Psychological Test and Assessment Modeling, Volume 54, 2012 (3), 257-283

Examinee-centered standard setting for large-scale assessments: The prototype group method

Thomas Eckes[1]

Abstract

This paper presents the prototype group method (PGM) of standard setting within the context of a large-scale language assessment project. The PGM combines a Rasch measurement approach to the analysis of examinee proficiency with the concept of prototypes drawn from research on human judgment and categorization. Experts first identify learners typical of each of five levels of language proficiency as specified by the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). Based on the distributions of proficiency estimates for learner prototypes belonging to adjacent levels, cut scores are computed by means of a logistic regression procedure. These cut scores define the language proficiency level a particular examinee has achieved. Data from 39 independent samples of examinees (total N = 8,721) covering a range of German language proficiency levels are used to illustrate the PGM. Rasch analysis and logistic regression results corroborate the adequacy of this approach. The discussion focuses on the method's distinctive features, practical requirements of its implementation, and issues of cut-score validation.

Key words: standard setting, Rasch measurement, prototypes, large-scale assessment, language proficiency

[1] *Correspondence concerning this article should be addressed to:* Thomas Eckes, PhD, TestDaF Institute, Massenbergstr. 13 b, 44787 Bochum, Germany; email: thomas.eckes@testdaf.de

Standard setting refers to the process of establishing one or more cut scores on a test (Cizek, 2006; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kaftandjieva, 2004, 2010). Cut scores are used to divide a distribution of test scores into two or more categories of performance, representing distinct levels of knowledge, competence, or proficiency in a given domain. Thus, examinees may be categorized as *pass* or *fail*, or may be placed into a greater number of ordered performance categories, with labels such as *basic*, *proficient*, and *advanced*. The focus of the present research is on setting cut scores on language tests. Here, the categories are typically taken from the global scale of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The CEFR scale comprises six levels of communicative proficiency in three bands: A1, A2 (*basic user*), B1, B2 (*intermediate user*), and C1, C2 (*proficient user*).

Building on previous research on standard setting in the context of large-scale language assessment (Eckes, 2010a, 2010b), this study elaborates on the prototype group method for setting multiple cut scores. In essence, this method combines a Rasch measurement approach to test and assessment modeling with advances in research on human judgment and decision making. More specifically, the prototype group method uses expert judgments of the most typical or representative members of categories of examinees at various levels of proficiency, that is, judgments of *category prototypes*, as the basis for objectively determining cut scores on the latent proficiency dimension. To set the stage for discussing the unique features of the method and empirical evidence on its efficiency, a brief overview of commonly-used approaches to standard setting is provided next.

Standard-setting methods

Basic distinctions

Hambleton and Pitoniak (2006, p. 235) characterized standard setting as a "blend of judgment, psychometrics, and practicality," attributing to judgment a "critical role" in that judgments provided by standard-setting participants are the "cornerstone" on which cut scores are based. Not surprisingly, therefore, methods of standard setting differ in the kind of judgments provided and the way these judgments are processed to yield the desired cut scores.

From a systematic point of view, standard-setting methods can be distinguished with respect to (a) the focus of the judgmental task, (b) the process of providing judgments, and (c) the procedure of determining cut scores (e.g., Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kaftandjieva, 2004, 2010). In terms of task focus, standard-setting methods differ in a number of ways: judges may be asked to rate test items or examinees, the items being rated may be dichotomously or polytomously scored, and the judgment itself may consist of estimates of response probabilities, classes of items or examinees, or some other kind of data suitable for deriving cut scores. Similarly, the judgmental process may differ, such as the kind and frequency of feedback given to judges, the number of judgmental rounds per session, and whether the judges work on an individual basis, providing judgments independently, or work together, providing judgments after more or

less extensive group discussion. The third distinction refers to the decision rule or statistical method applied to judgmental data in order to determine cut scores, such as averaging across judgments, performing multiple regression analysis, or estimating item and examinee parameters based on item response theory (IRT). Making use of these and related distinctions, Kaftandjieva (2010) arrived at a list containing no less than 62 methods of standard setting.

One particularly simple classification system was suggested by Jaeger (1989). This system refers to the focus of the judgmental task: Do judges provide judgments of test items or judgments of examinees? When they judge items, the method is called *test centered*; when they judge examinees, the method is called *examinee centered*. The following section briefly discusses four frequently-used standard-setting procedures, two test-centered and two examinee-centered methods.

Test-centered methods

The *Angoff method* (Angoff, 1971), along with its many modifications and extensions that have been proposed over the years, is one of the most widely used and best researched methods of standard setting (e.g., Brandon, 2004; Hurtz & Auerbach, 2003; Plake & Cizek, 2012). Following this method, judges are instructed to imagine a hypothetical *borderline examinee*, for example, an examinee on the borderline between two adjacent performance levels, and to indicate, for each item in a test, the probability (between 0.00 and 1.00) that the examinee will answer the item correctly. Alternatively, judges may be asked to imagine a group of 100 borderline examinees, and to indicate how many of them will answer the item correctly. For each judge, the probabilities, or proportions, are averaged across items to determine each individual judge's cut score; then, these cut scores are averaged across judges to obtain the final cut score. To illustrate, if the average proportion across items and judges is 75%, the cut score on a 20-item test would be a number-correct score of 15.

The concept of borderline examinee, also known as *minimally competent examinee*, is of central importance to the Angoff method. Yet, exactly what is meant by "minimally competent" may not be sufficiently clear to judges. Hence, those in charge of the standard-setting procedure generally need to devote some time to promote an understanding of "borderline" knowledge, skills, and abilities before judges can start rating items. In order to accomplish this rating task, Zieky and Perie (2006) recommended that the judges' focus "should be at the point at which the best performing student within a category becomes indistinguishable from the worst performing student in the next higher category" (p. 8).

Quite obviously, the task of providing probability judgments with a hypothetical borderline examinee in mind is conceptually rather complex. Therefore, the quality of the final cut scores critically hinges on the extent to which judges are able to construct a precise mental picture of the borderline examinee, and on the extent to which judges agree with one another in their individual constructions. Moreover, even if judges have developed a clear, consensual belief of what constitutes borderline competency at the performance

levels in question, the task of determining the probability of a borderline examinee's correct response to a given test item remains cognitively highly demanding and may be too difficult for judges to accomplish in an accurate, unbiased manner (e.g., Brandon, 2004; Morgan & Michaelides, 2005; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Skorupski, 2012; see also Angoff, 1988).

The *bookmark method* (Mitzel, Lewis, Patz, & Green, 2001) has been developed to overcome some of the limitations of the Angoff method. In particular, the bookmark method addresses the need to set multiple cut scores on a single test, to simplify the rating task required of judges, and to accommodate both selected-response items, where examinees have to choose the correct answer from alternatives given (e.g., multiple-choice items), and constructed-response items, where examinees are asked to create a response (e.g., writing an essay). In addition, the method is relatively easy to implement, time-efficient, and has a sound methodological basis in terms of the psychometric approach used for cut-score computation (i.e., IRT-based methods). For these reasons, the bookmark method has become quite popular in recent years (Lewis, Mitzel, Mercado, & Schulz, 2012).

The basic bookmark standard-setting procedure can be illustrated as follows. Consider a test containing 40 multiple-choice items, where each item is scored dichotomously. Let the test be designed to assess three language performance levels (e.g., CEFR levels A2, B1, and B2). Then, judges are presented with a booklet consisting of the set of 40 items, one item per page, with items ordered from easiest to hardest (called the *ordered item booklet*, OIB). The ordering of items in the OIB is based on item difficulty estimates resulting from an item analysis using a suitable IRT model.

