Psychological Methods
2017, Vol. 22, No. 4, 705–724

1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000124

Using Expert Knowledge for Test Linking

Maria Bolsinova and Herbert Hoijtink
Utrecht University and Cito, Dutch Institute for Educational Measurement

Jorine Adinda Vermeulen
Cito, Dutch Institute for Educational Measurement and University of Twente

Anton Béguin
Cito, Dutch Institute for Educational Measurement

Abstract

Linking and equating procedures are used to make the results of different test forms comparable. In the cases where no assumption of random equivalent groups can be made some form of linking design is used. In practice the amount of data available to link the two tests is often very limited due to logistic and security reasons, which affects the precision of linking procedures. This study proposes to enhance the quality of linking procedures based on sparse data by using Bayesian methods which combine the information in the linking data with background information captured in informative prior distributions. We propose two methods for the elicitation of prior knowledge about the difference in difficulty of two tests from subject-matter experts and explain how these results can be used in the specification of priors. To illustrate the proposed methods and evaluate the quality of linking with and without informative priors, an empirical example of linking primary school mathematics tests is presented. The results suggest that informative priors can increase the precision of linking without decreasing the accuracy.

Translational Abstract

If each year a new version of an educational test is used then the results of the new version are not directly comparable to the results of the reference test version due to the difference in difficulty of the 2 tests and the differences in the ability of the new and reference populations of students. Typically extra data (other than the examination data) are collected to compare the difficulty of the items in the two tests. However, in high stakes testing, where the amount of data that are available to link the reference and the new test are limited due to security reasons, these extra data do not provide enough information to obtain the desired levels of certainty. In this article we argue that linking data are not the only source of information about the difference in the difficulty of the two test forms. Experts may also provide information about this difference. In the study we propose and evaluate two methods for elicitation of the prior knowledge about the difference in difficulty of two tests from subject-matter experts. The results prove the utility of the proposed methodology, since the precision of the linking results increases without the increase in bias.

Keywords: elicitation, expert knowledge, informative priors, test equating, test linking

Supplemental materials: http://dx.doi.org/10.1037/met0000124.supp

If different test forms of an educational test measuring the same ability are administered to different populations of students (e.g., from different years), their results are not directly comparable because of the differences in the difficulty of the tests and the differences in the ability in the populations. Linking and equating techniques are ways to make the scores on the tests comparable. For linking a current test to the reference test form different linking designs can be used (Angoff, 1971; Kolen & Brennan, 2004; Lord, 1980; Petersen, Kolen, & Hoover, 1989; Wright & Stone, 1979).

In high-stakes testing (e.g., examinations) different test forms often do not have items in common due to security reasons. If the forms are administered under the assumption of nonequivalent groups it is necessary to collect additional data to link the different test forms (Mittelhaëuser, Béguin, & Sijtsma, 2015). Most commonly a type of anchor test is used, but the administration of anchor tests under appropriate conditions is challenging and expensive (Keizer-Mittelhaëuser, 2014). In this article we consider a situation in which two test forms can be connected through the so called linking groups in a pretest nonequivalent group design

This article was published Online First April 3, 2017.

Maria Bolsinova and Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University, and Psychometric Research Center, Cito, Dutch Institute for Educational Measurement; Jorine Adinda Vermeulen, Psychometric Research Center, Cito, Dutch Institute for Educational Measurement, and Department of Educational Measurement and Data analysis, University of Twente; Anton Béguin, Psychometric Research Center, Cito, Dutch Institute for Educational Measurement.

Maria Bolsinova is now at the Department of Psychology, University of Amsterdam.

Correspondence concerning this article should be addressed to Maria Bolsinova, Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam, the Netherlands. E-mail: m.a.bolsinova@uva.nl

(Béguin, 2000), because that is common for high-stakes examinations in the Netherlands, but the methodology developed in this article can be also used with different linking designs.

When item response theory (IRT) is used for linking, item parameters of the items in the current and the reference tests have to be placed on the same scale (Kolen & Brennan, 2004; von Davier, 2011). This can be done either by estimating the IRT parameters in the two tests separately and then placing them on the same scale using scale transformation methods (Haebara, 1980; Loyd & Hoover, 1980; Marco, 1977; Stocking & Lord, 1983), or by estimating the parameters of the two test forms together in concurrent calibration (Wingersky & Lord, 1984). Once the item parameters are put on the same scale, the predicted score distribution of the reference population on the current test can be estimated. A cut-score (i.e., a minimum number of correct responses needed to pass the test) for the current test can be determined using IRT observed score equating using equipercentile linking (Lord & Wingersky, 1984).

Unlike examination data, which usually consist of responses of thousands of students to test items, the linking data are often not sufficiently large, such that the uncertainty about the difference between the difficulties of the two tests and, hence, about the new cut-score is rather large. These data are often collected in nonexamination conditions (with different levels of stress and motivation) and not from the populations of interest. Thus, the linking data often do not provide a high enough quality of linking (in terms of uncertainty and bias of the cut-scores).

From the Bayesian perspective, the data are not the only source of information about the item parameters. A second source is the background information which can be captured in the prior distributions. It has been advocated that using informative priors is useful in practical applications (Goldstein, 2006; Vanpaemel, 2011). The purpose of this study is to develop methods to improve the quality of linking by combining the information from the linking data and the informative priors. We explore different ways of eliciting the prior distributions about the difference in difficulty of the two tests from subject-matter experts and using them to link the two tests. In this study we focus on tests that involve the use of cut-scores to render classification decisions (e.g., pass/fail or different levels of mastery), but the methods for elicitation of prior distributions are applicable to other contexts as well.

There have been studies with a focus on specification of informative priors for the parameters of IRT models. Item features (e.g., operations involved in solving the item, number of response alternatives, or use of negation in the formulation of the item) can be used to predict the item parameters (Fisher, 1973; Tatsuoka, 1987), which can be included as prior knowledge for Bayesian estimation of the item parameters (Mislevy, 1988). This source of prior information has been also used in the context of test equating and linking (Mislevy, Sheehan, & Wingersky, 1993). However, information about item features that are good predictors of the difficulty is not always available. Other authors include judgmental information from subject-matter experts in the estimation of the item parameters (Bejar, 1983; Ozaki & Toyoda, 2006; Swaminathan, Hambleton, Sireci, Xing, & Rivazi, 2003; Wauters, Desmet, & van der Noordgate, 2012). The latter has not been done in the context of test linking, and the expert judgments were only used to improve the estimation of the individual item parameters. Judgmental information about the items difficulties is also collected in the context of standard setting (Cizek & Bunch, 2007; Geisinger, 1991; Shepard, 1980). In some standard setting procedures the cut-score for the test is selected based solely on expert judgments, in others after the experts' judgments are collected experts are informed about the results from the linking data and can update their judgments before establishing the cut-scores. However, this updating is not done in a formalized way of quantifying expert knowledge as prior distributions which are combined with data using Bayesian statistical inference. In our study we develop a formal way of combining prior knowledge elicited from subject-matter experts with information from linking data.

Experts' judgments of individual items are often not reliable, however judgments about sets of items are more reliable because there is less variation in the means than in individual observations. Hence, combined on the test level expert judgments can provide valuable information about the relations between two tests. Therefore, we argue that the expert knowledge about the item difficulties is especially useful for test equating and linking. Another reason for a special interest in using experts' judgments in the context of test linking is that from the examination data we can estimate the differences between the item difficulties within the reference test and within the current test with high precision and the only thing that is missing is the information about the relations between the tests. Therefore, the information available from the examination data can be used to help in obtaining more valid and reliable judgments with respect to the relations between the two tests.

This article is structured as follows. First, the measurement model and the equating design used throughout the article are discussed. Then, in Section Elicitation of prior knowledge about the difference between the difficulty of two tests, we propose two methods for elicitation of the prior knowledge about the test difficulty from experts. The first one is an adaptation of the Angoff standard setting procedure (Angoff, 1971). The second method was designed by us for more direct elicitation of the experts' knowledge about the differences between the difficulties of the two tests. In Section 4 Empirical example, the two elicitation methods are compared in terms of the quality of linking with the elicited priors using an empirical example based on the primary school mathematics test "Entreetoets Groep 7." The article is concluded with a discussion.

Measurement Model and Equating Design

In this study the marginal Rasch model (Rasch, 1960) is used assuming a normal distribution for proficiency. It models the probability of a correct response to an item in a population:

$$\Pr(X_i = 1) = \int_{\mathbb{R}} \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)} \mathcal{N}(\theta;\ \mu, \sigma^2) d\theta, \qquad (1)$$

where X_i denotes a binary coded response (1 for correct and 0 for incorrect) to item i with difficulty δ_i of a person randomly sampled from the population with the mean and the variance of proficiency θ equal to μ and σ^2. The Rasch model was chosen because it has a clear interpretation of the item difficulty. If $\delta_i > \delta_j$, then both the conditional (i.e., given a particular value of θ) and the marginal probability (Equation 1) of a correct response to item i is smaller than to item j. This is important when translating experts' judgments of the type "Item i is more difficult than item j" into statements about the model parameters ($\delta_i > \delta_j$). This is not possible if an item discrimination parameter is added to the model, like is done in the two parameter logistic model (Lord & Novick, 1968). We assume that all the items, both in the current and in the reference test, have the same discriminative power. Although the Rasch model is a rather restrictive model, it has been shown that equating results using the Rasch model are rather robust to model violations (Béguin, 2000).

We consider a pretest nonequivalent group equating design with *G* linking groups. This design is visualized in Figure 1, where rows represent persons and columns represent items. We denote the data matrix of the reference exam by **X**, the data of the current exam by **Y** and the data of the *G* linking groups by $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_G$. By $\mathbf{X}^*$ we denote the unobserved responses of the reference population to the current test.

When concurrent Bayesian calibration is used for linking the two tests, samples from the joint posterior of the model parameters need to be obtained:

$$p(\boldsymbol{\delta}_r, \boldsymbol{\delta}_c, \mu_r, \sigma_r^2, \mu_c, \sigma_c^2, \mu_1, \sigma_1^2, \ldots, \mu_G, \sigma_G^2 \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_G), \tag{2}$$

where $\boldsymbol{\delta}_r$ and $\boldsymbol{\delta}_c$ are the vectors of item difficulties of the items in the reference and the current tests, respectively; μ_r and μ_c are the means of proficiency of the reference and the current populations respectively, σ_e^2 and σ_c^2 are the corresponding population variances, and $\mu_1, \sigma_1^2, \ldots, \mu_G, \sigma_G^2$ are the population parameters in the linking groups. A zero point for the IRT scale is fixed by setting the average difficulty of the items in the reference test equal to zero: $\bar{\delta}_r$. Using samples from the posterior in (Equation 2) the score distribution of the reference population on the current test (score distribution in $\mathbf{X}^*$) is estimated and the new cut-score is determined using equipercentile equating (see Appendix A, section Estimating the cut-score s_{pass}).

The linking data are the only data that provide information about the difference between the average difficulty of the items in the current test and the average difficulty of the items in the reference test, denoted by $\tau = \bar{\delta}_c - \bar{\delta}_r$; under the stated identification that $\bar{\delta}_r = 0$, this renders $\tau = \bar{\delta}_c$. Since the linking data are sparse, the largest part of the uncertainty about what the new cut-score should be is coming from the uncertainty about τ. We aim to increase the precision of the estimate of the new cut-score by including prior information about τ in the estimation. The following reparametrization is used throughout the article (see Figure 2):

$$\boldsymbol{\delta}_c^* = \boldsymbol{\delta}_c - \bar{\delta}_c = \boldsymbol{\delta}_c - \tau \tag{3}$$

$$\mu_c^* = \mu_c - \bar{\delta}_c = \mu_c - \tau, \tag{4}$$

The rest of the article is focused on the specification of the prior distribution of τ.

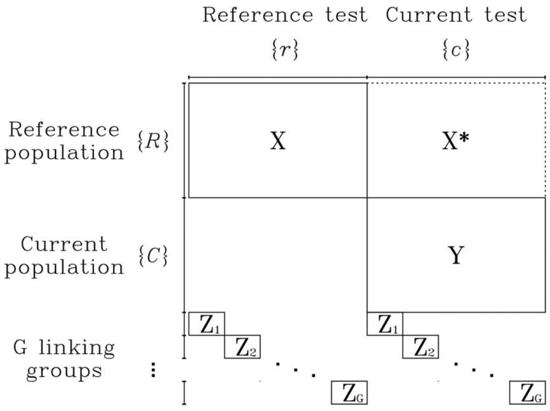


Figure 1. Equating design with *G* linking groups: rows represent persons, columns represent items.

Elicitation of Prior Knowledge About the Difference Between the Difficulty of Two Tests

Information about the difference between the average difficulty of the items in the current test and the average difficulty of the items in the reference test can be collected from subject-matter experts who can judge the difficulty of the items in the two tests. In this section we describe the two methods developed for the elicitation of the prior knowledge about τ. In the following section we compare the performance of these methods in an empirical elicitation study.

