Open access • Journal Article • DOI:10.1111/J.1745-3984.2007.00051.X

Modeling Judgments in the Angoff and Contrasting-Groups Method of Standard Setting — Source link

Daniël Van Nijlen, Rianne Janssen

Institutions: Katholieke Universiteit Leuven

Published on: 01 Mar 2008 - Journal of Educational Measurement (Blackwell Publishing Inc)

Journal of Educational Measurement
Spring 2008, Vol. 45, No. 1, pp. 45–63

Modeling Judgments in the Angoff and Contrasting-Groups Method of Standard Setting

Daniël Van Nijlen and Rianne Janssen
Katholieke Universiteit Leuven

Essential for the validity of the judgments in a standard-setting study is that they follow the implicit task assumptions. In the Angoff method, judgments are assumed to be inversely related to the difficulty of the items; contrasting-groups judgments are assumed to be positively related to the ability of the students. In the present study, judgments from both procedures were modeled with a random-effects probit regression model. The Angoff judgments showed a weaker link with the position of the items on the latent scale than the contrasting-groups judgments with the position of the students. Hence, in the specific context of the study, the contrasting-groups judgments were more aligned with the underlying assumptions of the method than the Angoff judgments.

In standard setting, two main components can be distinguished (Kane, 1998a): the development of the performance standard and the identification of the cut score. The performance standard describes what it means to meet a standard in a certain domain (Haertel & Lorié, 2004). Although this description ultimately refers to some observable performance or behavior, it is typically formulated at an abstract level and must be made concrete when using tests to assess respondents on a performance standard. More specifically, a score should be set that can be considered as a cut-off to distinguish among the examinees who have the characteristic described in the performance standard and those who do not. This cut score is a minimum test score that is an operationalization of the performance standard (Kane, 2001) and can be used to make a classification of the test takers. The identification of the cut score can be made using an item- and an examinee-centered method of standard setting (see e.g., Berk, 1996; Cizek, 1996; Haertel & Lorié, 2004; Jaeger, 1989; Kane, 1994, 2001).

The field of standard setting is not without controversy. Zieky (2001) summarized the field by stating that there is no such thing as a true cut score and, hence, that it is useless to evaluate the correctness of a cut score. Likewise, Kane (1994) stated: "There is no gold standard. There is not even a silver standard" (pp. 448–449). Moreover, there are a large number of different methods available, but they tend to lead to different results (Jaeger, 1989; Zieky, 2001), and there is no agreement on which method should be preferred. In response to this conclusion, Green, Trimble, and Lewis (2003) described a procedure in which the results of three standard-setting methods were integrated into one judgment.

Despite the controversy, it is clear that in high-stakes testing and national assessments, standard setting is an indispensable step to decide which test takers can be considered as masters. Kane (1998a) proposed a possible solution to this paradox by evaluating standard-setting procedures according to criteria used for evaluating

policy decisions. These policy-based criteria refer to the appropriateness and the defensibility of the standard. Along these lines, Kane (1998a) discussed possible criteria to evaluate item- and examinee-centered standard-setting methods, but he concluded that there is not really an empirical criterion—neither procedural, nor internal or external—to prefer one approach over the other. This absence of empirical criteria led Kane (1998a) to put forward two analytic criteria to choose for an item- or an examinee-centered approach: consistency with the model of achievement and feasibility, given the assessment format.

In a special issue of *Applied Measurement in Education* on "Qualitative inquiries of participants' experiences with standard setting," McGinty (2005) made a plea for more emphasis on validity in the evaluation of standard-setting methods, rather than limiting research to reliability. Reliability criteria like replicability and interjudge consistency can be "artificially" achieved through procedural features like iterative processes and the use of normative data by panelists and, thus, only provide limited support for the appropriateness of a standard. Validity can be studied by looking at different standard-setting procedures from an empirical point of view by investigating how well the judgments made in an examinee- and item-centered standard-setting procedure follow the implicit task assumptions. Van der Linden (1982) referred to this as "intrajudge consistency."

The purpose of the present paper is to compare, in an empirical way, an item- and an examinee-centered method. More specifically, the contrasting-groups method of standard setting and the dichotomous version of the Angoff method were studied. These methods were chosen because of their widespread use and because in both methods the judges have to make a dichotomous classification, either on the person side or on the item side.

In the following, both standard-setting methods are subsequently described. On the basis of their link with the continuum view of mastery (Meskauskas, 1976), a logistic regression model is proposed to analyze the dichotomous classifications made by the judges in each method. Using a model from item response theory (IRT), both the classifications made in the Angoff and contrasting-groups method can be regressed on the same scale, namely on the difficulties of the items and the abilities of the students, respectively. In this way, the results of the logistic regression for both standard-setting procedures are comparable. The logistic regression model will be estimated in a Bayesian way.

Examinee- and Item-Centered Standard-Setting Methods

Examinee-Centered Methods

Livingston and Zieky (1982) proposed two examinee-centered methods of standard setting. In the borderline-group method, judges have to identify students who are performing at the borderline with regard to the performance standard. The median test score of this borderline group is then used as the cut score. In the contrasting-groups method, judges classify the students in two contrasting groups: students who reached the standard and students who did not reach the standard. The cut score is then set in such a way as to get the best discrimination between the two groups and to make sure that the fewest students are wrongly classified by applying the cutoff.