Judges are asked, for each level of performance, to place a bookmark on the first page in the booklet at which they believe the probability of a borderline examinee answering the item correctly drops below .67 (i.e., below a 2/3 chance). Thus, judges have to place two bookmarks in their booklet, each one identifying a cut-off between two adjacent levels. Judges usually repeat this marking procedure two times, each marking session constituting a separate round. Final cut scores are determined on the basis of the examinee proficiency estimate that corresponds with the difficulty of the item at which the bookmark was placed in the last round (for a detailed description of the bookmark method, see Cizek & Bunch, 2007; Karantonis & Sireci, 2006).

The most controversial issue about this method concerns the choice of an appropriate response probability (RP) value. In the above example, the RP value was set at .67. Yet, other values (e.g., RP = .50 or RP = .80) have also been suggested for various methodological or practical reasons (e.g., Beretvas, 2004; Huynh, 2006; Wang, 2003). Whatever the particular RP value may be that is used in a bookmark standard-setting procedure, the problem remains that the value's meaning needs to be conveyed adequately to judges. More generally speaking, it is a defining characteristic of test-centered approaches to standard setting that the judgment task is hypothetical in nature, requiring judges in some way or other to estimate an imagined examinee's probability of getting an item correct.

The two methods outlined next make use of a fundamentally different approach: they involve direct judgments of real examinees well known to judges (e.g., teachers judging each of their students in a class).

Examinee-centered methods

In the *contrasting groups method* (Berk, 1976), judges are asked to identify one group of examinees who are clearly above a particular performance level ("masters"), and another group of examinees who are clearly below that level ("non-masters"). When placing examinees into one of the two groups, judges are unaware of the actual test scores the examinees achieved on the test. Then, the test score distributions of these groups are analyzed to determine a cut score that distinguishes between examinees judged to be masters and those judged to be non-masters. One way is to compute, for each score distribution separately, the median (or mean) score, and to define the cut score as the midpoint between these two medians (or means). Another way involves the use of logistic regression (Livingston & Zieky, 1989). This statistical procedure yields the raw score at which the probability of membership in one of the two groups is .50. When more than two groups, or levels of performance, need to be distinguished, the approach to data analysis is basically the same as for two groups; that is, in each case adjacent levels are used to derive cut scores.

The contrasting groups method requires judges to provide clear, unambiguous decisions as to the membership status of each examinee. Yet, such decisions may be difficult to make for some examinees, in particular for those examinees perceived as falling in between the two groups considered. For example, when the construct being measured is multi-componential, as is typically the case with language proficiency, examinees may be judged to belong to the higher level on some components (e.g., grammar, vocabulary), but to some lower level on other components (e.g., pronunciation, fluency).

The *borderline group method* (Livingston & Zieky, 1982) addresses the need to provide in-between decisions. This method allows judges to classify examinees into more than two groups, one of which is a "borderline" group. Typically, judges are asked to sort examinees into three groups, one group of examinees they believe to be clearly at (or above) the performance level in question (masters), one group of examinees they believe to fall clearly below that level (non-masters), and one group of examinees they believe to fall in between (borderline examinees). Then, the test score distribution for the group of borderline examinees is formed, and the median of this distribution is computed to yield the desired cut score.

Though the borderline group method has some intuitive appeal, it has also been noted that one major limitation of this method concerns the size of the borderline group; that is, in many practical circumstances the borderline group may turn out to be too small to warrant sufficiently stable estimates of cut scores (e.g., Cizek, 2006; Hambleton & Pitoniak, 2006). In particular, when the test scores of the borderline group are spread widely over the range of possible scores, the median is ill-suited to yield a precise cut score estimate. Moreover, judges are likely to show less agreement when it comes to placing

examinees into the borderline group, as compared to placing examinees into the more clearly-defined group of masters or non-masters, respectively. This, of course, would further reduce the stability, or precision, of cut score estimation.

The prototype group method

Examinee categorization

The prototype group method (PGM) builds on an examinee-centered approach to standard setting. Hence, the method avoids some of the intricacies that are typically involved when judges are asked to estimate response probabilities. The method also helps to resolve some of the difficulties associated with the concept of borderline examinee.

To begin with, providing estimates of probability is a prime example of *judgment under uncertainty* (Tversky & Kahneman, 1974): Lacking any guidance by the formal rules of probability, judges have to generate *subjective probabilities* in their own minds reflecting their knowledge and expertise in the field. However, research on human judgment and decision making has provided ample evidence that under conditions of uncertainty humans often are subject to a whole range of heuristics and biases, which may lead to erroneous results (e.g., Bar-Hillel, 2001; Gilovich, Griffin, & Kahneman, 2002; Kahneman, 2011; Newell, Lagnado, & Shanks, 2007). Moreover, within the context of standard setting, judges have been shown to employ different standards when estimating the difficulty of items or placing examinees into performance categories (e.g., Longford, 1996; Van Nijlen & Janssen, 2008; see also Eckes, 2011a; Engelhard, 2009).

As discussed above, the judges' task is complicated by the fact that the Angoff and bookmark procedures imply judgments about a hypothetical target, that is, a borderline examinee's response to an item. The picture that each judge has in mind when asked to think of a borderline or minimally competent performance may vary widely, with unknown and possibly adverse consequences for the probability judgment provided in any given instance (Hein & Skaggs, 2010; Impara & Plake, 1997; Skaggs & Hein, 2011). For example, Hein and Skaggs (2010) found that judges in their study used alternative cognitive strategies such as thinking of individual students whom they had taught. The authors therefore recommended that judges "consider a single target student (or a small number of target students) whom they know rather than an entire classroom (or other large group) of hypothetical target students" (Hein & Skaggs, 2010, p. 42).

In contrast, the PGM makes no reference at all to the fuzzy concept of borderline performance. Rather, this method basically rests on re-considering the problem of setting cut scores on a test as a classification problem (Sireci, 2001; Sireci, Robin, & Patelis, 1999; see also Hess, Subhiyah, & Giordano, 2007): Examinees are to be sorted into groups or classes based on a set of relevant attributes, such that the classes are internally homogeneous and externally well-separated; the points of intersection between adjacent classes constitute the cut scores. However, different from Sireci (2001) who adopted a cluster-analytic approach to solving this kind of classification problem, the present method rests

on expert judgments of examinees representing each of a number of proficiency levels, and employs logistic regression on Rasch-based estimates of these examinees' proficiency to distinguish between adjacent levels.

Prototype approach

As implied by its name, the PGM adopts a prototype approach to standard setting. Specifically, it is assumed that judges with increasing professional experience explicitly or implicitly develop ever clearer views of those examinees being *most typical members* or *best examples* of the proficiency level or performance category in question. For example, language professionals who have conducted language courses for many years will presumably be able to identify individual learners who are typical examples of the basic level, and to distinguish these from other learners who are typical examples of the next higher, proficient level. The most typical or representative member of a given category is called *category prototype* (e.g., Hampton, 2006; Rosch, 1978; Eckes, 1989).