Both methods use item difficulties estimated from the examination data. Because we also used the data to calibrate the two test forms on the same scale, we need to divide the examination data (both from the reference and from the current exams) into two halves: the first half which is used to facilitate the elicitation of experts' knowledge about the mutual order of the items and to construct priors for the item and the population parameters, here called the training data, and the second half which is used for the estimation of the new cut-score, here called the estimation data (see Appendix A for technical details).

Adaptation of the Angoff Method for Elicitation of the Prior Knowledge About τ

The first method that we developed is an adapted version of the Angoff method of standard setting (Angoff, 1971). Unlike the regular use of the Angoff method and other standard setting procedures we use the experts' judgments not to set the cut-scores directly but use these judgments for the specification of the informative prior for τ. In this way the cut-scores can be estimated based on both the experts' knowledge and the linking data.

Traditionally in the Angoff method, each expert $e \in \{1{:}\,E\}$ from a panel of *E* experts is asked for each test item to give the probability that a minimally competent (borderline) candidate will answer this item correctly:

$$p_{ie} = \Pr(X_{ip} = 1 \mid \theta_p = \theta^*), \tag{5}$$

where θ^* is proficiency of a borderline candidate. These probabilities are then added over items to obtain the expected score of a borderline candidate which is chosen as a cut-score. In our study, we use the experts' evaluations of the probabilities p_{ie} differently. If each expert evaluates all items, then based on the Rasch model her/his estimate of the difference between the average difficulties of the items in the current and the reference test denoted by τ_e can be computed as

$$\tau_e = \frac{\sum_{i \in \{c\}} \ln\left(\frac{1-p_{ie}}{p_{ie}}\right)}{|c|} - \frac{\sum_{j \in \{r\}} \ln\left(\frac{1-p_{je}}{p_{je}}\right)}{|r|}, \tag{6}$$

where $\{r\}$ and $\{c\}$ are the sets of items in the reference and the

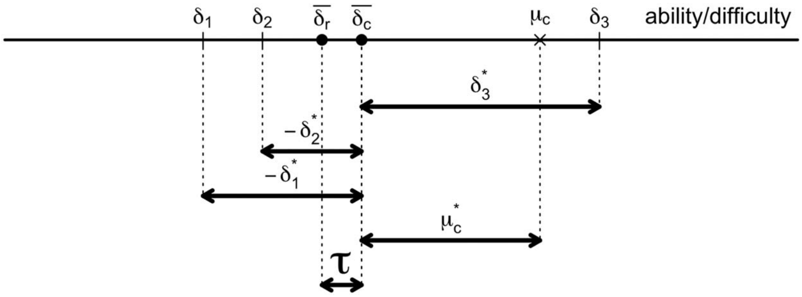


Figure 2. Reparametrization of the model parameters.

current tests, respectively. It can be seen that Equation 6 which defines τ_e does not include the level of the proficiency θ^* which means that the expert's estimate τ_e does not have a systematic upward bias if the expert overestimates the level of the borderline candidate or a downward bias if θ^* is underestimated. However, when giving instructions to experts it is important how θ^* is defined, as it may be easier for experts to make predictions about the performance of a borderline candidate when the level of proficiency is moderate (e.g., producing probabilities of a correct response between .2 and .8) than for more extreme definitions of θ^*.

Letting each expert evaluate all items is very time consuming and can lead to experts' judgments being less valid and reliable due to fatigue and loss of motivation. In our adapted procedure only a subset of items $\{r^*\}$ from the reference test and a subset of items $\{c^*\}$ from the current test are used. Then, for each expert we compute the difference between the average difficulties of the items in the subsets $\{c^*\}$ and $\{r^*\}$, denoted by τ_e^*:

$$\tau_e^* = \frac{\sum_{i\in\{c^*\}}\ln\left(\frac{1-p_{ie}}{p_{ie}}\right)}{|c^*|} - \frac{\sum_{j\in\{r^*\}}\ln\left(\frac{1-p_{je}}{p_{je}}\right)}{|r^*|}. \tag{7}$$

τ_e^* is not equal to τ_e since the items in the subsets are not fully representative of the full tests:

$$\tau_e = (\hat{d}_r - \hat{d}_c) - \tau_e^*, \tag{8}$$

where $\hat{d}_r$ is the difference between the average difficulty in the subset $\{r^*\}$ and the average difficulty in the set $\{r\}$, and $\hat{d}_c$ is the difference between the average difficulty in the subset $\{c^*\}$ and the average difficulty in the set $\{c\}$. These two quantities can be estimated from the training data.

The prior distribution of τ is chosen to be a normal distribution

$$p_1(\tau) = \mathcal{N}\big((\hat{d}_r - \hat{d}_c) - \mu_w, \sigma_w^2\big) \tag{9}$$

where $\mu_w = \frac{\Sigma_e w_e \tau_e^*}{\Sigma_e w_e}$ is the weighted mean of τ_e^* across the experts and $\sigma_w^2 = \frac{\Sigma_e w_e(\tau_e^* - \mu_w)^2}{1 - \Sigma_e w_e^2}$ is the weighted variance. The weights are determined by how well the estimated p_{ie} from each expert correlate with the observed proportions of correct responses to the items within each test in the training data. Because the probabilities are bounded between 0 and 1, before computing the correlation the estimated and the observed probability are logit-transformed, that is, $l_{ie} = \ln\left(\frac{p_{ie}}{1-p_{ie}}\right)$ and $l_i = \ln\left(\frac{p_i}{1-p_i}\right)$ and the weights are computed as follows:

$$w_e = \frac{1}{2}(Cor(\mathbf{l}_{re}, \mathbf{l}_r) + Cor(\mathbf{l}_{ce}, \mathbf{l}_c))\mathcal{I}_{Cor(\mathbf{l}_{re},\mathbf{l}_r)>0}\mathcal{I}_{Cor(\mathbf{l}_{ce},\mathbf{l}_c)>0}, \tag{10}$$

where $\mathbf{l}_{re}$ and $\mathbf{l}_{ce}$ are the vectors of logits of probabilities of length $|r^*|$ and $|c^*|$, respectively, evaluated by expert e, and $\mathbf{l}_c$ and $\mathbf{l}_r$ are the observed logit-transformed proportions of correct responses to the items in the training data. If one of these correlations is negative for a particular expert then this expert gets a weight of zero. The weights defined in Equation 10 are in a way themselves a sort of prior information, regarding the experts, obtained based on the analyses of training data.

Rulers Method for Direct Elicitation of the Prior Knowledge About τ

One could imagine two rulers on which the positions of the items within each test are indicated (see Figure 3). As the arrows

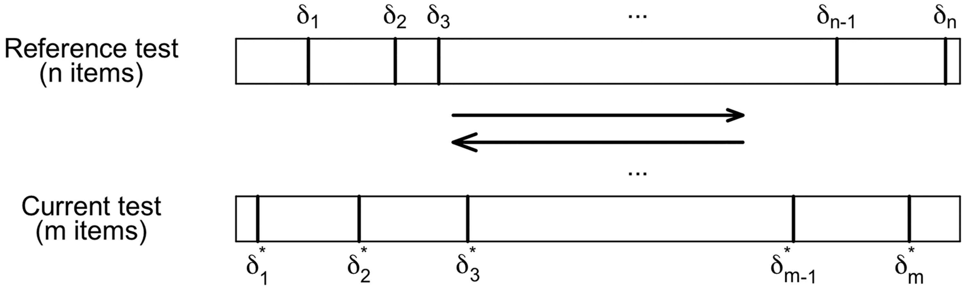


Figure 3. Two rulers with item difficulties estimated within each test: The arrows indicate that relative position of the rulers needs to be determined by experts.

show, the rulers can be arbitrarily shifted to the left or to the right relative to each other, because the examination data do not tell us anything about the relative position of these two rulers. Prior knowledge about how these rulers should be positioned relative to each other can be elicited from experts. The most direct way would be to give an expert the rulers with the empirical item positions within each test and ask her/him to determine the proper mutual position of the two rulers. But comparing two complete tests with a large number of items is a very complicated task which is practically impossible to complete. For that reason, we developed a procedure in which experts are asked to compare smaller sets of items which have to be carefully selected. Our method is different from just asking the experts to specify the mutual order of the items in the two tests, because the empirical within-test items orders and the distances between the item difficulties within each test make many of the orders impossible, for example in Figure 3 the order $\delta_1 < \delta_1^* < \delta_2^* < \delta_3^* < \delta_2$ is not possible since the distance between δ_3^* and δ_1^* is larger than the distance between δ_2 and δ_1.

When developing the elicitation procedure, we carried out a pilot study with one expert to figure out what problems experts might experience when comparing sets of items. First, we observed that it was much easier for the expert to compare items of similar content. Another observation was that sometimes the expert did not agree with the empirical order of the item difficulties within a test which made specifying the mutual order of the items in the two sets senseless. Based on these observations, we developed the elicitation procedure consisting of several steps. In the first step, items are selected from the reference and the current test so that they are similar to each other in content and differ from each other in difficulty. In the second step, each expert $e \in \{1 : E\}$ orders the item difficulties within each test and only those items for which the expert's order and the order observed in the training data match are retained. In the third step, each expert aligns the remaining items from the reference test to the items from the current test. In this way, we make sure that the problems observed in the pilot study will not occur. Below we describe the three steps of the procedure in detail.

Preliminary item selection.

1. Divide the items within each test into homogenous groups based on the content, for example in a language test items may be divided in spelling, grammar, and punctuation subgroups. Often tests consist of clearly defined subdomains. If that is not the case, then experts can be asked to help divide the items in homogeneous subgroups. The subgroups should not be made too small, six to eight items per subgroup should be sufficient.

2. Estimate the item difficulties with a Rasch model separately for the items in the reference test and in the current test given the training data.

3. Select the largest subset of items from each homogenous group, such that the posterior probability of each pair of items within the subset to have a certain order of item difficulties is larger than 95%.[1] If multiple subsets can be constructed, then select one of them at random. The elicitation procedure cannot be used if these subsets cannot be constructed, that is if either there is not enough variation in item difficulty or if there are not enough data to be certain about the order of the item difficulties.

Final item selection (performed separately for each expert *e*).

1. An expert orders the items within each homogeneous group based on their difficulty, for the two tests separately.

2. A subset of items from a set is retained if the expert's order of this set does not contradict the order observed in the training data. For example, if the observed order is:
$$\hat{\delta}_1 < \hat{\delta}_2 < \hat{\delta}_3 < \hat{\delta}_4, \tag{11}$$
and the expert's order is:
$$\delta_{1e} < \delta_{3e} < \delta_{2e} < \delta_{4e}, \tag{12}$$
then the expert's order within the subsets {1, 2, 4} and {1, 3, 4} do not contradict the empirical order. Both subsets can be used in the procedure. To make the selection of items automatic, one of these subsets is chosen randomly.

3. If for one of the content groups of items, in one of the tests there is no pair of items to be retained, then this group is discarded from both tests. Therefore, after the final selection of items different experts might have different number of item groups to compare.

4. The quality of the judgments of expert *e* is quantified by the average proportion of items for which the expert's order and the empirical order were the same, denoted by p_e. In the case of empirical and expert orders in (Equation 11) and (Equation 12) this proportion is equal to.75. p_e used to weigh the expert's judgments, such that the effect of the prior distribution elicited from an expert on the combined prior decreases if her/his judgments rarely match the observed data. The weight of expert *e* is equal to
$$w_e = \frac{p_e - p_0}{\sum_e (p_e - p_0)}, \tag{13}$$
where p_0 is the expected average proportion of items which would be obtained if the order produced by a hypothetical expert is random.

Shifting two rulers. This stage is done for all content groups $j \in [1 : J_e]$, where J_e is the number of the content groups retained for expert *e* and consequently the number of judgments about the parameter τ which are elicited from her/him. Starting from the third step of the procedure the rulers method is illustrated in Figure 4.

1. The expert is presented with a set of items from the reference test from one of the content groups ordered

[1] This posterior probability can be approximated by the proportion of samples from the posterior distribution (given in Equations 23 and 24 in Appendix A) in which a certain order holds for the sampled values of the item difficulties.