Item-Centered Methods

According to Hurtz and Auerbach (2003), the method introduced by Angoff (1971) is the most used and best known example of an item-centered standard-setting method. In the Angoff method, judges have to assess the expected performance on each item of the minimally competent student (MCS). The MCS is a person who is just performing at an acceptable level to pass the performance standard. In the first formulation of the method, the expected performance is expressed dichotomously, namely as whether or not this person should be able to answer the item correctly. In a second version, which was introduced by Angoff (1971) in a footnote, judges have to express the expected performance as the probability that the MCS would solve the item. In both versions, the cut score on the test score scale is set at the sum of the expected performances on all items. Impara and Plake (1997) found that both versions produced comparable cut scores, but that judges found the dichotomous version more comfortable to use. Different modifications to the Angoff method were proposed over the years. Hurtz and Auerbach (2003) and Brandon (2004) gave an overview and an evaluation of the most frequent modifications.

Standard Setting and the Continuum View of Mastery

The item-centered as well as the examinee-centered methods can be situated within a continuum view of mastery (Meskauskas, 1976). This view implies that within a domain, students can be ordered along a continuously distributed ability dimension. In line with this view, it can be expected that at the same time the items can be ordered along this continuum, according to their difficulty. Items on the lower end of the continuum are items that students with a low level of ability are expected to solve correctly. Students at the higher end of the continuum are also expected to solve these items, and, in addition, they are expected to solve items that are at the higher end of the continuum. The continuum view of mastery with a complementary ordering of persons and items in terms of ability and difficulty forms the core of different models of IRT. In addition, the judgments in an examinee- or item-centered method of standard setting and their resulting cut scores can be framed within a model representing the continuum view, as shown below.

Examinee-Centered Methods

In an examinee-centered method, the cutoff is a function of the position of a selected group of students along the continuum. Students' mastery in a domain is judged. Independently from the judgments, the students' performances on the test are assumed to be indicators of their position on the same continuum. In this way, the classification of the examinees can be used as a basis to elicit the implicit standards of the judges and a cut score can be set on the test continuum.

Livingston and Zieky (1989) proposed the logistic regression model to link the contrasting-groups judgments to the test score continuum. In this model, the dichotomous judgments can be regressed on the ability estimates of the students, and a cut score can be derived from this regression. More specifically, the probability that student p is classified as a master is modeled as a function of the ability of the student

(θ_p) and the severity of the judges (α). The strength of the relationship between the classification and the ability estimate is expressed by the slope parameter δ. The model reads as follows:

$$P(y_p = 1 \mid \theta_p) = \frac{\exp(\alpha + \delta\,\theta_p)}{1 + \exp(\alpha + \delta\,\theta_p)}. \tag{1}$$

Using the model in Equation (1), the cut score is set at the point where the probability of being classified as a master is .50, which can be calculated as $-\alpha/\delta$.

The model in Equation (1) is identical for all judges. It is possible to extend the model in such a way that differences among the judges can be modeled (Longford, 1996). Differences in severity can be modeled by allowing the intercept of the regression to vary over judges. In the following equation, a random effect a_g is added for each judge. The probability that student p is classified as a master by judge g then becomes:

$$P(y_{pg} = 1 \mid \theta_p, a_g) = \frac{\exp(\alpha + a_g + \delta\,\theta_p)}{1 + \exp(\alpha + a_g + \delta\,\theta_p)}, \tag{2}$$

where the random coefficients a_g are assumed to follow a normal distribution with a mean of 0 and a variance σ_a^2. Equation 2 implies that each judge may have a different standard in mind when classifying students. However, all judges rely to the same extent on the ability continuum to make their classification, as indicated by the fixed slope parameter δ.

In a fully random logistic regression, differences among judges in the slope parameter are also allowed. The probability of student p being classified as a master by judge g then becomes:

$$P(y_{pg} = 1 \mid \theta_p, a_g, d_g) = \frac{\exp(\alpha + a_g + (\delta + d_g)\,\theta_p)}{1 + \exp(\alpha + a_g + (\delta + d_g)\,\theta_p)}, \tag{3}$$

where the random parameters d_g are assumed to be normally distributed with a mean 0 and a variance σ_d^2. As before, it is assumed that $a_g \sim \mathrm{N}(0, \sigma_a^2)$. According to the fully random logistic regression, the cut score for judge g equals: $-(\alpha + a_g)/(\delta + d_g)$.

Note that the distribution for the discrimination parameters allows for negative slope parameters, which would indicate that a student with a higher θ_p has a lower probability of being classified as a master than a student with a lower test score. Given that it is possible that some teachers take into account a characteristic of the pupil that is inversely related to the ability, the model should be able to include negative slope parameters (Longford, 1996).