Expressed in terms of a spatial metaphor, prototypes are the centers of clusters of similar objects or persons (Hampton, 2006). According to the prototype view of category formation, prototypes build up through abstraction over previously encountered category members, retaining information on their most characteristic and distinctive attributes. Categorization of newly encountered exemplars is based on a comparison of the exemplars with a range of category prototypes. Exemplars are assigned to the category that yields the highest exemplar–prototype similarity.

Research on the structure of categories and processes of categorization has shown that prototypes, as compared to less typical category members, especially members located closer to the category boundary, offer a number of distinct advantages for information processing in general, and judgment and decision-making in particular: Prototypic category members are easier recognized, more rapidly and accurately accessed and retrieved from memory, contain richer and more strongly interrelated sets of attribute information, and are rated or evaluated with higher within-rater consistency and higher between-rater agreement (e.g., Barsalou, 1992; Eckes, 1991, 1996; Homa, 1984; Minda & Smith, 2001; Murphy, 2004).

The PGM-based process of standard setting builds on these findings. Specifically, experts are asked to identify examinees (students, learners) they know very well and view as best examples of a particular level of proficiency. These examinees are called *examinee prototypes* (*level prototypes*), or, in the context of language learning, *learner prototypes*. As noted before, the focus of the present research was on setting cut scores on a language test. Therefore, the experts were language teachers, trainers, or course leaders with extensive professional experience in the field of language testing and assessment. Accordingly, the definition of language proficiency levels was based on information taken from the global CEFR scale (Council of Europe, 2001).

The identification of examinee prototypes, that is, best examples of a given performance category, through expert judgment is one of two major components of the PGM. The

other component refers to the estimation of examinee proficiency based on Rasch measurement. That is, the total sample of examinees, which includes the examinee prototypes, is calibrated together with the set of test items on a single latent dimension. As a result, for each examinee prototype two data points are available: (a) category membership information (e.g., A2 or B1), and (b) an estimate of examinee proficiency (in logits). To determine cut scores between adjacent proficiency levels, a binary logistic regression procedure is used, where the category membership of examinee prototypes is predicted on the basis of the proficiency measures of these prototypes (for more detail on this procedure, see the Data Analysis section).

Method

Overview

Data collection followed the same general design as described in an earlier study on Rasch-based item banking (Eckes, 2011b). The data were collected as part of an ongoing process of developing an Internet-delivered placement test of German as a foreign language. Data input for the present study was provided by 39 independent samples of examinees. In each sample, examinees worked on a different version of a 10-item language test measuring general language proficiency. To jointly calibrate items across all sets, and to estimate examinee proficiency across all samples, a Rasch measurement approach was employed. In addition, for each sample of examinees, experts provided judgments of prototypes of each of five performance categories defined by the global CEFR scale (i.e., A1, A2, B1, B2, and C1). The proficiency estimates of the examinee prototypes were used to predict membership in adjacent levels of language proficiency through a logistic regression procedure.

Examinees

A total of 8,721 examinees took part in the process of data collection. There were 5,736 (65.8%) females and 2,909 (33.4%) males; 76 (0.9%) participants did not indicate their gender. The age of 82.5% of the total sample of participants ranged from 18 to 28 years (M = 23.24, SD = 5.98), 4.3% of the participants were younger than 18 years, 13.2% were older than 28 years.

At the time of testing, participants were either attending German language courses as part of a preparatory study program in Germany or planning to study at a German university while still in their home country. Texts were administered at test centers of the TestDaF Institute (TestDaF is short for "Test Deutsch als Fremdsprache"; www.testdaf.de) or at so-called *Lektorships* ("Lektorate") of the German Academic Exchange Service (DAAD; www.daad.de) in more than 50 countries from around the world.

Participants came from 133 different countries. In terms of the number of participants, the following 10 national groups ranked highest (percentage in parentheses): Russia

(11.0%), Indonesia (7.9%), Lithuania (5.6%), People's Republic of China (5.3%), Poland (4.8%), Ukraine (3.8%), Bulgaria (3.5%), Turkey (2.8%), Romania (2.6%), and France (2.2%).

Following data analysis, each participant received feedback on his or her performance. Feedback consisted of the test score earned in a given set of texts and the percentile rank achieved in the sample the participant belonged to.

Test material

In each sample, examinees were presented with a different version of a written language test conforming to the C-test format. C-tests are gap-filling tests widely used to assess general language proficiency for purposes of placement, screening, or provision of feedback to language learners (e.g., Eckes & Grotjahn, 2006a; Grotjahn, Klein-Braley, & Raatz, 2002; Klein-Braley, 1997). The test versions used in the present study comprised 10 short authentic texts in which parts of words were missing. Examinees had to insert the missing parts, that is, to restore the original words in each text.

In a series of trial studies spanning a six-year period (March 2005 – May 2011), a total of 314 mutilated texts, each containing 20 gaps, were subjected to detailed examination. Texts were compiled in sets of 10 texts, making a total of 39 sets. Within each set, texts were arranged in ascending order of supposed difficulty based on findings from pre-testing and expert judgment. The main purpose of this tentative difficulty-based ordering of texts was to keep low-proficient participants from becoming deterred when having to start with a text that was unduly hard. As usual, each text within a given set dealt with a different topic.

Gaps were inserted according to the classic deletion rule (the "rule of two"); that is, words were mutilated by deleting the second half of every second word, beginning with the second word of the second sentence. If a word had an odd number of letters, the larger part was deleted (see Grotjahn et al., 2002). Throughout the texts, the missing part of each word was indicated by a single underline of constant length. The instruction read: "Complete the gaps in the following texts in a meaningful way. You have five minutes for each text". The time allowed was also printed above each text. Test administrators strictly controlled adherence to this time limit.

Across all trial sets, two texts were the same. These common texts served to provide the link between the different sets, following a non-equivalent groups with anchor test design (also called common-item non-equivalent groups design; see, e.g., Kolen & Brennan, 2004; Wolfe, 2000; Wright & Stone, 1979). In each set, the common texts appeared at the third and eighth position, respectively.

Procedure and scoring

Trial sets were administered by TestDaF examination officers or DAAD-*Lektors*. All test administrators were highly experienced in teaching German as a foreign language at a

range of proficiency levels. They had been trained in the use of the global CEFR scale at TestDaF workshops and/or had become familiar with this scale in their professional work. In addition to the test booklets, each administrator received a questionnaire they were to respond to in order to gain the information required for implementing the PGM approach.

First, administrators had to indicate which CEFR proficiency levels were present in their respective language course. The scale reached from A1 (Breakthrough) to C2 (Mastery). Multiple answers were allowed. For easier orientation, Table 1 of the CEFR publication (German version; Europarat, 2001) was printed on a separate page of the booklet. In this table, each level was described briefly with respect to characteristic language skills using can-do statements. For example, at level A2 (Waystage), the scale states that learners "can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment)"; at level B2 (Vantage), learners "can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation".

Then, administrators had to identify those learners they believed they were able to reliably judge as to their communicative language skills, taking into account both reception (listening, reading) and production (writing, speaking). They were to strictly base their judgments on course observations of learners' language behavior across a period of several months. From this learner sample, administrators had to select those learners whom they considered best examples of the relevant CEFR proficiency levels A1 through C1. Level A1 was included because it was needed to determine the lower cut score for level A2; level C2 was not included, because it fell outside the range of levels for which cut scores were to be determined.

Specifically, administrators listed the names of maximally three learners who were (in their personal view) best examples of learners at level A1, best examples of learners at level A2, and so on. In case a particular level was not present in their language course, they were not to list any names at that level. Administrators were free to provide these listings shortly before, during, or shortly after the learners took the C-test. Learner listings were sent back to the TestDaF Institute along with the complete set of test booklets for scoring and data analysis.

