998b). Criterion bias in examinee-centred standard setting: Some thought experiments. *Educational Measurement: Issues and Practice*, *17*, 23–30.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) *Setting performance standards. Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*, 4–12.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), IRT-based standard setting procedures using behaviorial anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

Livingstone, S. A., & Zieky, M. J. (1989). A comparative study of standard setting methods. *Applied Measurement in Education*, *2*, 121–141.

McLaughlin, D. H. (1993). Validity of the 1992 NAEP achievement-level setting process. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Setting Performance Standards for Student Achievement: Background Studies* (pp. 81–122). Stanford, CA: National Academy of Education.

Näsström, G., & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research & Evaluation*, *13*(9), Retrieved from http://pareonline.net/getvn.asp?v=13&n=9

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing specific passing standard for the IELTS examination. *Language Assessment Quarterly*, *4*, 295–317.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*, 261–282.

Peterson, C. H., Schulz, E. M., & Engelhard Jr., G. (2011). Reliability and validity of bookmark-based methods for standard setting: Comparisons to Angoff-based methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *30*, 3–14.

Reckase, M. D. (2006a). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, *25*, 4–18.

Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice*, *25*, 14–17.

Santos, J. M., & Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification, *Proceedings of the 19th International Conference on Artificial Neural Networks: Part II*, 175–184.

Shin, S. K. (2004). How much is good enough? Setting and validating performance standards and cut scores for the web-based English as a second language placement exam at UCLA (Unpublished doctoral dissertation.) University of California Los Angeles.

Sireci, S. G. (2001). Standard setting using cluster analysis. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 339–354). Mahwah, NJ: Lawrence Erlbaum.

Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education*, *12*, 301–325.

SPSS Ins. (2011). PASW statistics for Windows (version 20.0) [computer software]. Chicago, IL: SPSS Inc.

Swets, J. A. (2014). *Signal Detection Theory and ROC analysis in psychology and diagnostics: Collected papers*. Philadelphia: Psychology Press.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*, 233–249.

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing*, *11*, 321–344.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, *40*, 231–253.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Standard setting: concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Erlbaum.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Zimoski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [computer software]. Lincolnwood, IL. Scientific Software International.