Figure 1 gives a graphical representation of Equation 3 for two judges with different values for both a_g and d_g. Both regression curves follow the well-known S-shaped curve. As both slope parameters are positive, the probability of being classified as a master increases with increasing test scores. However, for Judge 1, the relationship between the test score and the mastery classification of students is stronger

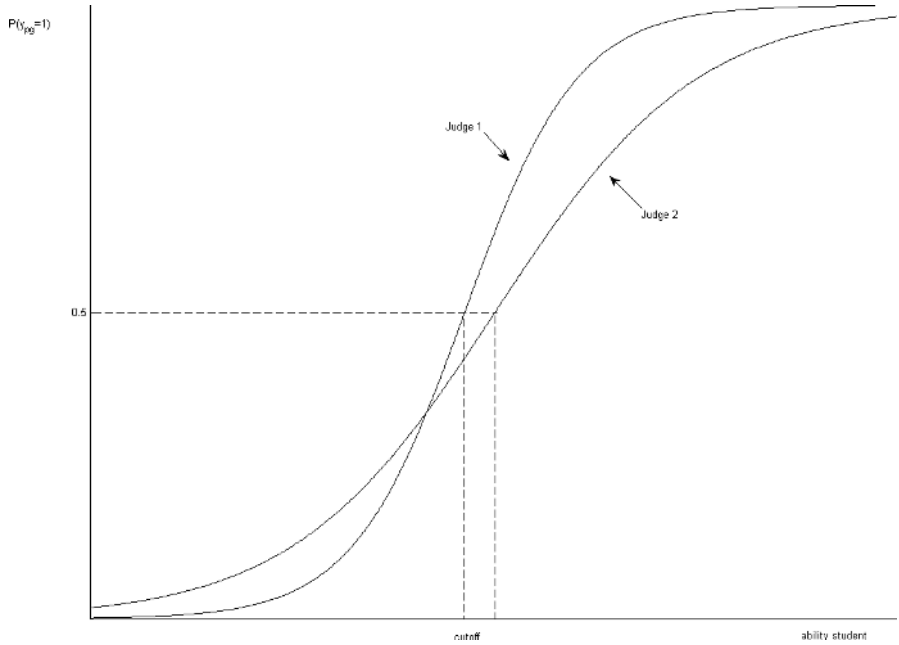

FIGURE 1. *Graphical illustration of a fully random logistic regression model.*

than that for Judge 2. Note that in the fully random logistic regression model, the intercept cannot be strictly interpreted as a measure for the judge's severity. Differences in a_g only denote the difference in severity for the two judges for a student with $\theta_p = 0$. Because of differences in the slope parameter among the judges, and the corresponding crossing of the regression curves, one judge can be more lenient (as indicated by a higher mastery classification probability) than the other judge on one part of the continuum, while being less lenient (as indicated by a lower mastery classification probability) on another part of the continuum.

Item-Centered Methods

The basic assumption of item-centered methods of standard setting is that items differ in their position along the latent continuum and are situated both above and below the position of the MCS. In order to infer the position of the MCS, judges have to predict the performance of the MCS on the items.

In the Angoff method of standard setting, the judge has to start from the conceptualization of the MCS and compare it with the items' expected difficulty. Items that are positioned above the MCS on the latent continuum should get a low probability of success, while items that are positioned below the MCS should get a high probability of success. Van der Linden (1982) showed that the IRT solution of the test can be used as a means of investigating the intrajudge consistency of Angoff judgments. Using the test characteristic curve, the Angoff cutoff score on the test score scale can be linked to the IRT scale, and the position of the MCS can be inferred. Given the position of the MCS on the IRT scale, the consistency of the judgments with the

characteristics of the items can be investigated. A similar use of the IRT continuum can be found in the "Process feedback" procedure in standard setting proposed by Reckase (2001).

For dichotomous Angoff data, a realistic model to link the judgments to the continuum would be that the probability of rating an item as solvable by the MCS decreases with an increasing difficulty of the item. This can be modeled with a logistic regression model where the classification of an item i is modeled as a function of the difficulty of the item (β_i), the severity of the judge g (a_g) and the strength of the link with the difficulty estimate (d_g). When interjudge differences are allowed for both parameters, the model reads as:

$$P(y_{ig} = 1 \mid \beta_i, a_g, d_g) = \frac{\exp\left(\alpha + a_g + (\delta + d_g)\,\beta_i\right)}{1 + \exp\left(\alpha + a_g + (\delta + d_g)\,\beta_i\right)}. \tag{4}$$

Again it is assumed that $a_g \sim \mathrm{N}(0, \sigma_a^2)$ and $d_g \sim \mathrm{N}(0, \sigma_d^2)$. The cut score can be calculated as $-(\alpha + a_g)/(\delta + d_g)$.

In Equation (4), the classification of an item is modeled in an equivalent way as the classification of the students in Equation (3). However, the signs of the slope parameters are likely to be opposite. For the Angoff judgments, the probability that an item is classified as an item that should be mastered by the MCS is modeled. Hence, a negative slope parameter can be expected, indicating a lower probability of classifying an item as able to be mastered by the MCS when β_i increases. For the contrasting-groups judgments, a positive slope parameter can be expected, indicating a higher probability of a student being classified as a master with an increasing θ_p. If the Angoff judgments were coded in the opposite way, the same sign of the slope is expected.

Goal of the Present Study

Since it has been widely accepted that different methods of standard setting lead to different results, the present study wants to go further than merely comparing the results of an examinee- and item-centered method of standard setting. The study investigated to what extent the judgments of both standard-setting methods are made according to the implicit assumptions made in both methods. In this way, an empirical criterion to evaluate the performance of an item- and examinee-centered method of standard setting might be provided.

A within-subjects design was used to investigate the differences in model parameters between the Angoff and the contrasting-groups method of standard setting. Given that each judge performed both the item- and examinee-centered classification, the implicit value of the MCS along the latent continuum can be expected to remain the same during both classification tasks for each judge. Each judge classified the same set of items. However, since the judges were teachers, each judge rated only the limited subsample of their own students. The use of such a split-plot design was not a problem for the study, as all students took the same test, and, hence, received a comparable ability estimate θ_p. However, the use of a different subsample for each rater may create an extra source of variation for the examinee classification

data, which may show up in the random effect parameters in the logistic regression model.