The test booklet contained instructions on the first page and 10 gap-filling texts, with each text appearing on a separate page. Texts had to be worked on in the order and within the time limits given (i.e., five minutes per text); paging up and down the booklet was not allowed.

Each correctly restored word, or each acceptable variant (e.g., use of a plural form instead of the singular), was scored one point. Each incorrectly restored word, including spelling errors, was scored zero points (for a discussion of different scoring procedures for C-tests, see Eckes & Grotjahn, 2006b). Thus, the total score, computed across all 10 texts within a given set, could range from 0 to 200 points.

Data analysis

Since a considerable degree of dependence typically exists between the gaps within a text, each text was considered as a polytomous item or, alternatively, as a (discrete) rating scale. A text containing 20 gaps thus corresponded to a single item with item values ranging from 0 to 20. More generally, a text that contains m gaps could take on $m + 1$ item values, corresponding to $m + 1$ successive categories of the rating scale.

To calibrate the texts (or items) and estimate examinee proficiency measures, a polytomous Rasch modeling approach was used. Previous research had shown that Andrich's (1978) rating scale model (RSM) was highly suitable for that purpose (e.g., Eckes, 2006, 2007, 2011b). This model adds parameters to the basic Rasch model for dichotomous data (Rasch, 1960/1980) that describe the functioning of the rating scale; that is, the RSM adds a threshold parameter to represent the relative difficulty of the transition between adjacent response categories.

Specifically, in the RSM, the probability that person n with ability θ_n will obtain a score of k ($k = 0, \ldots, m$) on item i is expressed as

$$p_{ik}(\theta_n) = p(x_{in} = k \,|\, \theta_n) = \frac{\exp \sum_{j=0}^{k} \left[\theta_n - (\beta_i + \tau_j) \right]}{\sum_{r=0}^{m} \exp \sum_{j=0}^{r} \left[\theta_n - (\beta_i + \tau_j) \right]}, \qquad (1)$$

where $\tau_0 \equiv 0$ (Wright & Masters, 1982).

In Equation 1, β_i represents the difficulty parameter for item (or rating scale) i, and τ_j represents the threshold parameter for scale category j, that is, the parameter for the transition from category $j - 1$ to category j. Item i has m categories, and k is the count of the number of successfully completed categories for that item. That is, in the present context, k is the count of gaps within text i that person n filled in correctly.

The data analysis that was to yield estimates of the examinee prototype proficiencies finally needed for the PGM standard setting was run in two stages. In the first stage, the data within each sample of examinees were analyzed separately based on the RSM. As judged by statistical indicators of model fit, texts that did not function properly were excluded from further consideration. In the second stage, all texts that came through the first stage were put on the same difficulty scale using a concurrent estimation procedure. Generally speaking, a concurrent estimation procedure involves estimating item parameters using the data from two (or more) test forms, linked by a set of common items, simultaneously in a single run to achieve a common scale. In the present study, the concurrent Rasch analysis arranged examinees from all 39 samples along a single proficiency scale. The linking between these samples was provided by the two texts (anchor texts) that were common to all trial sets. All Rasch analyses were performed using the computer program WINSTEPS (Version 3.73; Linacre, 2011).

As mentioned earlier, cut scores between adjacent proficiency levels were determined using a binary logistic regression procedure. In the regression analysis, the dependent

variable was the category membership of the examinee prototypes (e.g., A2 vs. B1), and the independent variable was the Rasch-based prototype proficiency measure.

Specifically, the logarithmic form of the logistic regression equation for predicting the probability of an examinee prototype i belonging to the higher-level category y, that is, $p(y_i = 1)$, from a single predictor x is given by:

$$\ln\left(\frac{p(y_i = 1)}{1 - p(y_i = 1)}\right) = b_0 + b_1 x_i. \quad (2)$$

Here, x_i is the proficiency estimate of examinee prototype i in units of the logit scale, b_0 is the regression constant, and b_1 is the regression coefficient (e.g., Cohen, Cohen, West, & Aiken 2003).

Setting $p(y_i = 1) = .50$ in Equation 1 and solving for $x_i = x_c$, that is, for the logit value of the examinee prototype c located exactly between the two categories, gives the following result:

$$x_c = \frac{-b_0}{b_1}. \quad (3)$$

Equation 3 defines the cut score in terms of logits.

Linear transformation of the logit scale into the raw-score scale yields the final cut score that can be applied to the test score distribution to provide the desired classification of examinees.

Results

Separate Rasch analyses

To examine the psychometric quality of each of the 39 data sets, two kinds of Rasch statistics were considered: fit statistics and separation statistics. WINSTEPS provides users with an unweighted mean-square fit statistic (Wright & Masters, 1982). This statistic, also called *outfit* (Linacre, 2002), has an expected value of 1 and can range from 0 to infinity. Linacre (2002, 2011) suggested 0.50 as a lower-control limit and 1.50 as an upper-control limit for the outfit mean-square statistic. That is, Linacre considered mean-square values in the range between 0.50 and 1.50 as "productive for measurement" or as indicative of "useful fit" (see also Linacre, 2003). Other researchers suggested to use a narrower range defined by a lower-control limit of 0.70 (or 0.75) and an upper-control limit of 1.30 (see, e.g., Bond & Fox, 2007; R. M. Smith, 2004). The actual definition of lower- and upper-control limits for mean-square fit statistics will mainly depend on the nature of the assessment purpose (e.g., high-stakes vs. low-stakes decisions).

WINSTEPS also provides users with a weighted mean-square statistic (Wright & Masters, 1982). This statistic, also called *infit* (Linacre, 2002), has the same expected value and the same range of values as the outfit statistic. Whereas outfit is more sensitive to

unexpected responses on items located away from a person's proficiency level, infit is more sensitive to unexpected responses on items near a person's proficiency level. Fit values greater than 1 (*misfit*) are generally deemed to be more problematic than fit values smaller than 1 (*overfit*), because misfit can change the substantive meaning of the resulting parameter estimates (Wright & Linacre, 1994).

Table 1 presents the frequencies of mean square fit indices (infit, outfit) for three different fit intervals, computed at the level of individual texts within samples.

There were only two (out of the 390) cases in which the analysis yielded infit and outfit statistics exceeding the upper-control limit of 1.50. Concerning the narrower 0.70/1.30 interval, the vast majority of texts (i.e., about 96%) could still be considered fitting the RSM. Finally, applying the overly strict 0.90/1.10 interval resulted in about 13% of the texts being diagnosed as misfitting.

WINSTEPS produces various separation statistics as proposed by Wright and Masters (1982): The person separation index, from which the number of person strata index *H* can be computed, and the test reliability of person separation *R*. Index *H* is of special importance when a measurement instrument is to be used for placing examinees in a number of different levels of proficiency. In the present study, *H* indicates the number of statistically distinct levels of examinee proficiency in a given sample of examinees. In general, the number of proficiency levels that a measurement instrument can reliably distinguish should be at least as high as the number of proficiency levels that the instrument purports to distinguish. Since the language test that was the focus of this research is intended to sort examinees into one of four levels or ordered classes of language proficiency, the *H* index should take on values of 4.0 or higher.