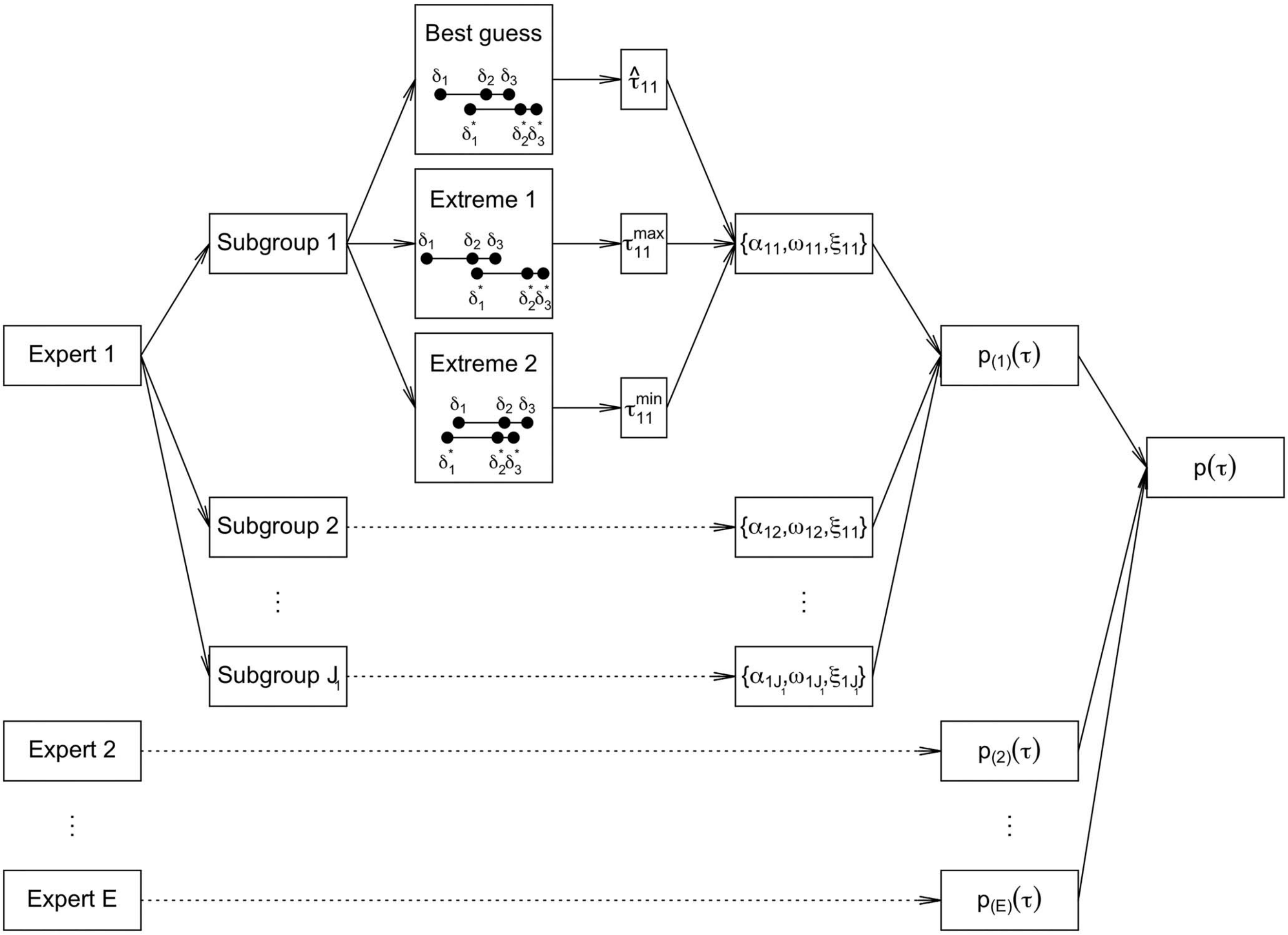

Best guess
δ_1 δ_2 δ_3
δ_1^* $\delta_2^*\delta_3^*$
Extreme 1
δ_1 δ_2 δ_3
δ_1^* $\delta_2^*\delta_3^*$
Extreme 2
δ_1 δ_2 δ_3
δ_1^* $\delta_2^*\delta_3^*$
Expert 1
Subgroup 1
Subgroup 2
⋮
Subgroup J_1
$\hat{\tau}_{11}$
τ_{11}^{max}
τ_{11}^{min}
$\{\alpha_{11},\omega_{11},\xi_{11}\}$
$\{\alpha_{12},\omega_{12},\xi_{11}\}$
⋮
$\{\alpha_{1J_1},\omega_{1J_1},\xi_{1J_1}\}$
$p_{(1)}(\tau)$
$p(\tau)$
Expert 2
$p_{(2)}(\tau)$
⋮
⋮
Expert E
$p_{(E)}(\tau)$

Figure 4. Scheme of the rulers method after the final selection of items.

from the easiest to the most difficult one, and with an ordered set of the items from the current test (from the same content group). The items are positioned according to the distances between the item difficulties estimated from the training data. The expert can shift these two sets of items relative to each other to the right or to the left, but cannot change the positions of the items within a test, since they are fixed at the estimated difficulties.

2. The expert places the two sets together on the common scale according to her/his best guess. In each of the content groups we obtain the mutual order of the items and the estimate of the mode of the prior distribution, denoted by $\hat{\tau}_{ej}$ (see "Best guess" in Figure 4).

3. To evaluate the expert's uncertainty about the estimate, she/he is asked to place the item sets in the two most extreme positions which the expert still considers plausible: first in which the set of items from the current test has the rightmost position on the scale (resulting in an upper bound $\tau_{ej}^{\max}$) and second in which it has the leftmost position in the scale (resulting in a lower bound $\tau_{ej}^{\min}$; see "Extreme 1" and "Extreme 2" in Figure 4).

The above described procedure of experts' judgments collection for each expert e results in multiple sets $\{\hat{\tau}_{ej}, \tau_{ej}^{\min}, \tau_{ej}^{\max}\}, \forall j \in [1 : J_e]$. Next, we describe how this information can be translated into the prior distribution of τ.

Each pair of sets of items to be ordered presents the expert with an opportunity to express her/his beliefs about the relative difficulties of the items in the two sets and her/his uncertainty about the assessment of these relative difficulties. Through our method, each pair of sets produces an estimate of the mode of the expert's prior, as well as an estimate of the lower and the upper bound of the region that the expert still considers to be credible. This credible range was operationalized as the 90% credible interval for that expert's prior, and hence the lower and the upper bound that were specified by the expert are used as an estimate of her/his 5% and 95% percentiles, respectively. Because the extreme positions do not have to be symmetric around the mode, to approximate the prior knowledge elicited from expert e in judgment j we need a distribution which allows for skewness. A skew-normal distribution (Azzalini, 2005) is used:

$$p_{ej}(\tau) = \text{Skew-normal}\ (\xi_{ej}, \omega_{ej}, \alpha_{ej}), \qquad (14)$$

where ξ_{ej} specifies the location, ω_{ej} specifies the spread and α_{ej}

specifies the degree of skewness of the distribution. The skew-normal distribution includes a normal distribution as its special case when $\alpha_{ej} = 0$. The parameters of the distribution are chosen in such a way that $\tau_{ej}^{\min}$ and $\tau_{ej}^{\max}$ are the fifth and the 95th quantiles and $\hat{\tau}_{ej}$ is the mode (see Appendix B for the details).

Each judgment of expert e adds extra information about which values of τ are plausible according to this expert. Separate judgments of the same expert are assumed to be independent; therefore, to combine the information from several judgments, we use a product of the distributions $p_{ej}(\tau)$:

$$p_e(\tau) = \frac{\prod_j p_{ej}(\tau)}{\int_{-\infty}^{\infty} \prod_j p_{ej}(\tau) d\tau}, \tag{15}$$

where the normalizing constant in the denominator ensures that $p_e(\tau)$ is a proper distribution. This integral does not have a closed form solution and is therefore approximated here by numerical integration with Gauss-Hermite quadrature (see Equation 58 in Appendix B). The motivation for the independence assumption is that each judgment refers to a different set of items with a unique combination of item features influencing the item difficulty, which an expert takes into account.

When combining the information from the different experts, we use linear opinion pooling (O'Hagan et al., 2006; Stone, 1961):

$$p_2(\tau) = \sum_e w_e p_{(e)}(\tau), \tag{16}$$

where the weights w_e computed using (Equation 13) make sure that the results of the experts with higher quality judgments have a larger influence on the prior $p_2(\tau)$. We prefer linear opinion pooling over logarithmic opinion pool because the latter being a geometric mean of individual distributions leads to unrealistically strong aggregated believes. Moreover, linear opinion pool does not rule out the low or the high values of the parameter that are supported by a minority of the experts (O'Hagan et al., 2006, p. 184).

Differences Between the Two Elicitation Methods

From a practical perspective there are some important differences between the two proposed methods. While the Angoff method is simple to implement—the items need to be presented to the experts, either in a paper-and-pencil or in a computerized format, the rulers method requires more elaborate preparations. It has to be implemented in a computerized procedure which may need some adaptation to a specific set of items. The second difference between the methods is that the experts' judgments can be collected using Angoff method prior to exam administration, but for collecting the judgments using the rulers method the examination data have to be available for the specification of item locations in the computerized procedure. However, the procedure can be prepared beforehand such that as soon as the examination data come in, one is able to collect the experts' judgments.

The methods also differ from the cognitive perspective. First, in the Angoff method experts provide absolute judgments about each of the items, while in the rulers method experts provide comparative judgments, ordering sets of items. The latter might provide better results since people are often more reliable when comparing objects than when giving absolute judgments (Laming, 2004). Second, in the Angoff method the items are presented sequentially, while in the rulers methods sets of items are presented simultaneously, which puts higher cognitive demand on the experts. The number of items presented simultaneously should not exceed the capacity of working memory (7 ± 2; Miller, 1956). Third, unlike in the Angoff method, in the rulers method partial feedback is provided to the experts with respect to the item order within each test. When combining the item orders from two tests they can use the information in the order of the items within each test as a reference for aligning the two sets. This kind of external information is not available in the Angoff method.

Empirical Example

Data

For illustrating and comparing the methods of test linking using prior knowledge, we used the data from the test of mathematics for primary school "Entreetoets Group 7" taken by students in the Netherlands at the end of the 5th grade. The same test consisting of 120 items was administered in 2008 and 2009. The test was divided into 10 groups based on content: (a) mental arithmetics; (b) mental arithmetics—estimation; (c) arithmetic operations; (d) number relations; (e) geometry; (f) measurement of length and surface; (g) measurement of weight and volume; (h) percentages, fractions and ratios; (i) time; (j) money and then each subgroup was randomly divided into two parts. We treated the first part as the reference test and the second part as the current test. The populations of 2008 and 2009 were treated as the reference and the current population, respectively. Hence, an equating problem was artificially created for the data set in which the responses of the persons from what we labeled as "reference population" to the items from what we labeled "current test" were actually observed (see Figure 5). This makes it possible not only to illustrate the procedures introduced in Section Elicitation of prior knowledge about the difference between the difficulty of two tests but also to evaluate them by comparing the estimate of the new cut-score obtained with the different priors based on the *predicted* responses of the reference population to the current test (see Figure 5b denoted by "?") with the cut-score, derived from the *observed* responses of the reference population to the current test in the complete data (see Figure 5a). Hence, the latter is used as a proxy of the true cut-score.

The data set of each year consisted of responses of more than 100,000 students. Because the linking procedures developed in this study are meant for tests administered to smaller samples, responses of 2,000 persons from each year were randomly selected as examination data. The data for three linking groups with responses to randomly selected eight items from the reference test and eight items from the current test were selected from the data of 2008. Although they were sampled from the same population as the examination data, this fact was ignored in the estimation, assigning separate parameters for the mean and the variance of proficiency in each linking group.

Because based on this particular arithmetics test students are assigned to one of five levels of proficiency (from A to E), four cut-scores need to be estimated. The cut-scores between the levels in the reference test (denoted by s_{ref}) were 49, 42, 35, and 24 correct responses. The corresponding cut-scores for the current test

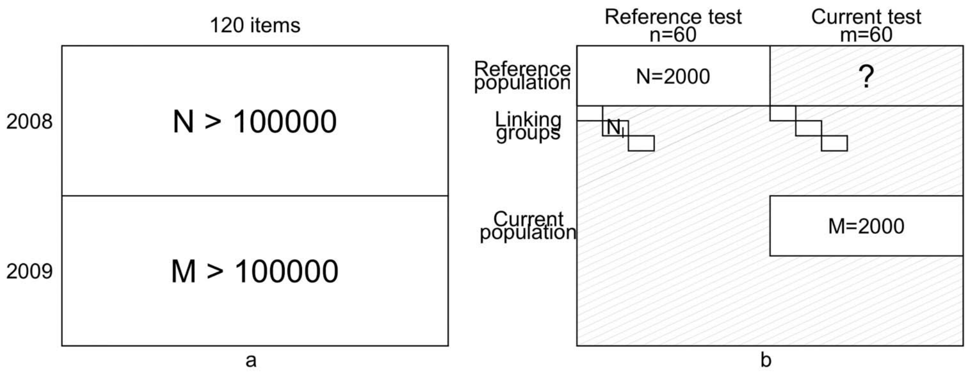


Figure 5. Creating an equating problem from a complete design. The completely dashed areas indicate data that were not used in the analysis.

maintaining the same standard need to be estimated. A true cut-score for the current test (denoted by $s_{pass,true}$) was determined from the complete data (see Figure 5a) as such a score that the proportion of persons from the reference population with scores on the current test below this score is as close as possible to the proportion of persons from the reference population with scores on the reference test below the corresponding s_{ref}:

$$s_{pass,true} = \text{argmin}_s\left(\left(\sum_p\left(\mathcal{I}\left(\sum_{i\in\{r\}} x_{pi} < s_{ref}\right) - \mathcal{I}\left(\sum_{i\in\{c\}} x^*_{pi} < s\right)\right)\right)^2\right). \tag{17}$$

In this way the true cut-scores for the current test were determined to be equal to 50, 43, 37, and 26 correct responses.