Apart from looking at the differences in slope and intercept for the two standard-setting methods, the within-subjects design makes it possible to compare the resulting cut scores, both with respect to their mean and with respect to their correlation across judges. Differing cut scores having a high correlation would only indicate that there is a main effect of the method: one method elicits stricter judgments than the other. A low correlation, even when the overall resulting cut scores are similar, would indicate that there is some kind of interaction between the judge and the method, such as the use of other criteria depending on the judgment task or a difference in complexity of both judgment tasks.

Method

Context

In 1997, the Flemish Parliament issued a set of attainment targets for primary education, which specify the basic competencies children should master when they leave primary education at the age of 12, after 6 years of school. The attainment targets are minimum objectives the educational authorities consider both necessary and feasible. Through national assessments, the authorities investigate at the level of the educational system the mastery of a particular set of attainment targets.

For the children who do not reach the attainment targets at the end of primary education, there is the possibility to follow 1 year of remedial teaching at the beginning of secondary education. The decision is made by the parents, advised by the teacher of the final grade, the headmaster, and the educational counselling office of the primary school.

Hence, the attainment targets for primary education are a familiar frame of reference to teachers in the sixth grade of primary education. They function as guidelines for the teachers for day-to-day educational practice, both with respect to the content of what they teach and with respect to the guidance given to their students. The present study pertains to the attainment targets for biology in primary education.

Test

A test was composed to measure the students' performance on the attainment targets for biology in primary education. The total item set consisted of different subsets of items, each referring to a cluster of attainment targets (e.g., all attainment targets related to health care). The 162 items of the test were administered at the end of the school year using an incomplete block matrix design. The items were of mixed format. Open-ended questions as well as multiple-choice items were used. All items were scored dichotomously. Using the overlap in the test design, it was possible to construct an IRT-scale using the two-parameter normal ogive (2PNO) item response model (Lord & Novick, 1968). In this model, the probability that a person p ($p = 1, \ldots, n$) answers item i correctly equals:

$$\Pr(\mathrm{X}_{pi} = 1) = \Phi\left(\alpha_i(\theta_p - \beta_i)\right), \tag{5}$$

where $\Phi(.)$ denotes the cumulative standard normal distribution, α_i the item discrimination parameter, θ_p the person ability parameter, and β_i the item difficulty parameter.

Participants

For a calibration study of the test, a stratified random sample of Flemish schools offering primary education was drawn. Three stratification variables were used: educational sector, province, and school size. Within a sampled school all classes of the sixth grade participated in the study. In Flanders, each class usually has only one teacher. Within a class, the different test booklets were distributed according to a spiraling design.

Fifty-one out of the 61 schools in the sample drawn participated in the study. Eighty-seven classes and their respective teachers were involved. The Angoff judgments were completed by 69 teachers, 73 teachers performed the contrasting-groups judgments, and 63 teachers completed both tasks. Only the data of the pupils for whom a contrasting-groups judgment was available were included. Because of missing data in their test booklets, a limited number of pupils were excluded from the analysis. The analyses reported here were performed on the data of 1,321 pupils.

Standard-Setting Tasks

While the students were taking the test, the teacher of each participating class performed the two standard-setting tasks. On the one hand, they had to make a classification of their students in a group of masters and nonmasters. Because there is a clear parallel between this person-focused version of the contrasting-groups method (Brandon, 2002) and the possible orientation toward remedial teaching decided at the end of primary education, one can assume that teachers are quite familiar with this kind of classification. In this study, however, the teachers had to confine their judgment to the attainment targets in biology.

On the other hand, the teachers had to make a dichotomous Angoff judgment of all the test items, including those that were not presented to their pupils. The teachers were asked to indicate for each item whether the student that just reaches the attainment targets in biology *should*[1] be able to solve it. The Angoff procedure was applied in one judgment round only, for three reasons. Firstly, it was not possible, in practice, to bring the teachers together for other rounds. Secondly, the initial individual judgment is not confounded with possible effects of social interactions, group processes, or additional information provided between rounds. Finally, the effect of an extra judgment round is shown not to be unequivocal. In a meta-analysis, Hurtz and Auerbach (2003) showed that the effect depended on what was done in between the judgments. If judges were allowed to discuss their judgments, the average judgment increased. If they were provided normative data about the performance of the students, the average judgment declined.

The instructions for performing both judgmental tasks were given on paper. The definition of what minimal competency in the domain means (the performance standard), was provided by the attainment targets for biology. As a reminder, the teachers

were provided with the relevant attainment targets. Since teachers are familiar with this performance standard, preparatory training was not provided.

Estimation

A Bayesian Approach

The models for the students' data and for the teachers' judgments were estimated in a Bayesian way, using a Markov Chain Monte Carlo (MCMC) estimation procedure (Gelman, Carlin, Stern, & Rubin, 2004; Gilks, Richardson, & Spiegelhalter, 1996; Tanner, 1996). MCMC methods are devised to simulate random observations from complex and high-dimensional probability distributions. In Bayesian estimation, they are used to evaluate the posterior distribution of all the model parameters. By simulating a large sample from this distribution, the characteristics of the posterior distribution can be estimated. The posterior distribution of a parameter gives an idea of the posterior uncertainty of the parameter at hand, given the data and the prior distributions. The mean of the posterior distribution can be used to summarize the distribution and serves as an estimate of the parameter. The standard deviation of the posterior distribution of a parameter gives an indication of the standard error of estimation.