In most samples, the number of person strata index *H* ranged from 6.50 to 8.50. There was not a single sample, in which this index fell below 5.60. The high measurement precision of each individual trial set was also indicated by the reliability index of person separation. This index ranged from .94 to .98.

Table 1:
Frequency of infit and outfit statistics of texts from 39 trial sets using different fit intervals

	Infit		Outfit	
Interval	**Freq.**	**%**	**Freq.**	**%**
Fit < 0.50	0	0.0	0	0.0
0.50 ≤ Fit ≤ 1.50	388	99.5	388	99.5
Fit > 1.50	2	0.5	2	0.5
Fit < 0.70	9	2.3	7	1.8
0.70 ≤ Fit ≤ 1.30	376	96.4	374	95.9
Fit > 1.30	5	1.3	9	2.3
Fit < 0.90	147	37.7	139	35.6
0.90 ≤ Fit ≤ 1.10	192	49.2	199	51.0
Fit > 1.10	51	13.1	52	13.3

Note. Each of the 39 trial sets contained 10 texts.

In order to select texts for inclusion in the concurrent Rasch analysis, two kinds of criteria were employed. The first criterion was based on the fit analysis conducted separately in each trial sample. Given that the language test would be associated with low to medium stakes, the 1.30 upper control limit was applied for infit and outfit statistics. This led to the elimination of 16 texts. The second criterion made use of the results from an analysis of differential item functioning (DIF) related to (a) examinee gender and (b) region of origin (European vs. non-European).

For purposes of the DIF analysis, the procedure implemented in WINSTEPS (Linacre, 2011) was adopted. This procedure involved first a joint run of all examinees to produce anchor values for examinee proficiency measures and for the rating scale structure (i.e., the threshold measures). Then, separate analyses were run with female and male (or European and non-European) examinee proficiency measures and the rating scale structure anchored at the values obtained in the previous analysis to produce item difficulty estimates separately for both groups of examinees. Finally, pairwise item difficulty difference *t*-tests were conducted between the two sets of item difficulty estimates (for more detail, see Linacre, 2011; see also Eckes, in press).

Since this method of DIF detection required 10 comparisons to be made per sample, critical significance levels had to be adjusted to guard against falsely rejecting the null hypothesis that no DIF was present. To this purpose, methods such as those based on the Bonferroni inequality (see Myers & Well, 2003) or the Benjamini – Hochberg procedure (see Thissen, Steinberg, & Kuang, 2002) can be used. Adopting the Benjamini – Hochberg approach, 13 gender-related DIF texts were identified. Seven of these texts were significantly more difficult for females than for males, the remaining six texts were significantly more difficult for males than for females. Four texts showed DIF related to region of origin, with two texts each favoring European and non-European groups of examinees, respectively. Expert review of item content did not suggest any unintended factor that could be hypothesized to account for the observed group differences in item difficulty. Any item showing DIF was excluded from further analysis.

Concurrent Rasch analysis

Figure 1 displays the result of the concurrent analysis in form of an examinee-item map, also called distribution map (Linacre, 2011) or Wright map (Wilson, 2005). The map illustrates that, through this analysis, all 8,721 examinees and all 281 items (texts) selected on the basis of the separate Rasch analyses were put on a common scale. This scale, the logit scale, is shown on the left-hand side. For ease of presentation, the logit scale was truncated at ± 4.0 logits.

Immediately to the right of the logit scale, the locations of the examinees are shown. These locations correspond to the estimates of the examinee proficiency measures. Each "#" in the examinee column stands for 25 examinees, and each dot stands for 1 to 24 examinees. The horizontal lines inserted in the examinee column indicate the location of the cut scores (in logits). Each logit value defines the boundary between two adjacent

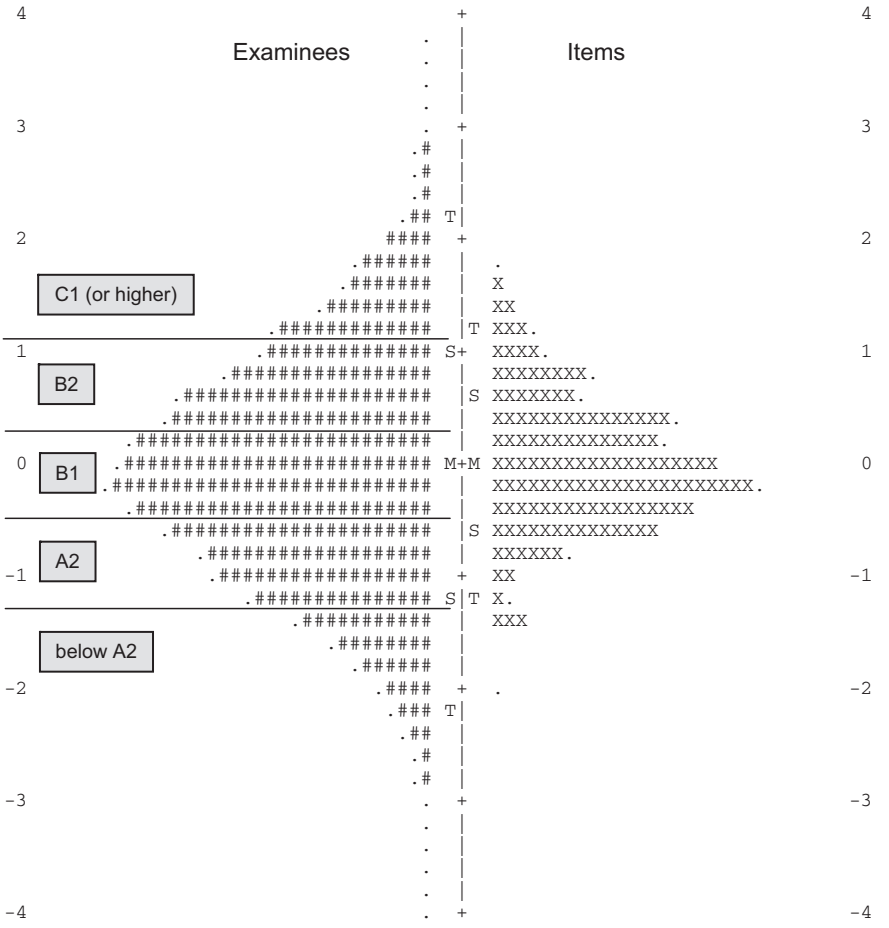

Figure 1:
Examinee-item map showing the cut scores (horizontal lines)

performance categories in terms of CEFR levels. Precisely how these boundaries were determined is explained later.

On the right-hand side, the locations of the items, corresponding to the item (or text) difficulty, are shown. Each "X" in the item column stands for 2 items, and each dot stands for 1 item. Along the line in the middle, markers summarize the distribution of examinee and item measures, respectively. An "M" marker represents the location of the mean measure, "S" markers are placed one sample standard deviation away from the mean, and "T" markers are placed two sample standard deviations away from the mean.

Three features of the map deserve particular attention: (a) the distribution of the examinee proficiency measures lines up nicely with that of the text difficulty measures, that is, the overall level of text difficulty is correctly targeted at the overall level of examinee proficiency, (b) the vast majority of texts are located in the middle range of the logit scale, that is, within one sample standard deviation around the mean, and (c) the distribution of the examinee measures approximates the normal distribution. Together these features attest to a high potential of the item pool to differentiate between examinees in terms of the construct being measured, that is, general language proficiency.