Method

To evaluate the linking results based on different priors we analyzed how the estimated cut-scores for the current test (s_{pass}) changed depending on the prior specification.

For each cut-score in the reference test we estimated a cut-score in the current test using different priors and compared it to the true cut-score $s_{pass,true}$. To take into account the sampling variability, we resampled the data (both the persons and the items) for the linking groups 100 times to examine the distribution of the estimated cut-scores across these resampled data sets, to evaluate how often each of the cut-scores is correctly estimated:

$$k_{true} = \sum_{k=1}^{100} \mathcal{I}(s_{pass,k} = s_{pass,true}) \tag{18}$$

and how large the mean squared error (*MSE*) is:

$$MSE = \frac{1}{100}\sum_{k=1}^{100}(s_{pass,k} - s_{pass,true})^2, \tag{19}$$

where $s_{pass,k}$ is the estimated cut-score from the k-th resampled data set.

We also used the average number of misclassified students across the resampled data sets to compare the quality of linking based on different priors. In each resampled data set the number of persons from the current population who were assigned to an incorrect level of proficiency due over- or underestimation of the cut-scores was counted. If the estimates of the cut-scores are centered around the true values and do not vary a lot then this number would be small.

To analyze how the influence of the prior distribution changes depending on the size of the linking data, we varied the number of persons per linking group: N_l = 100, 200, 300, and 500.

In addition to $p_1(\tau)$ defined in Equation 9 and $p_2(\tau)$ defined in Equation 16, we also used a vague prior:

$$p_0(\tau) = \mathcal{N}(0, 100) \tag{20}$$

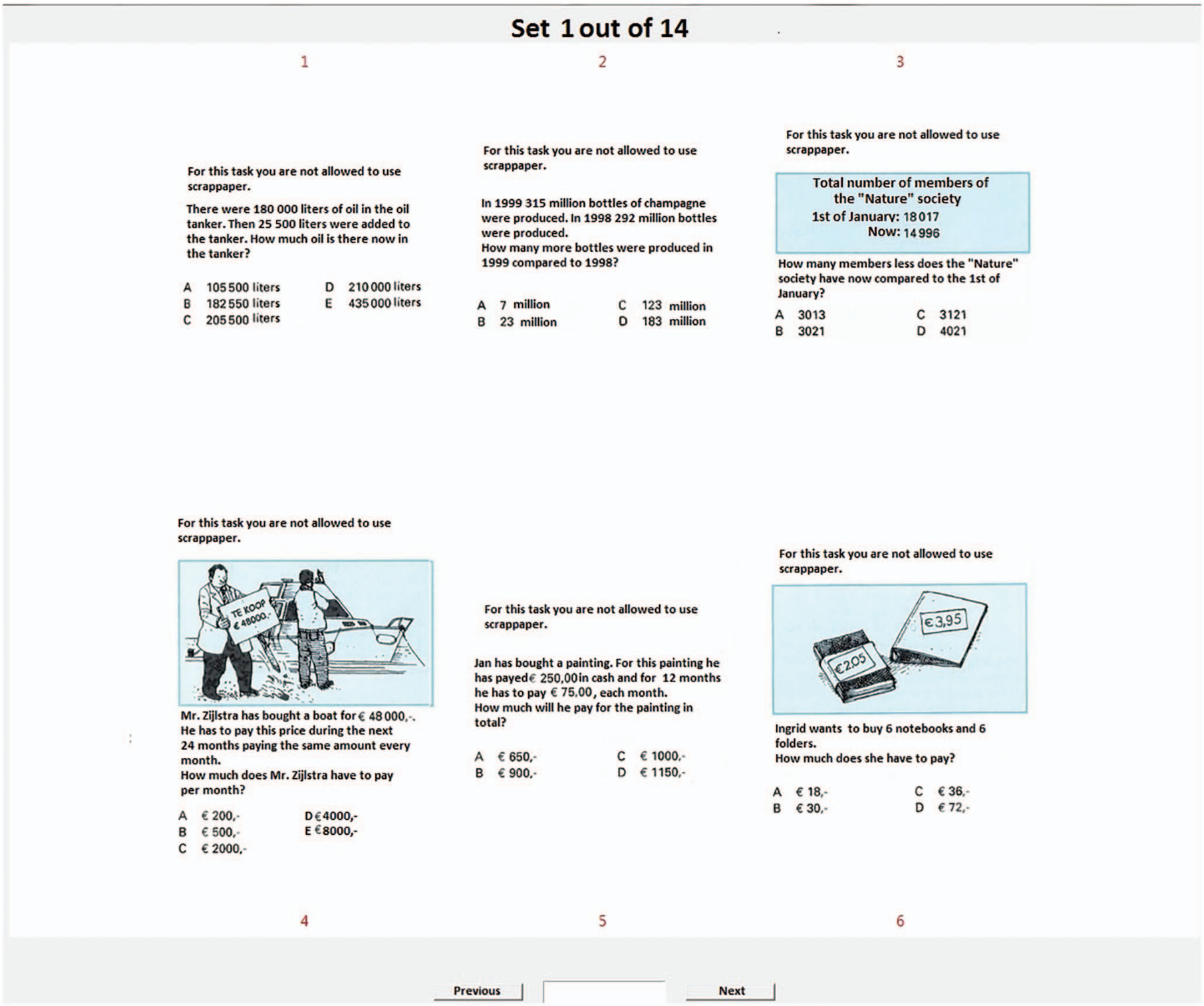
to show the added value of using prior knowledge.

Pilot Study

Prior to the main elicitation study a pilot study was conducted with a group of nine experts who were members of the construction groups, who develop items for mathematical tests at the Dutch National Institute for educational measurement. A computerized procedure implementing the rulers method and a paper-and-pencil Angoff procedure were tried out during a group meeting in which each expert evaluated the items using both methods.

Based on the results of the pilot study several decisions were made. Minor adjustments were implemented in the computerized procedure to make sure that the instructions were properly followed. Furthermore, we decided that it would be better to organize individual face-to-face sessions with experts instead of group sessions because in this way any questions that the participants might have can be answered immediately and instructions can be clarified if needed (see O'Hagan et al., 2006).

The pilot study demonstrated that it was rather difficult for the participants to evaluate the item difficulty and order the items along a single dimension of mathematics ability. Because most of the experts who participated in the elicitation study were teachers, they had a lot of practical experience with primary school mathematics but lacked more theoretical knowledge in item analysis. Having observed this, we were confronted with two options: develop a specialized training session to familiarize the experts with basic psychometric concepts and to teach them use these abstracts concepts in item evaluation, or to search for a different group of primary school mathematics experts who already have this knowledge. As O'Hagan et al. (2006) note, it is important to evaluate what kind of experience and expertise potential reviewers have and decide on training based on that. We decided



Set 1 out of 14

1

For this task you are not allowed to use scrappaper.

There were 180 000 liters of oil in the oil tanker. Then 25 500 liters were added to the tanker. How much oil is there now in the tanker?

A 105 500 liters
B 182 550 liters
C 205 500 liters
D 210 000 liters
E 435 000 liters

2

For this task you are not allowed to use scrappaper.

In 1999 315 million bottles of champagne were produced. In 1998 292 million bottles were produced.
How many more bottles were produced in 1999 compared to 1998?

A 7 million
B 23 million
C 123 million
D 183 million

3

For this task you are not allowed to use scrappaper.

Total number of members of the "Nature" society
1st of January: 18 017
Now: 14 996

How many members less does the "Nature" society have now compared to the 1st of January?

A 3013
B 3021
C 3121
D 4021

4

For this task you are not allowed to use scrappaper.

Mr. Zijlstra has bought a boat for € 48 000,-.
He has to pay this price during the next 24 months paying the same amount every month.
How much does Mr. Zijlstra have to pay per month?

A € 200,-
B € 500,-
C € 2000,-
D € 4000,-
E € 8000,-

5

For this task you are not allowed to use scrappaper.

Jan has bought a painting. For this painting he has payed € 250,00 in cash and for 12 months he has to pay € 75,00, each month.
How much will he pay for the painting in total?

A € 650,-
B € 900,-
C € 1000,-
D € 1150,-

6

For this task you are not allowed to use scrappaper.

Ingrid wants to buy 6 notebooks and 6 folders.
How much does she have to pay?

A € 18,-
B € 30,-
C € 36,-
D € 72,-

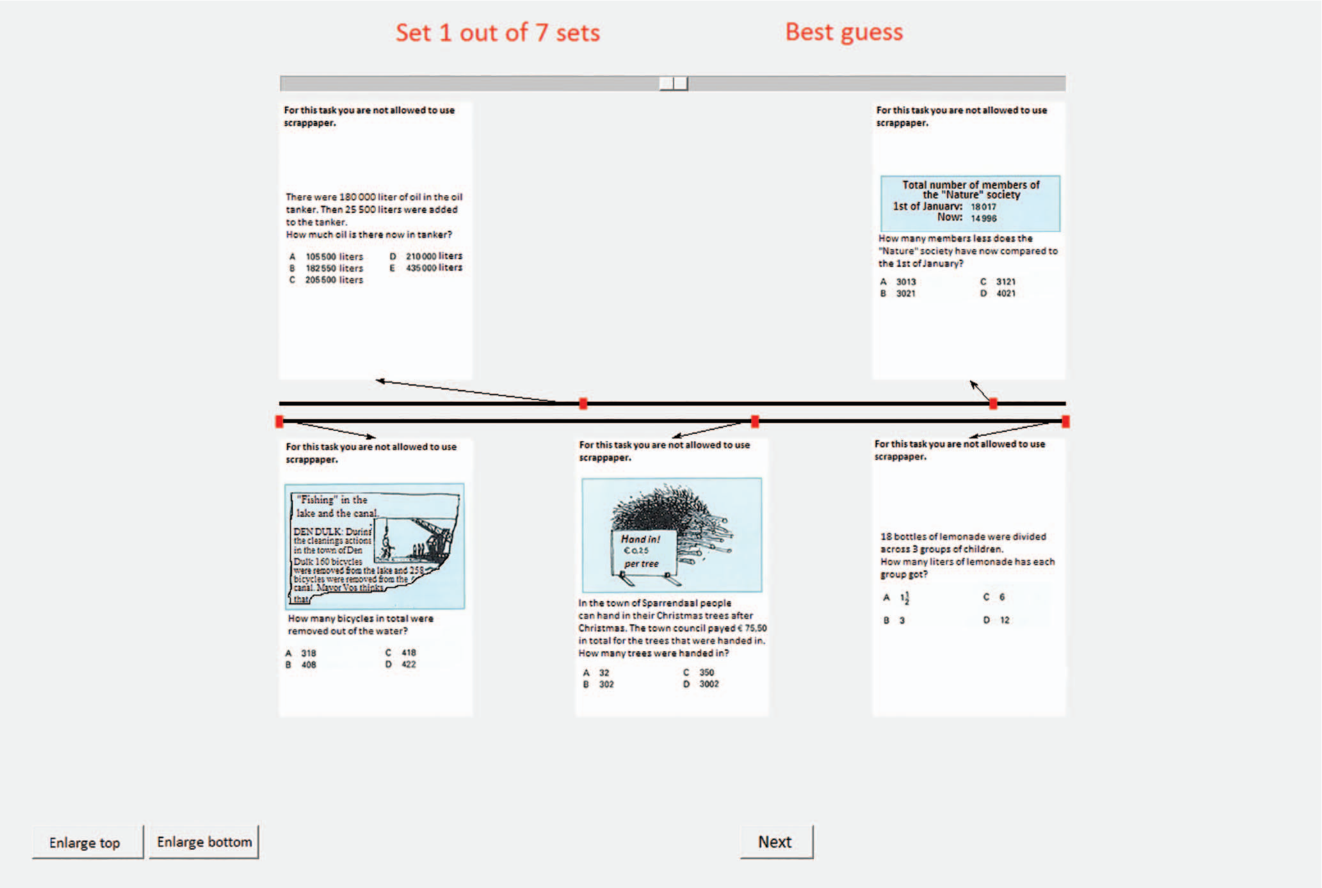
Figure 6. Illustration of the first part of the computerized procedure for the rulers method: Six items have to be ordered based on their difficulty (translated from Dutch). See the online article for the color version of this figure.

to choose the second option because it was easier to find educational researchers and test experts specialized in mathematics than develop and implement an extensive training program. Our choice was also supported by the fact that the results of one of the participants who was not only an item writer but also a primary school mathematics researcher with considerable psychometrics training were closer to the true value of τ than those of the other participants. The results of the pilot study were generally promising, especially for the rulers method, and therefore we decided that it was worthwhile continuing with a new group of carefully selected persons.

Expert Elicitation

Participants. Seven experts (four females and three males) participated in the elicitation study. These were four primary schools mathematics researchers from two Dutch universities and three employees of the Dutch National Institute for educational measurement working with primary school mathematics tests. These experts had previous experience in standard setting procedures and in psychometric item analysis. Most of them also had worked as item writers and had received feedback on item performance (on item difficulties among other properties). Moreover, they had theoretical knowledge about primary school mathematics and understanding of fundamental psychometric concepts. For these reasons, we decided not to include a training session as part of the procedure and rely on their extensive previous training.