The reasons for using a Bayesian estimation procedure were threefold. First, the Bayesian method allows for estimating the models for the students' test responses and the teachers' judgments jointly. In this way, the estimation error of the predictors (student ability and item difficulty) is built into the estimation uncertainty for the regression weights.

A second advantage of the Bayesian approach is that the random-effects parameters themselves are sampled from the posterior distribution together with the hyperparameters describing the distribution of the random effects. In maximum likelihood estimation, a two-step procedure is needed to get estimates of the random effects. In the first step, the distribution of the random effect is estimated. In a second step, an empirical Bayes procedure can be used to get estimates of the individual random effects.

Finally, a Bayesian estimation procedure can easily implement a design with missing data. Because of the use of an incomplete block matrix design, every student only responded to a subset of the items. The data that are missing can be considered to be missing by design or missing completely at random (Rubin, 1976). Hence, a missing data imputation mechanism can be used for estimation purposes. In the context of Bayesian estimation of IRT models, Patz and Junker (1999) proposed to impute data by generating for each draw, from the posterior, the probability of a correct response for student p on item i using the estimated person and item parameter. This probability is then dichotomized by comparing it to a random draw u from the uniform distribution on the interval [0,1]. The value of the missing response is set to 1 if the calculated probability of a correct response exceeds u and set to 0 otherwise.

Data Augmented Gibbs Sampling

The most widely used form of MCMC in statistical applications is Gibbs sampling. The Gibbs sampler takes a modular approach to sampling from the posterior

distribution. It creates a Markov chain by successively sampling from the set of full conditional distributions, so that once this Markov chain has converged to its equilibrium distribution, the sampling is done from the posterior distribution. Albert (1992) showed that for IRT models (and logistic regression functions in general) the Gibbs conditionals can be derived analytically using a data augmentation step. This is true only if a probit link function is used. Therefore, probit regression was used for the analyses of the teacher judgments, and the normal ogive formulation of the two-parameter IRT model was used for the students' data (as was indicated in Equation 5).

Distributional Specifications

For the estimation of the 2PNO model for the students' data, it was assumed that all students come from a normal distribution with mean 0 and variance 1 to identify the latent scale. For the items, it was assumed that all items come from one population with separate normal distributions for item difficulty $N(\mu_\beta, \sigma_\beta^2)$ and for item discrimination $N(\mu_\alpha, \sigma_\alpha^2)$. By specifying a normal distribution for the item discrimination parameter, the model incorporates the theoretical possibility of negative item discrimination. The distributional specifications correspond to a hierarchical IRT model at the item side (Janssen, Tuerlinckx, Meulders, & De Boeck, 2000) with one population, as is implicitly assumed when specifying a single prior distribution for each set of item parameters.

Using Bayesian estimation, prior distributions were needed for all hyperparameters. All priors were chosen to be noninformative. For the hyperparameters indicating a mean (μ_β, μ_α, α, and δ), a flat prior was used. For the variance parameters (σ_β^2, σ_α^2, σ_a^2, and σ_d^2), an inverse-χ^2 distribution with one degree of freedom was taken (Gelman et al., 2004).

Gibbs Conditionals

Given the prior distributions and the data augmentation step in both the model for the students' data and teachers' data, it was possible to derive Gibbs conditionals in closed form for all model parameters. More specifically, the conditional distributions were truncated normal distributions, normal distributions, or scaled inverse-χ^2 distributions, from which one can sample directly using standard routines.[2]

Convergence of the Markov Chain

A critical issue in applying MCMC techniques is the question how long the Markov chain must be in order to be confident that it has converged, and, hence, that one actually samples from the posterior distribution. Two checks of convergence were applied. Firstly, the convergence measure $\sqrt{\hat{R}}$ proposed by Gelman and Rubin (1992) was calculated for each model parameter. This measure is based on the idea that chains, running from different starting points in the parameter space, have reached convergence at the point where the variance of the sampled parameter values between the chains approximates the variance of the sampled parameter values within the chains. The measure $\hat{R}$ quantifies the estimated ratio of between-

chain variation to within-chain variation of a single parameter. $\sqrt{\hat{R}}$ should be smaller than 1.1 for all parameters of the model for the MCMC algorithm to converge (Kass, Carlin, Gelman, & Neal, 1998). Secondly, for different batches of consecutive draws from the posterior, the posterior mean and posterior standard deviation were calculated for each parameter (Hoijtink & Molenaar, 1997). Small differences among the batches indicate that the Gibbs sampler converged.

Specifications of the Markov Chains

Five Markov chains were run using random starting points. To estimate the 2PNO model for each θ_p and each β_i, a starting value was sampled from the standard normal distribution. For α_i a starting value was sampled from the uniform distribution with a minimum of 0 and a maximum of 2. The starting points for the hyperparameters consisted in the mean and variance of the starting values for the item difficulty and item discrimination parameters. For the probit regression models, the starting values for the intercept and the slope were randomly sampled from the standard normal distribution. The hyperparameters representing the mean were also sampled from a standard normal distribution, while the variance hyperparameters were drawn from the uniform distribution. Each chain was run for 10,000 iterations. The first 5,000 iterations of each chain were discarded. The remaining 5,000 iterations were used to calculate the convergence measure $\sqrt{\hat{R}}$.