This conclusion was corroborated by the Rasch summary statistics *H* and *R*. Considering the total sample of examinees, the results were $H = 7.72$, and $R = .97$. Thus, about seven-and-a-half classes of examinees were reliably distinguished by the pool of texts studied here (i.e., almost twice as much as the final instrument is supposed to differentiate).

Determining cut scores

Analysis of prototype categorizations. As mentioned before, the present application of the PGM approach to standard setting calls for two distinct kinds of input data: (a) examinee proficiency measures, constructed in the concurrent Rasch analysis, and (b) the categorization of examinees as best examples (prototypes) of CEFR proficiency levels A1 to C1. However, before computing cut scores using the logistic regression procedure discussed above, it had to be ascertained that the examinee categorizations provided a sufficiently valid data base. Thus, the question was: To what extent were the expert judges, that is, language teachers, course leaders, and DAAD-Lektors, able to reliably identify examinee prototypes at each of the intended proficiency levels? The basic requirement was that the subjective teacher assignments of examinees to language performance categories be closely related to the objective examinee proficiency measures, independently estimated through the concurrent Rasch analysis. In other words, level A2 prototypes were predicted to show lower proficiency estimates than level B1 prototypes, and level B1 prototypes were predicted to show lower proficiency estimates than level B2 prototypes, and so on.

Table 2 presents descriptive statistics for the proficiency distributions of examinee prototypes (in logits) at each performance category considered by the experts. As can be seen, the majority of judgments were provided at levels B1 and B2, much less at level A2; not surprisingly, the levels which were represented by relatively small numbers of examinee prototypes were A1 and C1. Overall, 2,055 examinees (23.6%) were judged to be best examples of their respective performance category.

The mean logits showed a strictly monotonic increase from level A1 to level C1, with mean logit differences between levels falling within the narrow range of .65 to .73. The correlation between examinee prototype categorizations and examinee prototype measures was highly significant and substantial, $r(2055) = .67$, $p < .001$ (Spearman rank order correlation $r_s\,(2055) = .68$, $p < .001$). An analysis of variance on the logit values yielded a highly significant effect of performance category, $F(4, 2050) = 409.22$, $p < .001$, $\eta^2 = .48$. Simple comparisons revealed that all differences between logit means were highly significant (all p's $< .001$).

Table 2:
Descriptive statistics for examinee prototype logit distributions

Category	*n*	%	*M*	*SD*
A1	317	3.6	–1.53	0.96
A2	401	4.6	–0.80	0.75
B1	501	5.7	–0.15	0.72
B2	512	5.9	0.52	0.75
C1	324	3.7	1.17	0.84
Total	2,055	23.6	0.14	1.02

Note. n = Number of examinee prototypes per category. Percentage values in the third column refer to the total sample of examinees (N = 8,721). Logit means for categories A2 to C1 differ significantly from one another ($p < .001$).

Taken together, these findings suggested that the examinee prototype categorizations were carried out in an orderly fashion based on the performance categories A1 to C1. Put another way, experts' prototype judgments and the proficiency measures of examinee prototypes, which were independently constructed by means of a Rasch analysis of examinee responses to C-test items, were sufficiently congruent, attesting to the validity of the data used as input to the logistic regression procedure.

Logistic regression analysis. It should be noted first that results of a regression analysis can be markedly affected by a few extreme data points, so-called outliers. In the present case, though the prototype categorizations were in fine overall agreement with proficiency measures, individual expert judgments may have been subject to strong errors or biases leading to a highly proficient examinee being categorized as a typical B2 or even B1 learner instead of being categorized as a typical C1 learner. Conversely, again due to judgmental bias low-proficient examinees may have been categorized at a higher level than actually warranted. In fact, research has accumulated evidence indicating that a considerable amount of variance in ability or trait judgments can be accounted for by the influence of error (Hoyt & Kerns, 1999; see also Engelhard, 2002; Hoyt, 2000).

To examine the degree to which the regression results might have been influenced by outlying judgments, two kinds of analysis were performed. In the first analysis, extreme data points were deleted before running the analysis (e.g., Myers & Wells, 2003); that is, in each prototype logit distribution, the upper and lower 10% of examinee prototypes were excluded from the data set. All of the excluded prototypes were at least two standard deviations away from the mean of the respective distribution. In the second analysis, no data points were deleted, that is, the complete data set was analyzed. For the purposes of comparison, Table 3 presents results from both kinds of analysis.

All regression constants and coefficients proved to be highly significant. When outliers were deleted (upper part of Table 3), each of the four comparisons between adjacent proficiency levels showed model fit that was satisfactorily high. The rightmost column

Table 3:
Logistic regression results for determining cut scores

Comparison	*n*	b_0	$SE(b_0)$	b_1	$SE(b_1)$	Fit	x_c
			Outliers deleted				
A1 vs. A2	646	2.548**	0.193	2.082**	0.233	.401	-1.224
A2 vs. B1	812	1.572**	0.138	2.836**	0.206	.508	-0.554
B1 vs. B2	912	-0.479**	0.092	2.742**	0.188	.507	0.175
B2 vs. C1	753	-2.284**	0.177	2.207**	0.174	.436	1.035
			Outliers included				
A1 vs. A2	718	1.678*	0.113	1.211**	0.107	.181	-1.386
A2 vs. B1	902	0.806**	0.092	1.225**	0.110	.217	-0.658
B1 vs. B2	1,013	-0.208*	0.072	1.264**	0.104	.234	0.165
B2 vs. C1	836	-1.324**	0.119	1.034**	0.103	.187	1.280

Note. n = Number of examinee prototypes in the two categories considered. b_0 = Regression constant. b_1 = Regression coefficient. *SE* = Standard error (regression coefficient). Fit = Nagelkerke-R^2 index. x_c = Cut score (logits). * $p < .01$. ** $p < .001$.

lists the cut scores in units of the logit scale. As can be seen, the cut scores are almost evenly spaced across the logit scale.

In contrast, when the analysis was run with outliers included (lower part of Table 3), the cut scores were located farther away from one another; that is, differentiation between examinees near the important middle area of the proficiency distribution was lowered. Moreover, model fit proved to be much lower when outliers were included, and in two cases (i.e., A1 vs. A2, B2 vs. C1) was unsatisfactorily low. Therefore, the cut scores resulting from the analysis with outliers deleted were deemed more appropriate. It is the location of these cut scores that is graphically shown in Figure 1.

Summary and discussion

The prototype group method (PGM) of standard setting discussed in this paper uses a version of an examinee-centered approach where experts provide judgments of best or prototypic examples of performance categories. These category prototypes form the basis of determining cut scores along the latent proficiency continuum. Thus, the PGM differs from the contrasting groups method (Berks, 1976) in that it does not require experts to provide dichotomous judgments categorizing examinees as masters or non-masters. Similarly, the PGM differs from the borderline group method (Livingston & Zieky, 1982) in that it does not invoke the notoriously vague concept of borderline examinees. Use of the concept of borderline or minimally competent examinees is the hallmark of most test-centered approaches to standard setting. These approaches further complicate the experts' task by calling for judgments of the probability that a hypothetical examinee (i.e.,

borderline examinee) gets a particular item correct. Reconsidering standard setting as a natural categorization task, the PGM approach does without hypothetical judgments involving high levels of uncertainty and vagueness. Of course, as any other standard-setting method, the PGM rests on subjective judgments, yet these judgments are rooted in experts' detailed and extended observation of real examinees' performance in the relevant domain.