Preliminary selection of items. From the 10 content groups only seven groups were selected for the expert elicitation, because the other three groups were too small (number relations, geometry, and money). The items within each test were ordered based on the observed proportions of correct responses in the training data.



For this task you are not allowed to use scrappaper.

There were 180 000 liter of oil in the oil tanker. Then 25 500 liters were added to the tanker.
How much oil is there now in tanker?

A	105 500 liters	D	210 000 liters
B	182 550 liters	E	435 000 liters
C	205 500 liters		

For this task you are not allowed to use scrappaper.

Total number of members of the "Nature" society
1st of Januarv: 18 017
Now: 14 996

How many members less does the "Nature" society have now compared to the 1st of January?

A	3013	C	3121
B	3021	D	4021

For this task you are not allowed to use scrappaper.

"Fishing" in the lake and the canal.
DEN DULK: Durinf the cleanings actions in the town of Den Dulk 160 bicycles were removed from the lake and 258 bicycles were removed from the canal. Mayor Vos thinks that

How many bicycles in total were removed out of the water?

A	318	C	418
B	408	D	422

For this task you are not allowed to use scrappaper.

In the town of Sparrendaal people can hand in their Christmas trees after Christmas. The town council payed € 75,50 in total for the trees that were handed in.
How many trees were handed in?

A	32	C	350
B	302	D	3002

For this task you are not allowed to use scrappaper.

18 bottles of lemonade were divided across 3 groups of children.
How many liters of lemonade has each group got?

A	$1\frac{1}{2}$	C	6
B	3	D	12

Figure 7. Illustration of the second part of the computerized procedure for the rulers method: best guess (translated from Dutch). The red blocks indicate the item positions on the difficulty scale. The slider at the top of the screen can be used to change the position of the item sets. See the online article for the color version of this figure.

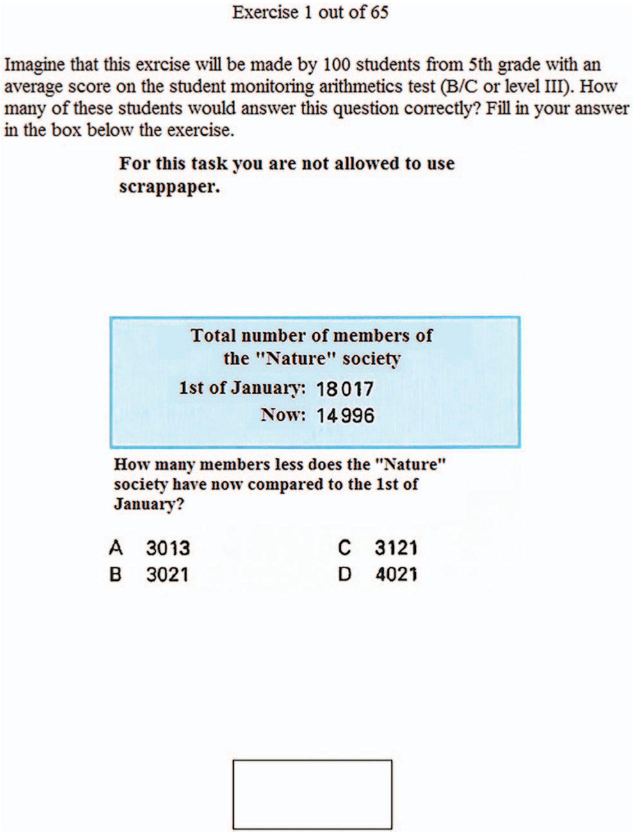
Within each of the seven content groups a subset was selected such that for each pair of items *i* and *j* within this subset the posterior probability of them having a certain order of item difficulties was larger than .95.[2] In total 62 items were selected which were used both for the Angoff and for the rulers methods. The sets of selected items is rather representative of the full sets of items in the reference and the current test in terms of difficulty: $\hat{d}_r = .10$ and $\hat{d}_c = .01$ (for comparison, the standard deviation of the estimates of the item difficulties in the reference and the current test were equal to .60 and .66, respectively).

The item difficulties within each test, which were used for specifying the locations of the items in the elicitation procedure, were estimated from the training data (see μ_{δ_i} and $\mu_{\delta_i^*}$ in Appendix A).

Procedure. The two methods were administered to each expert individually in two separate sessions. The period between the sessions was long enough for there to be no interference between the two elicitations. The sessions took about 30–45 min, roughly the same for the two methods. The rulers method was implemented in a computer application developed for this study. The application had two parts corresponding to the second and the third steps of the elicitation procedure described in Section Rulers method for direct elicitation of the prior knowledge about τ. Each expert got a prerecorded audio-instruction accompanied by a power-point presentation illustrating the procedures. In the first part of the procedure, experts were presented with sets of three, four, five, or six items from each content group and each test (see Figure 6). The content groups were presented in the same order to everyone, but the order of the item sets within a content group was randomly chosen for each expert (either reference test–current test or current test–reference test). Within each set items were presented in a random order. For each set experts had to fill in the order of the items based on their difficulty starting from the easiest item.

After the first part, experts received the instruction for the second part. In the second part of the procedure experts were presented with two sets of items: one at the top and one at the bottom. The items were located according to their estimated item

[2] This posterior probability is approximated by the number of samples from the posterior distribution defined in Equations 23 for the reference test and Equation 24 for the current test, respectively, in which the sampled values have that order, the specification of the prior distribution and the sampling scheme can be found in Appendix A.



Exercise 1 out of 65

Imagine that this exrcise will be made by 100 students from 5th grade with an average score on the student monitoring arithmetics test (B/C or level III). How many of these students would answer this question correctly? Fill in your answer in the box below the exercise.

For this task you are not allowed to use scrappaper.

Total number of members of the "Nature" society
1st of January: 18017
Now: 14996

How many members less does the "Nature" society have now compared to the 1st of January?

A 3013
B 3021
C 3121
D 4021

Figure 8. Illustration of the Angoff procedure: One item per page is presented (translated from Dutch). See the online article for the color version of this figure.

difficulties based on the training data. Each set contained at most four items. If after the first part there were more than four items in a set for which the expert’s order matched the empirical order, then four items were randomly selected to be retained. It was not possible for the experts to move the items within a set, but only sets as a whole relative to each other using the slider at the top. First, the experts had to place the sets in the most plausible position (see Figure 7). Second, they had to specify the first extreme position by moving the top set to the right away from the most likely position (in Figure 7 “Best guess” was substituted by “Extreme 1: Top to the right”) to the most extreme position which is still plausible. Third, they had to choose the second extreme position by moving the sets away from the most likely position in the opposite direction (in Figure 7 “Best guess” was substituted by “Extreme 2: Top to the left”). Although experts might vary in the interpretation of what “plausible” means, we argue that specifying the range of uncertainty for the experts numerically (i.e., “90% chance that the order is between the extreme positions”) would not solve this problem because interpretations of “90% chance” might also differ between persons, because probabilistic statements of this kind are difficult to make in general (O’Hagan et al., 2006). For this reason in this study we decided not to quantify the uncertainty in the instructions for the experts, but rather use a verbal description and only use a quantification after the expert judgments are collected.

In the Angoff procedure the same 62 items as used in the rulers method were given to each expert in a random order. The experts filled in their responses to a question “Imagine a group of students from the fifth grade with an average level of proficiency (B/C or Level 3 of the student monitoring system). How many of these students will answer this item correctly?” in a booklet with one item per page (see Figure 8). Note, that instead of asking the experts about model parameters (item difficulty or probability of a correct response) which is often difficult for experts to provide direct information on (Christensen et al., 2010), we translated these questions into questions that are more familiar to experts and, therefore, easier to work with. Three extra items were included in the beginning of the booklet to familiarize the experts with the procedure, such that the total number of items was 65 but only the results of the 62 items were taken into account.

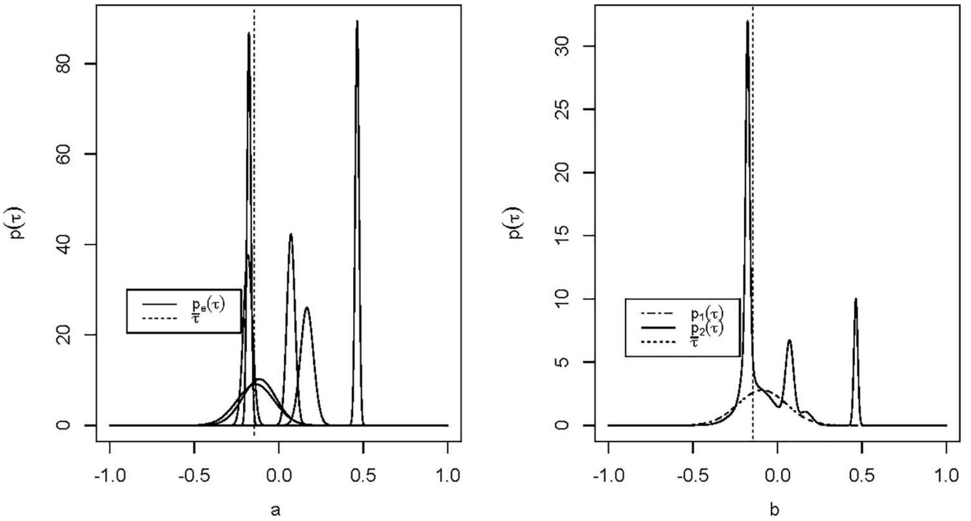


Figure 9. Prior distributions elicited from individual experts using the rulers method (a) and combined prior distributions: $p_1(\tau)$ - Angoff prior, $p_2(\tau)$ - rulers prior (b); $\bar{\tau}$ - proxy of the true value of τ.

Results

The Angoff method resulted in a prior

$$p_1(\tau) = \mathcal{N}(\mu = -0.09, \sigma^2 = 0.02), \tag{21}$$

which is shown in Figure 9b. Figure 9a shows the priors elicited from individual experts in the computerized procedure using the rulers method. In Figure 9b the combined prior $p_2(\tau)$ and the prior elicited with the Angoff method $p_1(\tau)$ are shown. Figure 9 also includes $\bar{\tau} = -0.145$ which was estimated from the observed responses of almost 20,000 persons from both 2008 and 2009 to all 120 mathematics items (see Figure 5a). A Rasch model was fitted to the complete data set and the difference between the average estimates of the difficulties of the items in the current and the reference test was computed. We used it as a proxy for the true value of τ to evaluate how close to the truth the expert's judgments were. The mean of the prior distribution $p_1(\tau)$ is slightly above the true value of τ and this prior assigns relatively high density to $\bar{\tau}$. The largest mode of the distribution $p_2(\tau)$ is very close to $\bar{\tau}$. Four of the distributions $p_e(\tau)$ are concentrated around the proxy of the true value $\bar{\tau}$. One of the experts provided a distribution with a mode far away from the judgments of other experts. This was the expert with the lowest quality of the judgment and the smallest weight $w_e = .04$ (for comparison, $w_e = .14$ would be the weight of each expert if equal weights were assigned).