Results

Descriptive Analyses

For the Angoff task, all the judges found that the majority of the items should be solved by the MCS. The percentage of items to be solved by the MCS varied considerably across judges, namely from 56% to 99%. This variation reflects differences in severity, since every teacher had to judge the same items. The median percentage was 83%.

For the contrasting-groups task, the median percentage of students classified as masters by their teacher was 79%. Again there were large differences among judges. One teacher classified only 20% of the students as masters and nine teachers classified all their students as masters. This does not mean that the former teacher is the strictest teacher and the latter teachers were the most lenient, since it is possible that the former teacher had the weakest performing students. Variation in the ratings was somewhat larger for the contrasting-groups judgments than for the Angoff judgments, which may be due to the fact that the former judgments reflect not only differences in severity, but also differences in the ability of the students who were judged by the teacher. The probit regression model takes these differences in ability into account.

Convergence of the Markov Chain

The convergence of the Markov chains was satisfactory. The convergence measure $\sqrt{\hat{R}}$ varied between .999 and 1.047 with a median of 1.001 for all the parameters in the analysis. An analysis comparing parameter estimates across different batches

TABLE 1
Estimates of the Hyperparameters (with Standard Errors) for both Random-Effects Probit Regression Models

	Contrasting Groups	Angoff
Intercept	1.17 (.09)	.64 (.06)
Slope	.96 (.09)	−.22 (.02)
$\sigma^2_{\text{intercept}}$	.25 (.08)	.23 (.05)
σ^2_{slope}	.17 (.06)	.02 (.00)

was in line with the above results. The last 5,000 iterations of the five chains were divided in 50 batches of 500 subsequent draws from the posterior distribution. The means and standard deviations of each parameter calculated in those batches were stable across batches.

Modeling the Students' Data

The posterior means of the person ability parameters varied between −3.10 and 2.98 with a mean of –.001. For the item difficulty parameters the posterior means varied between −5.61 and 4.63 with a mean of −1.45. Since the origin of the scale was fixed at the mean of the prior distribution of the ability parameters, this implies that most of the items were relatively easy for the students. The posterior means of the item discrimination parameters varied between .09 and 1.05 with a mean of .43.

Modeling Teachers' Judgments

Table 1 gives a summary of the resulting hyperparameters describing the distribution of the judges for the Angoff data as well as the contrasting-groups data. The intercept for the contrasting-groups judgments indicates that a student with an average ability (i.e., $\theta_p = 0$) has a probability of .76 to be classified as a master. According to the intercept for the Angoff data, an item with a difficulty of 0 has a probability of being judged as to be mastered by a MCS is .65. For the average item, this probability is .72. Both the values of the hyperparameters for the slopes of the Angoff judgments and the contrasting-groups judgments were significant, but there is a much weaker link between the Angoff judgments and the position of the items on the IRT scale than between the contrasting-groups judgments and the ability of the students.

For both methods, the variance parameters for the slopes and the intercepts were significant, which indicates that it was crucial to include interjudge differences in the model. For the Angoff judgments, however, teachers did not seem to differ that much in the degree that their judgment is related to the difficulty of the item. For the contrasting-groups judgments, the variance for the slopes was somewhat larger. This might be a result of the split-plot design applied, as the judges did not classify the same students.

To get a finer picture of the interjudge differences, the posterior means of the regression parameters were calculated for each individual teacher and a cut

TABLE 2
Descriptive Statistics of the Posterior Means of the Probit Regression for the Contrasting Groups and the Angoff Data with the Resulting Cut Scores

	Contrasting Groups			Angoff			
	Intercept	Slope	Cutoff*	Intercept	Slope	Cutoff*	Cutoff†
Minimum	.24	.48	−3.42	−.06	−.36	−0.32	−1.00
Mean	1.17	.96	−1.29	.64	−.22	3.36	.84
Maximum	1.77	1.36	−.33	1.81	−.06	26.69	3.51

*Defined as $-(\alpha+a_g)/(\delta+d_g)$.
†Derived from van der Linden (1982).

score was derived on the basis of them. Table 2 presents a descriptive summary of the posterior means of both the intercept and slope and of the corresponding cut score for both methods. Note that results on one row are not necessarily referring to the results of the same judge.

Table 2 shows that for both standard-setting methods, the mean of the individual random effects for the intercept and slope were very similar to the estimated corresponding hyperparameters as given in Table 1. Comparing the slope parameters shows that the absolute values for the Angoff data were considerably smaller than those for the contrasting-groups data. From the range of the individual random effects, one can see that the lowest value for the slope parameter of contrasting-groups results (.48) was still considerably higher in absolute value than the highest value of the Angoff results (–.36). The cut scores for the contrasting-groups method were all below a student with an average ability ($\theta_p = 0$). For the highest cut score, the average student has a probability of .58 to be classified as a master. For the lowest cut score this probability was .97 and for the average cut score .78.

Because of the low values of the slope parameters for some judges for the Angoff method, the cut scores based on the regression parameters ($-\alpha/\delta$) were very extreme. The highest cutoff was set at 26.69 on an IRT-scale with an actual ability range from −3.10 to 2.98. Therefore, an alternative cut score was derived using an adaptation of the method proposed by van der Linden (1982). For each teacher, the judgments of the items were treated as the response pattern of a fictitious MCS and converted into a score on the IRT scale. This score can then be considered as the cutoff for mastery classification. The resulting cut scores are presented in the last column of Table 2, and are used in the next section to calculate the number of students reaching the cutoff.