Basically, data from two kinds of sources are needed to implement the PGM: (a) categorical data, that is, categorizations of examinees as best examples of each of a number of performance levels, (b) test data, that is, responses of examinees to items on the test on which cut scores are to be set. In the present application, categorical data were provided by language teachers or course instructors who identified learners they considered as prototypes of performance categories defined by the CEFR proficiency levels A1 through C1. Judges were able to categorize nearly one fourth of all learners (23.6%) as prototypic A1, A2, B1, B2, or C1 learners, respectively. The learners thus categorized, and all the others, less typical learners took a test measuring general language proficiency. Cut scores were to be set on that test (i.e., a gap-filling test following the C-test format).

Examinee responses to test items were analyzed using a Rasch measurement approach. First, data from 39 independent samples of examinees were linked through common items to produce one large, connected data set comprising a total of 8,721 examinees and 281 test items. Then, Rasch analysis of this data set was performed to obtain item calibrations and estimates of examinee proficiency. The examinee proficiency estimates defined the predictor variable in a logistic regression procedure, where the criterion was membership of the prototypic examinees in one of two adjacent performance categories.

Analysis of the judgmental data revealed that the prototype categorizations were in sufficient agreement with the objectively determined proficiency measures of prototypic examinees. The correlation between level assignments and Rasch estimates of examinee proficiency was .67, and the mean logits of the examinee prototype distributions increased in a strictly monotonic fashion from the lowest level (A1) to the highest level (C1), supporting the conclusion that experts were indeed able to reliably and accurately categorize the majority of learner prototypes according to the global CEFR scale. This finding was of critical importance, because the PGM requires that expert judgments of learner prototypes reflect the dimension that the test is intended to measure. Turning to the latent proficiency dimension, it was of similarly critical importance that the measurement precision was high enough to reliably distinguish at least four levels of proficiency. As indicated by the number of examinee strata index, this condition was easily met: In the total sample, the number of statistically distinct levels of examinee proficiency was almost twice as high as minimally required (i.e., $H = 7.72$).

The fit of the logistic regression model was satisfactorily high, particularly after deletion of outlying data points. For each pair of levels, estimates of regression coefficients were statistically highly significant, indicating that the cut scores (in logits) computed from the regression equation served to reliably distinguish between examinee prototypes from adjacent performance categories. Because the regression analysis run after deletion of

outliers was associated with higher model fit and yielded higher differentiation between examinees located in the middle range of the proficiency distribution (i.e., one standard deviation away from the center of the distribution), the cut scores derived from this analysis were considered better suited to the purpose of placing language learners into internally homogeneous and externally well-separated categories of performance. Hence, these cut scores formed the final result of the PGM-based process of standard setting.

As many researchers have noted, each standard-setting method has unique strengths and weaknesses, and no method will ever be suited to fit each and every assessment purpose (e.g., Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kaftandjieva, 2010). With respect to the PGM, the following issues may be of special concern.

First of all, the judges must be qualified to determine each examinee's level of proficiency with sufficient accuracy; that is, judges need to be experts in the relevant domain (e.g., language teaching), they need to know what levels of proficiency they are to distinguish, and they need to consider exclusively the proficiency in question, meaning that they explicitly disregard possible construct-irrelevant factors such as gender, ethnicity, or personality. The PGM only requires that judges focus on examinees or learners they feel they can surely identify as prototypic examples of a given performance category; hence, this issue may not be as critical as it is with other examinee-centered methods where judgments of borderline examinees are called for.

Second, unlike test-centered methods, the PGM requires not only judgmental data but also test or item-response data provided by the very sample of examinees on which the categorical judgments are based. These two kinds of interrelated data sets need to be analyzed separately to ascertain each data set's psychometric quality before running the logistic regression analysis. In some situations, this dual data set demand may seem to be a serious drawback. Yet, in the present context of collecting item responses using many different test forms administered to small groups of examinees (mostly 5 to 20 examinees per language course), these data were readily available. Moreover, considering the large number of test items successively trialed over an extended period of time in test centers spread across all parts of the world, a test-centered approach would simply not have been feasible.

Third, logistic regression analysis typically requires a large sample of examinees to yield sufficiently stable estimates of regression coefficients. In this study, the regression analysis rested on a sub-sample of 2,055 examinees categorized as prototypic category members. Even when outliers were deleted, the size of the sub-samples in each comparison between adjacent performance levels ranged from 646 (A1 vs. A2) to 912 (B1 vs. B2). These sample sizes can certainly be considered sufficient for the purposes of logistic regression analysis. At the same time, the sample size requirement imposed by the PGM (and by similar approaches like the contrasting groups method) is definitely an issue of concern.

Whichever method of standard setting has been chosen, it is understood that the final cut scores need to prove their utility in the operational setting for which they were developed. Within the present context, cut scores have been used with a completely web-based placement test of German as a foreign language, which (in German) is called the

"Online-Einstufungstest Deutsch als Fremdsprache" (onDaF, for short; see Eckes, 2010a; www.ondaf.de). The onDaF has served purposes of assigning L2 learners of German to language courses at institutions of higher education, providing feedback to L2 learners who plan to take the TestDaF, and assisting DAAD-Lektors in deciding on foreign students' eligibility for scholarships. Since its launch in October 2006, the onDaF has been taken by a steadily increasing number of examinees at hundreds of licensed test centers around the world (about 50,000 examinees, as of end of December 2011). Using the cut scores resulting from the current analysis, 12.1% of the examinees were placed below CEFR level A2, 23.8% were placed at level A2, 38.6% at level B1, 21.3% at level B2, and 4.2% at level C1 (or higher). Along with the consistently positive feedback received from DAAD-Lektors and language course instructors concerning the homogeneity of the learner groups formed on the basis of onDaF results, these numbers indicate that the cut scores have led to reasonable, well-differentiated placement decisions.

Furthermore, following the methodology advanced in the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2009; see also Figueras, North, Takala, Verhelst, & van Avermaet, 2005), PGM cut scores were subjected to an external validation procedure. Specifically, the onDaF placements of examinees were compared to placements of the same examinees generated by two language tests that had already undergone a standard-setting process relating the tests to the CEFR. The first of these tests was the German section of the online diagnostic language testing system DIALANG (Alderson, 2005; Alderson & Huhta, 2005), the second was the TestDaF (Kecker, 2011; Kecker & Eckes, 2010). Though onDaF and DIALANG differed in the construct being measured and in the overall test design and item format, both tests' placements of examinees along the CEFR scale were in fine agreement (Eckes, 2010a). Much the same finding was reported in another study comparing onDaF placements and the assignment of examinees to the categories of the TestDaF proficiency scale (Eckes, 2013). Hence, both external validation studies provided additional evidence attesting to the appropriateness of PGM cut scores for the onDaF.

Conclusion

It is commonly accepted that there is no one best method of setting cut scores. As stated in the *Standards for Educational and Psychological Testing*, "There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 53). It is similarly true that different standard-setting methods most likely result in different cut scores (e.g., Downing, 2006; Kaftandjieva, 2010; Zieky, 2012). Therefore, choosing a method of standard setting is itself a matter of judgment – a judgment that should take into account the specifics of the design, format, and purpose of the test under consideration as well as the technical and procedural requirements imposed by a particular method.