Figure 10 shows the distribution of the estimates of the cut-scores across the resampled data sets, and Table 1 shows how often each of the four cut-scores was correctly estimated with different priors. From Figure 10 one may see that with the vague prior there was a lot of variation in the estimated cut-scores, especially when $N_l = 100$. With the informative priors, the estimates of the cut-scores across the resampled data sets were less spread around the mode especially with the rulers prior. For all sample sizes and all cut-scores the rulers prior $p_2(\tau)$ resulted in correctly estimated cut-scores in more resampled data sets than the vague prior (see Table 1). The variance of the estimated cut-scores was reduced, while they were still concentrated around the true value (see Figure

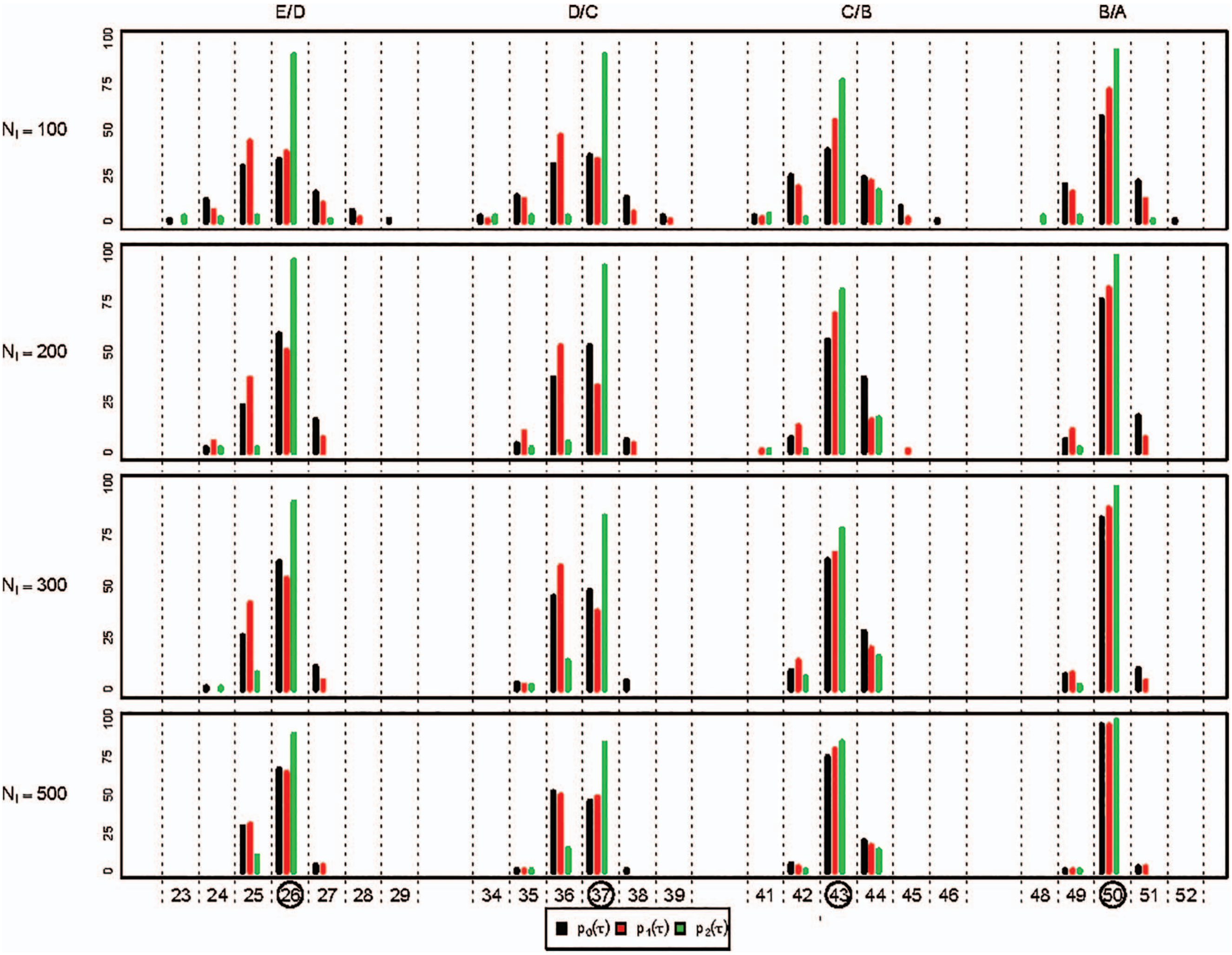


Figure 10. Distributions of the estimates of the cut-scores across the resampled data sets using different priors: $p_0(\tau)$ - vague prior, $p_1(\tau)$ - Angoff prior, $p_2(\tau)$ - rulers prior. The true cut-scores are marked with a circle. See the online article for the color version of this figure.

Table 1
Numbers of Data Sets in Which the Estimated Cut-Score was Equal to the True Cut-Score (k_{True}) and Mean Squared Error (MSE) of the Estimates of the Cut-Scores With the Vague Prior (p_0), Angoff Prior (p_1), and Rulers Prior (p_2) Given Different Number of Persons in the Linking Groups (N_l)

		N_l = 100		N_l = 200		N_l = 300		N_l = 500	
Cut-score	Prior	k_{true}	*MSE*	k_{true}	*MSE*	k_{true}	*MSE*	k_{true}	*MSE*
D/E	$p_0(\tau)$	34*	1.36	59*	.47	62*	.41	67*	.33
	$p_1(\tau)$	38	.86	51*	.64	54*	.46	65*	.35
	$p_2(\tau)$	91*	.39	96*	.10	91*	.12	90*	.10
C/D	$p_0(\tau)$	36*	1.39	53*	.59	48*	.61	46	.57
	$p_1(\tau)$	34	1.13	33	.97	38	.68	49	.54
	$p_2(\tau)$	91*	.42	93*	.13	84*	.22	84*	.19
B/C	$p_0(\tau)$	39*	1.02	56*	.44	63*	.37	75*	.25
	$p_1(\tau)$	55*	.57	69*	.37	66*	.34	80*	.20
	$p_2(\tau)$	77*	.35	81*	.22	78*	.22	85*	.15
A/B	$p_0(\tau)$	57*	.46	76*	.24	83*	.17	96*	.04
	$p_1(\tau)$	72*	.28	82*	.18	88*	.12	96*	.04
	$p_2(\tau)$	93*	.16	98*	.02	98*	.02	99*	.01

* The most frequent estimate of the cut-score is equal to the true cut-score.

10). For the rulers prior in all conditions and for all levels of proficiency, the most frequent estimate of the cut-score was equal to the true cut-score. The *MSE*s of the estimated cut-scores were the smallest for $p_2(\tau)$. The *MSE*s of the estimated cut-scores with the Angoff prior were in some conditions lower than those for the vague prior, but in some conditions higher, especially of the cut-score between categories C and D, because this cut-score is the most sensitive to the value of τ due to the fact that in the corresponding range of scores different scores are very close to each other in terms of proportions of persons below the score.

To illustrate the consequences of incorrectly estimating the cut-scores more concretely, we looked at the average number of persons misclassified when different priors were used (see Table 2). The use of the rulers method improves the proportion of misclassifications compared with the use of vague prior; however, the differences between the results decrease with the sample size N_l as expected. The use of the Angoff prior improves the results over the vague prior when the sample size is the smallest, but results in roughly the same proportion of misclassifications as with the vague prior when the number of persons in the linking groups is larger (N_l = 200, 300, and 500). As can be seen from Tables 1 and 2, among the two elicitation methods developed in this study the rulers method provided better results than the Angoff method.

What If the Experts Are Wrong?

The utility of including prior knowledge in test linking depends on the quality of the expert judgments. In our empirical elicitation study the pool of seven experts provided judgments relatively close to the true value of τ which to a large extent improved the linking results. However, the results of linking would be negatively affected if the expert judgments are far from the true value. When using informative priors based on subject-matter experts judgments it is important to compare the obtained results with the results based on the vague priors. The latter reflects the information contained in the data only. If the estimated cut-scores obtained with and without taking the expert judgments into account differ dramatically then one should decide whether to trust the linking data or the experts more. On the one hand, as Lunn, Jackson, Best, Thomas, and Spiegelhalter (2012) argue "there is no such thing as the true prior." On the other hand, in the context of test linking there is no such thing as perfect linking data. In the same way as the experts might be wrong, the data might introduce bias in the estimation of the difference between the test difficulties, since the test are usually administered to different populations of students (i.e., measurement invariance is not warranted) and in different conditions (e.g., respondents might not be motivated to do their best on the test especially on difficult and time intensive items). Using Bayesian methods for test linking allows to include all

Table 2
Average Number of Misclassified Persons (and Percentage from the Number of Persons in the Current Examination) Across 100 Resampled Data Sets Given Different Number of Person in the Linking Groups (N_l) With the Vague Prior (p_0), Angoff Prior (p_1), and Rulers Prior (p_2)

Prior	N_l = 100	N_l = 200	N_l = 300	N_l = 500
$p_0(\tau)$	147.21 (7.36%)	82.57 (4.13%)	73.54 (3.68%)	54.24 (2.71%)
$p_1(\tau)$	109.86 (5.49%)	86.37 (4.32%)	73.20 (3.66%)	49.77 (2.49%)
$p_2(\tau)$	37.10 (1.85%)	19.63 (.98%)	26.56 (1.33%)	21.34 (1.07%)

available information and provides possibilities to assign different weights to different sources of information depending on their credibility.

An important feature of our proposed elicitation methods is that they provide explicit methods for evaluating the quality of experts judgments, by comparing the experts judgments about the item difficulties within the reference and the current tests with the observed examination data (see expert weights defined by Equation 10 for the Angoff method and by Equation 13 for the rulers method). The procedure can be further extended with rules for including or excluding results of a certain expert to the prior distribution by setting thresholds on the minimal value for the correlation between the experts' probabilities of the correct response elicited in the Angoff method and the item difficulties estimated from the training data within each test (e.g., "the correlation has to be at least .2 for each test") and on the proportion of retained items in the rulers method (e.g., "at least 70% of the items have to be retained"). In this way not only the relative impact of one expert compared with another can be regulated, but also judgments of certain experts might be fully excluded if within the reference and the current test they do not perform better than if the judgments were random. However, although this could be done, we argue that it is not very likely that using a group of carefully selected experts would result in an "untrue" prior.

In this study the expert weights were assigned based on the match of expert judgments about item difficulties within the reference and the current test to the item difficulties estimated from the examination data. As an anonymous reviewer suggested, the choice of weights can be also (partly) based on the background information about the experts, for example by assigning higher weights to experts with more experience. Moreover, the expert weights for the current linking cycle may be determined by asking experts to compare items from tests that have gone through linking procedures in previous years and evaluating the quality of their judgments. These are all valuable ideas that should be considered when using the approach proposed in future applications.

Conclusions

In this article we introduced different procedures for elicitation of prior knowledge for test linking from subject-matter experts. The empirical elicitation study showed promising results of including expert knowledge in test linking. In our study the rulers method of elicitation based on first ordering items based on difficulties first within each test and then combining the item orders across the two tests performed better than the Angoff method based on absolute judgments about the proportions of students answering the items correctly. However, when deciding to prefer one method over the other one should also take the practical considerations into account because the Angoff method is easier to implement.

Although including expert knowledge in test linking was demonstrated to decrease the expected number of misclassified students, there is one limitation of the approach proposed. The approach is based on the Rasch model because unlike other IRT models it has a very clear interpretations of the item difficulties, which can be directly translated into statements that are clear to the experts. Because the Rasch model is often used in educational measurement applications and it has been shown to be rather robust in test linking situations, the applicability of the methods proposed in this study are still rather broad. However, in some applications the procedure would not be appropriate due to strong misfit of the Rasch model to the data.

The overall conclusion is that using informative priors can improve linking results. Expert judgments collected using the rulers method and included in the Bayesian estimation can increase the precision of linking without introducing a lot of bias and consequently decrease the expected proportion of misclassified persons.

References

Angoff, W. (1971). Scales, norms and equivalent scores. In R. Thorndike (Ed.), *Educational measurement* (pp. 508–597). Washington, DC: American Council of Education.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics, 32,* 159–188.

Béguin, A. A. (2000). *Robustness of equating high-stakes tests*. Unpublished doctoral dissertation, Enschede, the Netherlands. Retrieved from http://doc.utwente.nl/26293/

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

Christensen, R., Johnson, W., Brabscum, A., & Hanson, T. E. (2010). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. Boca Raton, FL: CRC press.

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

Fisher, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 36,* 359–374.

Geisinger, K. (1991). Using standard setting data to establish cutoff scores. *Educational Measurement: Issues and Practice, 10,* 17–22.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis, 1,* 403–420.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149.

Keizer-Mittelhaëuser, M. (2014). *Modeling the effect of differential motivation on linking educational tests* (Unpublished doctoral dissertation). Tilburg University, Tilburg, the Netherlands.

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

Laming, D. (2004). *Human judgement: The eye of the beholder*. London, UK: Thomson Learning.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 452–461.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement, 17,* 179–193.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton, FL: CRC Press.

Marco, G. L. (1977). Item characteristic curve solutions to thee intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

Marsman, M., Maris, G., Bechger, T., & Glas, C. A. (2014). *Composition algorithms for conditional distributions*. Manuscript submitted for publication.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21,* 1087–1092.

Miller, G. A. (1956). The magical number seven, plus or minus two. *The Psychological Review, 63,* 81–97.

Mislevy, R. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12,* 281–296.

Mislevy, R., Sheehan, K., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of educational measurement, 30,* 55–78.

Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (2015). Selecting a data collection design for linking in educational measurement: Taking differential motivation into account. In R. E. Milsap, L. A. van der Ark, D. M. Bolt, & W.-C. Wang (Eds.), *New developments in quantitative psychology: Presentations from the 78th annual psychometric society meeting* (pp. 181–193). New York, NY: Springer.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., . . . Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. New York, NY: Wiley.

Ozaki, K., & Toyoda, H. (2006). A paired comparison IRT model using 3-value judgement: Estimation of item difficulty parameters prior to administration of the test. *Behaviormetrika, 33,* 131–147.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (pp. 221–262). New York, NY: American Council of Education and Macmillan.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. (Expanded ed., 1980. Chicago, The University of Chicago Press)

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447–467.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Stone, M. (1961). The opinion pool. *Annals of Mathematical Statistics, 32,* 1339–1342.

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D., & Rivazi, S. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27,* 27–51.

Tatsuoka, K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24,* 233–245.

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology, 55,* 106–117.

von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.

Wauters, K., Desmet, P., & van der Noordgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers and Education, 58,* 1183–1193.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–464.