Mastery Classification

Table 3 presents the percentage of masters in the total sample that reached the cutoff for both methods of standard setting. The median of the Angoff judgments resulted in a cutoff that was reached by 18% of the students. Teachers differed quite drastically. The lowest cutoff resulted in 86% masters, while none of the students reached the level implied by the highest cutoff. The gap in percentage of masters between the 25th percentile cutoff and the 75th percentile cutoff was almost 50%.

TABLE 3
Percentage Masters based on the Cut Scores Derived from the Angoff and Contrasting-Groups Judgments

Cut score	Min	p25	Med	p75	Max
Angoff	86	50	18	4	0
Contrasting Groups	100	94	89	85	66

The cutoffs resulting from the contrasting-groups data resulted in a different picture. Even the highest cutoff for this method resulted in 66% masters. The lowest cutoff was reached by all the students. Based on the median of the resulting cutoffs, 89% of the students were classified as masters. The gap between the p25 cutoff and the p75 cutoff was only 9%, indicating that there was a much higher agreement among judges for this procedure. For 63 teachers, an Angoff cutoff as well as a contrasting-groups cutoff was available. The correlation between both was only .20.

Discussion

Toward a New Empirical Criterion?

Kane (1998a) pointed to the absence of an empirical criterion to choose between an examinee- or item-centered standard-setting method. In the present paper, a new approach to the comparison of an examinee- and an item-centered method of standard setting is presented. Essential for the validity of a standard-setting procedure is that the judgments follow the implicit assumptions of the standard-setting task. In the Angoff method, this concerns an inverse link of the judgments to the difficulty of the items; in the contrasting-groups method, a positive link of the judgments to the ability of the students is expected. Random-effects probit regression was applied on data of a contrasting groups and a dichotomous Angoff standard-setting procedure to investigate those assumptions.

A possible empirical criterion can be found in the slope parameter of the regression that reflects the degree to which judgments are consistent with the assumptions underlying the standard-setting method. Based on this criterion, the contrasting-groups method seemed to be the better one in the present study. The judgments expressed in the contrasting-groups method were more in line with the implicit assumption as to how the judgments are linked to the underlying continuum than the judgments collected using the dichotomous Angoff method.

The familiarity of the teachers with the students in the contrasting-groups method used in this study could have caused criterion bias (Kane, 1998b). Teachers might base their judgment on the student's knowledge and skills that only partly overlap with the actual content of the test (Brandon, 2004). The results of the regression made clear that the judgments were, to a significant degree, associated with the ability of the students in the domain of biology, so criterion bias does not seem to be at play to a large extent.

Several explanations can be proposed for the weak link between the dichotomous Angoff judgments and the difficulty of the item. First, it is possible that some

teachers interpreted the Angoff judgment instruction as a validity issue, rather than as an assessment of the required minimal competency level. That would mean that the judgments expressed by the teachers reflect whether they considered the items as good operationalizations of the attainment targets, rather than whether they thought the MCS should be able to master it. This might also account for the high cut scores being set, assuming the test consisted of valid items. Second, it may be the case that the teachers did use an estimation of the difficulty as a basis for their judgment but that the estimation was not quite accurate, including the estimation of the relative difficulty of an item. Bejar (1983) already stated that the ability of judges to estimate item difficulty is questionable. Impara and Plake (1998) found this was even the case when judges have a high degree of familiarity with the examinees and the test, which is the case for the teachers in this study. Finally, one might think of the high number of items scored "1" as a possible explanation for the low values of the slope parameter. In this case, there is not that much information available in the data to accurately estimate the slope parameter. This can only be a partial explanation though, since a high number of 1-scores was given for the contrasting-groups data as well.

The Resulting Cut Scores

The cut scores resulting from the Angoff judgments were quite high. The median cut score was reached by only 18% of the students. Several possible explanations for those high cut scores can be suggested. The first explanation refers to the phrasing of the Angoff task: in the present study, judges were asked to indicate which items the MCS *should* be able to solve. Although Angoff made no distinction between the phrasings would, could, and should, Impara and Plake (1997) stated that "should is typically interpreted as a target that is higher than how well examinees *will* perform" (p. 363). The phrasing might elicit the expression of an ideal, rather than an intended minimal performance standard.

Another possible explanation is related to the fact that the dichotomous version of the Angoff method was used. As Reckase (2004) showed, the obligation to use only 0 and 1 as a judgment can have an impact on the resulting cut score. The direction of the impact depends on the general difficulty of the test. As an artificial example, Reckase (2004) considered a test of five items, and for each item the intended MCS has a probability of .80 to solve it correctly. Assuming that a judge can estimate the actual success probability of the MCS accurately, because of the obligation to round this number into a dichotomous score, however, the judge would have to give a score of 1 to each item. This results in a cutoff being set at a score of 5 and not at a score of $4 (= 5 \times .80)$, which is the intended cut score (Reckase, 2006). Hence, if a test has a majority of easy items and if the judges are quite accurate in estimating the success probabilities of the MCS, one can expect the dichotomous Angoff procedure to result in too high a cut score. On the other hand, if the test has a majority of difficult items the opposite effect is expected, causing the cutoff to be too low. In our study the items were relatively easy for the students. Hence, according to the above reasoning one can expect the Angoff cutoffs to be set too high for the intended standard. The link between the two essential components (the performance standard and the cutoff) in standard setting is lost this way, which is a problem for the validity of the procedure

(Kane, 2001). Impara and Plake (1997) argued that in a group of judges, the intended probability could still be reflected using a dichotomous judgment. The proportion of 1s over judges and over items might reflect the performance of the MCS. An accurate estimate of the intended standard in the example of Reckase (2004) can only be accomplished if some judges would not make an accurate judgment of the performance of the MCS and assigned a 0-score to a clearly easy item. This would imply that the better judges would become at estimating the item difficulty, the worse the cut score would represent the standard.