The prototype group method (PGM) suggested here has been developed to meet the demands of an online placement test that measures general language proficiency, where test data are collected over extended periods of time, covering test administrations in test centers spread across all parts of the world. Adopting a categorization approach to standard setting, and directing the focus of the judgmental task on identifying prototypic members of performance categories, rather than requiring judges to think of borderline examinees responding to test items, has proved to yield empirically valid and practically useful results. It remains to be seen how well this approach can be transferred onto other tests and assessment situations.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22,* 301-320.

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. *Applied Measurement in Education, 1,* 215-222.

Bar-Hillel, M. (2001). Subjective probability judgments. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioural sciences* (pp. 15247-15251). Oxford, UK: Elsevier.

Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists.* Hillsdale, NJ: Erlbaum.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45,* 4-9.

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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 14,* 59-88.

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225-258). Mahwah, NJ: Erlbaum.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR): A manual.* Strasbourg: Language Policy Division.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Erlbaum.

Eckes, T. (1989). Knowledge structures and knowledge representation: Psychological models of conceptual order. In O. Opitz (Ed.), *Conceptual and numerical analysis of data* (pp. 269-277). Berlin: Springer.

Eckes, T. (1991). *Psychologie der Begriffe: Strukturen des Wissens und Prozesse der Kategorisierung* [The psychology of concepts: Structures of knowledge and processes of categorization]. Göttingen, Germany: Hogrefe.

Eckes, T. (1996). Begriffsbildung [Concept formation]. In J. Hoffmann & W. Kintsch (Eds.), *Lernen* (pp. 273-319). Göttingen, Germany: Hogrefe.

Eckes, T. (2006). Rasch-Modelle zur C-Test-Skalierung [Rasch models for scaling of C-tests]. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-test: Theory, empirical research, applications* (pp. 1-44). Frankfurt, Germany: Lang.

Eckes, T. (2007). Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen [Construction and analysis of C-tests with rating scale Rasch models]. *Diagnostica, 53,* 68-82.

Eckes, T. (2010a). Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung [The online placement test of German as a foreign language (onDaF): Theoretical foundations, construction, and validation]. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-test: Contributions from current research* (pp. 125-192). Frankfurt, Germany: Lang.

Eckes, T. (2010b). Standard-Setting bei C-Tests: Bestimmung von Kompetenzniveaus mit der Prototypgruppenmethode [Setting performance standards on C-tests: Definition of proficiency levels using the prototype group method]. *Diagnostica, 56*, 19-32.

Eckes, T. (2011a). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt, Germany: Lang.

Eckes, T. (2011b). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling, 53,* 414-439.

Eckes, T. (2013). Die onDaF – TestDaF-Vergleichsstudie: Wie gut sagen Ergebnisse im onDaF Erfolg oder Misserfolg beim TestDaF vorher? [The onDaF – TestDaF comparability study: How well do onDaF results predict success or failure at the TestDaF?]. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-test: Current trends* (pp. 1-26). Frankfurt, Germany: Lang.

Eckes, T. (in press). A study of differential item functioning in the TestDaF reading and listening sections. In E. D. Galaczi & C. J. Weir (Eds.), *Voices in language assessment: Exploring the impact of language frameworks on learning, teaching and assessment.* Cambridge, UK: Cambridge University Press.

Eckes, T., & Grotjahn, R. (2006a). A closer look at the construct validity of C-tests. *Language Testing, 23,* 290-325.

Eckes, T., & Grotjahn, R. (2006b). C-Tests als Anker für TestDaF: Rasch-Analysen mit dem kontinuierlichen Ratingskalen-Modell [C-tests as anchor for TestDaF: Rasch analyses using the continuous rating scale model]. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-test: Theory, empirical research, applications* (pp. 167-193). Frankfurt, Germany: Lang.

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen* [Common European framework of reference for languages: Learning, teaching, assessment]. Berlin: Langenscheidt.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2009). Evaluating the judgments of standard-setting panelists using Rasch measurement theory. In E. V. Smith & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 312-346). Maple Grove, MN: JAM Press.

Figueras, N., North, B., Takala, S., Verhelst, N., & van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22,* 261-279.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.) (2002). *Heuristics and biases: The psychology of intuitive judgment.* Cambridge: Cambridge University Press.

Grotjahn, R., Klein-Braley, C., & Raatz, U. (2002). C-tests: An overview. In J. A. Coleman, R. Grotjahn & U. Raatz (Eds.), *University language testing and the C-test* (pp. 93-114). Bochum: AKS-Verlag.

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.

Hampton, J. A. (2006). Concepts as prototypes. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 46, pp. 79-113). New York: Academic Press.

Hein, S. F., & Skaggs, G. (2010). Conceptualizing the classroom of target students: A qualitative investigation of panelists' experiences during standard setting. *Educational Measurement: Issues and Practice, 29*(2), 36-44.

Hess, B., Subhiyah, R. G., & Giordano, C. (2007). Convergence between cluster analysis and the Angoff method for setting minimum passing scores on credentialing examinations. *Evaluation and the Health Professions, 30,* 362-375.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, pp. 49-94). New York: Academic Press.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5,* 64-86.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4,* 403-424.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63,* 584-601.

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*(2), 19-20.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34,* 353-366.

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. In Council of Europe (Ed.), *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section B, pp. 1-43). Strasbourg: Language Policy Division.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL.* Arnhem, The Netherlands: EALTA.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12.

Kecker, G. (2011). *Validierung von Sprachprüfungen: Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen* [Validation of language examinations: Relating the TestDaF to the Common European Framework of Reference for Languages]. Frankfurt, Germany: Lang.

Kecker, G., & Eckes, T. (2010). Putting the Manual to the test: The TestDaF – CEFR linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual* (pp. 50-79). Cambridge, UK: Cambridge University Press.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14,* 47-84.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225-253). New York: Routledge.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16,* 878.

Linacre, J. M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions, 17,* 918.

Linacre, J. M. (2011). *A user's guide to WINSTEPS: Rasch-model computer programs* [Computer software manual]. Chicago: Winsteps.com.

Livingston, S. A., & Zieky, M. J. (1982*). Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

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

Longford, N. T. (1996). Reconciling experts' differences in setting cut scores for pass-fail decisions. *Journal of Educational and Behavioral Statistics, 21,* 203-213.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 775-799.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Morgan, D. L., & Michaelides, M. P. (2005). *Setting cut scores for college placement* (College Board Research Report No. 2005-9). New York: College Board.

Murphy, G. L. (2004). *The big book of concepts.* Cambridge, MA: MIT Press.

Myers, J. L., & Well, A. D. (2003*). Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Erlbaum.

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2007). *Straight choices: The psychology of decision making.* New York: Psychology Press.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181-199). New York: Routledge.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press. (Original work published 1960)

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.

Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement.* Stanford, CA: National Academy of Education.

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: Erlbaum.

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

Skaggs, G., & Hein, S. F. (2011). Reducing the cognitive complexity associated with standard setting: A comparison of the single-passage bookmark and yes/no methods. *Educational and Psychological Measurement, 71,* 571-592.

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 135-147). New York: Routledge.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73-92). Maple Grove, MN: JAM Press.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini – Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27,* 77-83.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Van Nijlen, D., & Janssen, R. (2008). Modeling judgments in the Angoff and contrasting-groups method of standard setting. *Journal of Educational Measurement, 45,* 45-63.

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

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement, 1,* 409-434.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370.

Wright, B. D., & Masters, G. N. (1982*). Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Zieky, M. (2012). So much has changed: An historical overview of setting cut scores. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 15-32). New York: Routledge.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service.