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

(*Appendices follow*)

Appendix A

Gibbs Sampler for Estimating the Cut-Score s_{pass}

Here we provide details about how to estimate the cut-score for the current test s_{pass}. To estimate s_{pass} we need to estimate the score distribution of the reference population on the current test (see $\mathbf{X}^*$ in Figure 1) for which samples from the joint posterior distribution

$$p(\delta_c^*, \mu_r, \sigma_r^2, \tau \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_G) \tag{22}$$

need to be obtained. To obtain samples from this multivariate distribution we use a Gibbs Sampler algorithm in which at each iteration each parameter of interest is sampled from its conditional posterior distribution given the current values of all other parameters. To simplify the conditional posterior distributions, not only the parameters μ_r, σ_r^2, $\boldsymbol{\delta}_c^*$ and τ are sampled but also the item parameters of the items in the reference test, the population parameters of the current population and the linking groups and the individual ability parameters of all the persons in the reference population (denoted by the vector $\boldsymbol{\theta}_r$), the current population (denoted by the vector $\boldsymbol{\theta}_c^*$) and the linking groups (denoted by the vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$).

The examination data sets $\mathbf{X}$ and $\mathbf{Y}$ (see Figure 1) are both split in two parts: one for constructing prior distributions, denoted by $\mathbf{X}^{(1)}$ and $\mathbf{Y}^{(1)}$, and another for the estimation of s_{pass}, denoted by $\mathbf{X}^{(2)}$ and $\mathbf{Y}^{(2)}$. The subsets of persons from the reference population whose responses are in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are denoted by $\{R^{(1)}\}$ and $\{R^{(2)}\}$, respectively. The subsets of persons from the current population whose responses are in $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$are denoted by $\{C^{(1)}\}$ and $\{C^{(2)}\}$, respectively. In the following subsections we will first describe how the training data $\mathbf{X}^{(1)}$ and $\mathbf{Y}^{(1)}$are used for constructing priors, second we will describe how using these priors, the estimation data $\mathbf{X}^{(2)}$ and $\mathbf{Y}^{(2)}$ and the data of the linking groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_G$, samples from the posterior distribution needed to determine the cut-score s_{pass} can be obtained, and finally we will show how to obtain the posterior distribution of s_{pass} and its estimate.

Using Training Data to Construct Priors

When constructing priors, we need to obtain samples from the posterior distributions:

$$\begin{aligned} p(\theta_r, \mu_r, \sigma_r^2, \delta_r \mid \mathbf{X}^{(1)}) \propto & \prod_{p\in\{R^{(1)}\}} \prod_{i\in\{r\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1+\exp(\theta_p - \delta_i)} \\ & \times \mathcal{N}(\mu_r; 0, 100)\text{Inv-}\mathcal{G}(\sigma_r^2; .001, .001) \\ & \times \prod_{i\in\{r\}} \mathcal{N}(\delta_i; 0, 100) \prod_{p\in\{R^{(1)}\}} \mathcal{N}(\theta_p; \mu_r, \sigma_r^2) \end{aligned} \tag{23}$$

and

$$\begin{aligned} p(\theta_c^*, \mu_c^*, \sigma_c^2, \delta_c^* \mid \mathbf{Y}^{(1)}) \propto & \prod_{p\in\{C^{(1)}\}} \prod_{i\in\{c\}} \frac{\exp(y_{pi}(\theta_p^* - \delta_i^*))}{1+\exp(\theta_p^* - \delta_i^*)} \\ & \times \mathcal{N}(\mu_c^*; 0, 100)\text{Inv-}\mathcal{G}(\sigma_c^2; .001, .001) \\ & \times \prod_{i\in\{c\}} \mathcal{N}(\delta_i^*; 0, 100) \prod_{p\in\{C^{(1)}\}} \mathcal{N}(\theta_p^*; \mu_c^*, \sigma_c^2). \end{aligned} \tag{24}$$

The normal priors for the means and the inverse-gamma priors for the variance are chosen because of the mathematical convenience of conditional conjugacy.

The initial values, denoted by a superscript (0), for all the parameters have to be chosen: $\mu_r^{(0)} = 0; \sigma_r^{2(0)} = 1; \mu_c^{*(0)} = 0, \sigma_c^{2(0)} = 1; \delta_i^{(0)} \sim U(-2, 2), \forall i \in \{r\}; \delta_i^{*(0)} \sim U(-2, 2), \forall i \in \{c\}, \tau^{(0)} = 0$. It is not needed to choose the initial values for the individual person parameters since they are sampled in the first step of the algorithm.

Below we describe how to sample from the posterior distribution in (Equation 23). Sampling from the posterior distribution in (Equation 24) is analogous to sampling from (Equation 23). The algorithm has five steps:

Step 1

$\forall p \in \{R^{(1)}\}$:

$$\theta_p \sim p(\theta_p \mid \ldots) = p(\theta_p \mid X_{p+}, \mu_r, \sigma_r^2, \delta_r), \tag{25}$$

which depends on the data only through the sumscore $X_{p+} = \sum_i X_{pi}$, because the Rasch model holds. Sampling from this distribution can be done using the conditional composition algorithm (Marsman, Maris, Bechger, & Glas, 2014):

a. Sample a candidate value from the population distribution $\theta \sim \mathcal{N}(\theta_p; \mu_r, \sigma_r^2)$
b. Simulate a vector of responses $\mathbf{X}$ to the items in the reference test

$$\Pr(X_i = 1 \mid \delta_i, \theta) = \frac{\exp(\theta - \delta_i)}{1+\exp(\theta - \delta_i)}. \tag{26}$$

c. Compute $X_+ = \sum_i X_i$. If $X_+ = X_{p+}$then θ is accepted as a sample from (25). Otherwise, Steps a, b, and c are repeated.

Step 2

$$\begin{aligned} \mu_r \sim p(\mu_r \mid \ldots) = p(\mu_r \mid \theta_r, \sigma_r^2) \propto \mathcal{N}(\mu_r; 0, 100) \\ \times \prod_{p\in\{R^{(1)}\}} \mathcal{N}(\theta_p; \mu_r, \sigma_r^2) = \mathcal{N}\left(\mu_r; \frac{\frac{\sum_{p\in\{R^{(1)}\}} \theta_p}{\sigma_r^2}}{\frac{1}{100} + \frac{N^{(1)}}{\sigma_r^2}}, \frac{1}{\frac{1}{100} + \frac{N^{(1)}}{\sigma_r^2}}\right), \end{aligned} \tag{27}$$

where $N^{(1)}$ is the number of persons in the subset $\{R^{(1)}\}$.

Step 3

$$\begin{aligned} \sigma_r^2 \sim p(\sigma_r^2 \mid \ldots) = p(\sigma_r^2 \mid \theta_r, \mu_r) \propto \text{Inv-}\mathcal{G}(\sigma_r^2; .001, .001) \prod_{p\in\{R^{(1)}\}} \mathcal{N}(\theta_p; \mu_r, \sigma_r^2) \\ = \text{Inv-}\mathcal{G}\left(\sigma_r^2; .001 + \frac{N^{(1)}}{2}, .001 + \frac{\sum_{p\in\{R^{(1)}\}}(\theta_p - \mu_r)^2}{2}\right). \end{aligned} \tag{28}$$

(*Appendices continue*)

Step 4

$\forall i \in \{r\}:$

$$\delta_i \sim p(\delta_i \mid \ldots) \propto \mathcal{N}(\delta_i; 0, 100) \prod_{p \in \{R^{(1)}\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)}. \tag{29}$$

The normalizing constant for this distribution does not have a closed form solution, therefore we use Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) to sample from this conditional posterior with a normal proposal density centered around the current value of the parameter.

Step 5

At the end of each iteration, the parameters have to be re-scaled to keep the chosen identification of the scale, namely $\bar{\delta}_r = 0$ $\mu_r = \mu_r - \bar{\delta}_r$, and $\delta_i = \delta_i - \bar{\delta}_r, \forall i \in \{r\}$. The individual person parameters do not have to be re-scaled since their values at the end of iteration t do not influence the values of the parameters in iteration $t + 1$.

By repeatedly going through these five steps, samples from the posterior in (Equation 23) and analogously in (Equation 24) are obtained. The posterior distributions of the population means and of the item difficulties can be approximated by normal distributions with the means and the variances equal the average values of these parameters across the samples from (Equation 23) and (Equation 24) and the variances of the sampled values of these parameters, respectively. The posterior distribution of the populations' variances can be approximated with inverse-gamma distributions with hyperparameters chosen based on the averages and the variance of the sampled values of these parameters. These approximations are used as priors for these parameters in the next step of the analysis (estimation of the cut-score s_{pass}):

$$\begin{aligned} p(\mu_r, \sigma_r^2, \delta_r, \mu_c^*, \sigma_c^2, \delta_c^*) = {} & \mathcal{N}(\mu_r; \mu_{r0}, \sigma_{r0}^2)\text{Inv-}\mathcal{G}(\sigma_r^2; \alpha_{r0}, \beta_{r0}) \\ & \times \prod_{i \in \{r\}} \mathcal{N}\left(\delta_i; \mu_{\delta_i}, \sigma_{\delta_i}^2\right)\mathcal{N}\left(\mu_c^*; \mu_{c0}^*, \sigma_{c0}^2\right) \\ & \times \text{Inv-}\mathcal{G}(\sigma_c^2; \alpha_{c0}, \beta_{c0}) \\ & \times \prod_{i \in \{c\}} \mathcal{N}\left(\delta_i^*; \mu_{\delta_i^*}, \sigma_{\delta_i^*}^2\right). \end{aligned} \tag{30}$$

Because the mean and the variance of a random variable with the inverse-gamma distribution with parameters α and β are equal to $\frac{\beta}{\alpha-1}$ and $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, respectively, we choose the following values for the hyperparameters:

$$\alpha_{r0} = \frac{(\bar{\sigma}_r^2)^2}{Var(\sigma_r^2)} + 2 \tag{31}$$

$$\beta_{r0} = (\bar{\sigma}_r^2)\left(\frac{(\bar{\sigma}_r^2)^2}{Var(\sigma_r^2)} + 1\right) \tag{32}$$

where $\bar{\sigma}_r^2$ and $Var(\sigma_r^2)$ are the average and the variance of the sampled values of σ_r^2. The hyper parameters for the distribution of σ_c^2 are chosen analogously.

The average sampled values of the item difficulties are used to facilitate the collection of expert judgements. In the Angoff method they are used to compute $\hat{\delta}_c$ and $\hat{\delta}_r$ (see Equation 8). In the rulers method they are used to select the items and to determine their position on the rulers which experts move in the third stage of the procedure.

Sampling From the Posterior Distribution Needed to Determine the Cut-Score

To estimate the cut-score s_{pass} we need to sample from the posterior:

$$\begin{aligned} & p(\theta_r, \mu_r, \sigma_r^2, \theta_c^*, \mu_c^*, \sigma_c^2, \theta_1, \mu_1, \sigma_1^2, \ldots, \\ & \theta_G, \mu_G, \sigma_G^2, \delta_r, \delta_c^*, \tau \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_G) \end{aligned} \tag{33}$$

which is proportional to the product of the density of the data:

$$\begin{aligned} f(\mathbf{X}^{(2)}, \mathbf{Y}^{(2)}, \mathbf{Z}_1, \ldots, \mathbf{Z}_G) = {} & \prod_{p \in \{R^{(2)}\}} \prod_{i \in \{r\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \\ & \times \prod_{p \in \{C^{(2)}\}} \prod_{i \in \{c\}} \frac{\exp\left(y_{pi}(\theta_p^* - \delta_i^*)\right)}{1 + \exp\left(\theta_p^* - \delta_i^*\right)} \\ & \times \prod_{g=1}^{G} \prod_{p \in \{E_g\}} \prod_{i \in \{e_g \cap r\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \prod_{i \in \{e_g \cap c\}} \frac{\exp\left(z_{gpi}(\theta_p - \delta_i^* - \tau)\right)}{1 + \exp\left(\theta_p - \delta_i^* - \tau\right)}, \end{aligned} \tag{34}$$

where $\{E_g\}$ denotes the set of persons in linking group G and $\{e_g\}$ denotes the set of items answered by linking group G; and the joint prior distribution

$$\begin{aligned} & p(\theta_r, \mu_r, \sigma_r^2, \theta_c^*, \mu_c^*, \sigma_c^2, \theta_1, \mu_1, \sigma_1^2, \ldots, \theta_G, \mu_G, \sigma_G^2, \delta_r, \delta_c^*, \tau) \\ & \quad = p(\tau)p(\mu_r, \sigma_r^2, \delta_r, \mu_c^*, \sigma_c^2, \delta_c^*) \\ & \quad\quad \times \prod_{p \in \{R^{(2)}\}} \mathcal{N}(\theta_p; \mu_r, \sigma_r^2) \prod_{p \in \{C^{(2)}\}} \mathcal{N}(\theta_p^*; \mu_c^*, \sigma_c^2) \\ & \quad\quad \times \prod_{g} \left(\mathcal{N}(\mu_g; 0, 100)\text{Inv-}\mathcal{G}\left(\sigma_g^2; .001, .001\right) \prod_{p \in \{E_g\}} \mathcal{N}\left(\theta_p; \mu_g, \sigma_g^2\right)\right), \end{aligned} \tag{35}$$

where the priors of the population means and variances of the reference and the current population, and the item difficulties are estimated from the training data (see Equation 30).