Compared to the Angoff results, the cutoffs for the contrasting-groups method were set lower. The median cut score was reached by 89% of the pupils. A criticism to the applied contrasting-groups method might be that teachers were very lenient in judging their own pupils and, thus, are setting the cutoff too low. A few elements contradict this position. First, the attainment targets were set in such a way that the vast majority of the students are supposed to be able to reach them by the end of primary school. Moreover, only about 10% of the pupils in Flanders continue to the remedial years in secondary education, and it is this group of pupils that is supposed not to master the attainment targets. Second, the median percentage of pupils classified as masters was 79%, and one teacher even classified only 20% of his pupils as masters. This shows that teachers were relatively uninhibited in classifying some of their students as nonmasters. Although not conclusive, these arguments indicate that some teachers were not overly lenient in judging their pupils.

The low correlation between the resulting cutoffs shows that there was not only a main effect of the method (which still could result in a high correlation) that was explaining the different standards, but also that other elements were influencing the judgments. It shows an interaction between the judge and the method, such as the use of other judgment criteria depending on the procedure or a difference in complexity of both procedures.

Limitations of the Present Study

In the present study, teachers judged their own students who made up the target population of the assessment. This was possible because of two characteristics of the context of the study. First, it concerned a "low-stake" testing situation with no consequences for the teachers or the students. Second, the teachers were already familiar with the performance standard as the attainment targets guide their day-to-day teaching practice. In another context, a specific training of the teachers may be needed, because the teachers are not acquainted enough with the performance standard involved.

The study also does not exclude that results might be different for the Angoff method if a more extensive procedure had been used (e.g., including training, multiple rounds with discussion, impact data). However, discussion and providing impact data will mainly affect the cut score, rather than the concordance with the implicit task assumptions. Extensive training, on the other hand, may be needed for improving the concordance and, hence, the performance of the Angoff procedure on the proposed empirical criterion. In any case, the presented regression model provides a way to evaluate these alternative Angoff procedures and the effect of training on the concordance with the implicit task assumptions.

The present study suggests that the contrasting-groups method might be a promising approach to standard setting if one wants to involve a large group of people from the educational field. This involvement can work in a low-stakes testing situation as a way of increasing the support for the resulting standard in the educational community, but limits the possibilities for training, instruction, and discussion of the judgments in several rounds.

Acknowledgments

The present research was supported by the OBPWO grant 01.08 and the 2006–2011 national assessments research grant of the Department of Education of the Ministry of the Flemish Community (Belgium).

Notes

[1]Zieky (1995, in Impara & Plake, 1997) indicated that Angoff made no distinction between could, would and should.

[2]The full conditional distributions are available on request from the authors.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

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.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303–310.

Berk, R. A. (1996). Standard-setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9,* 215–235.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: A review. *Measurement and Evaluation in Counseling and Development, 35,* 167–181.

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

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science, 7,* 457–511.

Gilks, W., Richardson, S., & Spiegelhalter, D. (Eds.) (1996). *Practical Markov Chain Monte Carlo*. New York: Chapman & Hall.

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

Haertel, E. H., & Lorié, W. A. (2004). Validating standards-based test score interpretations. *Measurement, 2,* 61–103.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62,* 171–189.

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.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34,* 355–368.

Impara, J. C., & Plake, B. S. (1998). Teacher's ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35,* 69–81.

Jaeger, R. M. (1989). Certification of student campetence. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed., pp. 485–514). New York: American Council on Education and Macmillan.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

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

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

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

Kane, M. T. (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.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov Chain Monte Carlo in practice: A roundtable discussion. *The American Statistician, 52,* 93–100.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores*. Princeton, NJ: Educational Testing Service.

Livingston, S., & Zieky, M. (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.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McGinty, D. (2005). Illuminating the "Black Box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education, 18,* 269–287.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery in standard setting. *Review of Educational Research, 45,* 133–158.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum.

Reckase, M. D. (2004). What if there were a "True Standard Theory" for standard setting like the "True Score Theory" for tests? *Measurement, 2,* 114–119.

Reckase, M. D. (2006). 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.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York: Springer.

Van der Linden, W. J. (1982). A latent trait model for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295–308.

Zieky, M. J. (2001). So much has changed: How the setting of cut-scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum.

Authors

DANIËL VAN NIJLEN is a Research Assistant at the Centre for Educational Effectiveness and Evaluation at the Katholieke Universiteit Leuven, Dekenstraat 2 (PB 3774), 3000 Leuven, Belgium; daniel.vannijlen@ped.kuleuven.be. His primary research interests include psychometric methods.

RIANNE JANSSEN is an Assistant Professor at the Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Dekenstraat 2 (PB 3774), 3000 Leuven, Belgium; rianne.janssen@ped.kuleuven.be. Her primary research interests include educational measurement.