The initial values, denoted by a superscript (0), for all the parameters are chosen in the same way as when sampling from (Equation 23) and (Equation 24) with the following initial values for the additional parameters $\mu_g^{(0)} = 0, \sigma_g^{2(0)} = 1, \forall g \in [1:G]$

Sampling from the conditional posteriors of the parameters in (Equation 33) is similar to sampling from the conditional posteriors of θ_r, μ_r, σ_r^2, and δ_r in (Equation 23). The following steps are involved:

Step 1a

$\forall p \in \{R^{(2)}\}:$

$$\theta_p \sim p(\theta_p \mid \ldots) \propto \mathcal{N}\left(\theta_p; \mu_r, \sigma_r^2\right) \prod_{i \in \{r\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \tag{36}$$

which is analogous to sampling from (Equation 25).

(*Appendices continue*)

Step 1b

$\forall p \in \{C^{(2)}\}:$

$$\theta_p^* \sim p(\theta_p^* \mid \ldots) \propto \mathcal{N}(\theta_p^*; \mu_c^*, \sigma_c^2) \prod_{i \in \{c\}} \frac{\exp\left(y_{pi}(\theta_p^* - \delta_i^*)\right)}{1 + \exp\left(\theta_p^* - \delta_i^*\right)} \tag{37}$$

which is analogous to sampling from (Equation 25).

Step 1c

$\forall g \in [1:G], \forall p \in \{E_g\}:$

$$\theta_p \sim p(\theta_p \mid \ldots) \propto \mathcal{N}(\theta_p; \mu_g, \sigma_g^2) \prod_{i \in \{r \cap e_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \times \prod_{i \in \{c \cap e_g\}} \frac{\exp\left(z_{gpi}(\theta_p - \delta_i^* - \tau)\right)}{1 + \exp\left(\theta_p - \delta_i^* - \tau\right)}, \tag{38}$$

which is analogous to sampling from (Equation 25).

Step 2a

$$\mu_r \sim p(\mu_r \mid \ldots) = \mathcal{N}\left(\mu_r; \frac{\frac{\mu_{r0}}{\sigma_{r0}^2} + \frac{\sum_{p \in \{R^{(2)}\}} \theta_p}{\sigma_r^2}}{\frac{1}{\sigma_{r0}^2} + \frac{N^{(2)}}{\sigma_r^2}}, \frac{1}{\frac{1}{\sigma_{r0}^2} + \frac{N^{(2)}}{\sigma_r^2}}\right), \tag{39}$$

where $N^{(2)}$ is the number of persons in the set $\{R^{(2)}\}$.

Step 2b

$$\mu_c^* \sim p(\mu_c^*, \mid \ldots) = \mathcal{N}\left(\mu_c^*; \frac{\frac{\mu_{c0}^*}{\sigma_{c0}^2} + \frac{\sum_{p \in \{C^{(2)}\}} \theta_p^*}{\sigma_c^2}}{\frac{1}{\sigma_{c0}^2} + \frac{M^{(2)}}{\sigma_c^2}}, \frac{1}{\frac{1}{\sigma_{c0}^2} + \frac{M^{(2)}}{\sigma_c^2}}\right), \tag{40}$$

where $M^{(2)}$ is the number of persons in the set $\{C^{(2)}\}$.

Step 2c

$\forall g \in [1:G]:$

$$\mu_g \sim p(\mu_g, \mid \ldots) = \mathcal{N}\left(\mu_g; \frac{\frac{\sum_{p \in \{E_g\}} \theta_p}{\sigma_g^2}}{\frac{1}{100} + \frac{N_e}{\sigma_g^2}}, \frac{1}{\frac{1}{100} + \frac{N_e}{\sigma_g^2}}\right), \tag{41}$$

Step 3a

$$\sigma_r^2 \sim p(\sigma_r^2 \mid \ldots) = \text{Inv-}\mathcal{G}\left(\sigma_r^2; \alpha_{r0} + \frac{N^{(2)}}{2}, \beta_{r0} + \frac{\sum_{p \in \{R^{(2)}\}} (\theta_p - \mu_r)^2}{2}\right). \tag{42}$$

Step 3b

$$\sigma_c^2 \sim p(\sigma_c^2 \mid \ldots) = \text{Inv-}\mathcal{G}\left(\sigma_c^2; \alpha_{c0} + \frac{M^{(2)}}{2}, \beta_{c0} + \frac{\sum_{p \in \{C^{(2)}\}} (\theta_p^* - \mu_c^*)^2}{2}\right). \tag{43}$$

Step 3c

$\forall g \in [1:G]:$

$$\sigma_g^2 \sim p(\sigma_g^2 \mid \ldots) = \text{Inv-}\mathcal{G}\left(\sigma_g^2; .001 + \frac{N_e}{2}, .001 + \frac{\sum_{p \in \{E_g\}} (\theta_p - \mu_g)^2}{2}\right). \tag{44}$$

Step 4a

$\forall i \in \{r/\{e_1 \cap \ldots \cap e_G\}\}:$

$$\delta_i \sim p(\delta_i \mid \ldots) \propto \mathcal{N}\left(\delta_i; \mu_{\delta_i}, \sigma_{\delta_i}^2\right) \prod_{p \in \{R^{(2)}\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)}, \tag{45}$$

which is analogous to sampling from (Equation 29).

Step 4b

$\forall i \in \{c/\{e_1 \cap \ldots \cap e_G\}\}:$

$$\delta_i^* \sim p(\delta_i \mid \ldots) \propto \mathcal{N}\left(\delta_i^*; \mu_{\delta_i^*}, \sigma_{\delta_i^*}^2\right) \prod_{p \in \{C^{(2)}\}} \frac{\exp\left(y_{pi}(\theta_p^* - \delta_i^*)\right)}{1 + \exp\left(\theta_p^* - \delta_i^*\right)}, \tag{46}$$

which is analogous to sampling from (Equation 29).

Step 4c

$\forall g \in [1:G], \forall i \in \{r \cap e_g\}:$

$$\delta_i \sim p(\delta_i \mid \ldots) \propto \mathcal{N}\left(\delta_i; \mu_{\delta_i}, \sigma_{\delta_i}^2\right) \times \prod_{p \in \{R^{(2)}\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \prod_{p \in \{E_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)}, \tag{47}$$

which is analogous to sampling from (Equation 29).

(*Appendices continue*)

Step 4d

$$\forall g \in [1:G], \forall i \in \{c \cap e_g\}:$$

$$\delta_i^* \sim p(\delta_i^* \mid \ldots) \propto \mathcal{N}\left(\delta_i^*; \mu_{\delta_i^*}, \sigma^2_{\delta_i^*}\right)$$
$$\times \prod_{p \in \{C^{(2)}\}} \frac{\exp\left(x_{pi}(\theta_p^* - \delta_i^*)\right)}{1 + \exp\left(\theta_p^* - \delta_i^*\right)} \prod_{p \in \{E_g\}} \frac{\exp\left(z_{gpi}(\theta_p - \delta_i^* - \tau)\right)}{1 + \exp\left(\theta_p - \delta_i^* - \tau\right)}, \tag{48}$$

which is analogous to sampling from (Equation 29).

Step 5

$$\tau \sim p(\tau \mid \ldots) \propto p(\tau) \prod_{g=1}^{G} \prod_{p \in E_g} \prod_{i \in \{c \cap e_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i^* - \tau))}{1 + \exp(\theta_p - \delta_i^* - \tau)}, \tag{49}$$

which is similar to sampling from the conditional posterior distributions of the item difficulties, the Metropolis algorithm is used to sample from this distribution.

Step 6

The parameters are re-scaled to make sure that $\bar{\delta}_r = 0$ and $\bar{\delta}_c^* = 0 : \mu_r = \mu_r - \bar{\delta}_r, \mu_g = \mu_g - \bar{\delta}_r, \forall g \in [1:G], \delta_i = \delta_i - \bar{\delta}_r, \forall i \in \{r\}, \mu_c^* = \mu_c^* - \bar{\delta}_c^*$, and $\delta_i^* = \delta_i^* - \bar{\delta}_c^*, \forall i \in \{c\}$.

Estimating the Cut-Score s_{pass}

After the burn-in, at each iteration t the unobserved responses of the persons from the reference population to the current exam ($\mathbf{X}^*$) are simulated according to the Rasch model using the values of the model parameters at iteration t sampled from (Equation 33) using the Gibbs sampler described in the previous subsection:

$$x_{pi}^{*(t)} \sim \text{Bernoulli}\left(\frac{\exp(\theta_p^{(t)} - (\delta_i^{*(t)} + \tau^{(t)}))}{1 + \exp(\theta_p^{(t)} - (\delta_i^{*(t)} + \tau^{(t)}))}\right),$$
$$\forall p \in \{R^{(2)}\}, \forall i \in \{c\}. \tag{50}$$

A sample from the posterior distribution of the cut-score, denoted by $s_{pass}^{(t)}$, is such a score that the number of students from the reference population with observed scores on the reference test below s_{ref} is as close as possible to the number of students from the reference population with simulated scores on the current test at iteration t below this score:

$$s_{pass}^{(t)} = \text{argmin}_s\left(\left(\sum_{p \in \{R^{(2)}\}} \left(\mathcal{I}\left(\sum_{i \in \{r\}} x_{pi} < s_{ref}\right) - \mathcal{I}\left(\sum_{i \in \{c\}} x_{pi}^{*(t)} < s\right)\right)\right)^2\right). \tag{51}$$

Using a large number of sampled values from the posterior in (Equation 33), a sequence of values $\{s_{pass}^{(1)}, s_{pass}^{(2)}, \ldots, s_{pass}^{(T)}\}$ is obtained, which is a sample from the posterior distribution of the cut-score is obtained. The maximum a posteriori estimate of s_{pass} is the mode of this sample. The posterior variance of s_{pass} is the variance in this sample.

(Appendices continue)

Appendix B

Approximating the Experts' Judgements With a Skew-Normal Distribution

In this section we describe how to choose the parameters of the skew-normal distribution $p_{ej}(\tau)$, such that its mode, fifth percentile and 95th percentile would match the values of $\hat{\tau}_{ej}$, $\tau_{ej}^{\min}$ and $\tau_{ej}^{\max}$, respectively. The skew-normal distribution with parameters α, ξ, ω is given by:

$$f(x) = \frac{2}{\sqrt{2\pi}\omega}\exp\left(\frac{-(x-\xi)^2}{2\omega^2}\right)\int_{-\infty}^{\alpha\frac{x-\xi}{\omega}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt. \quad (52)$$

First, we determine the value of the skewness parameter α. Let us by $q_p(\alpha, \xi, \omega)$ denote the p-th percentile of the skew-normal distribution with parameters α, ξ and ω and by $m(\alpha, \xi, \omega)$ the mode of this distribution. The larger the skewness is, the further away from 1 the ratio below is:

$$r(\alpha, \xi, \omega) = \frac{q_{95}(\alpha, \xi, \omega) - m(\alpha, \xi, \omega)}{m(\alpha, \xi, \omega) - q_5(\alpha, \xi, \omega)}. \quad (53)$$

The values of the parameters ξ and ω do not influence the value of this ratio. For all values of α ranging from −4 to 4, with equal interval steps of .001, we estimated the mode $m(\alpha, 0, 1)$—with the precision up to .0001, which is sufficient for the application at hand—and computed the ratio $r(\alpha, 0, 1)$, using the "sn" R-package. And then for each judgement of each expert we chose:

$$\alpha_{ej} = \arg\min_{\alpha} \left| \frac{\tau_{ej}^{\max} - \hat{\tau}_{ej}}{\hat{\tau}_{ej} - \tau_{ej}^{\min}} - r(\alpha, 0, 1) \right| . \quad (54)$$

Second, we choose the value of the parameter ω which determines the spread of the distribution:

$$\omega_{ej} = \frac{\tau_{ej}^{\max} - \tau_{ej}^{\min}}{q_{95}(\alpha_{ej}, 0, 1) - q_5(\alpha_{ej}, 0, 1)}. \quad (55)$$

And finally, we choose the value of the parameter ξ which determines the location of the distribution:

$$\xi_{ej} = \tau_{ej}^{\max} - q_{95}(\alpha_{ej}, 0, 1)\omega_{ej}. \quad (56)$$

Next, we show how to approximate the expert-specific normalizing constant, denoted by Z_e, for the product of skew-normal distributions in Equation 19:

$$Z_e = \int_{-\infty}^{+\infty}\left(\prod_{j=1}^{J_e}\frac{2}{\omega_{ej}}\phi\left(\frac{\tau - \xi_{ej}}{\omega_{ej}}\right)\Phi\left(\alpha_{ej}\frac{\tau - \xi_{ej}}{\omega_{ej}}\right)\right)d\tau, \quad (57)$$

where $\phi(x)$ denotes the standard normal density and $\Phi(x)$ denotes the standard normal cumulative distribution function. This integral can be approximated using the Gauss-Hermite the weights $\mathbf{w} = \{w_1, \ldots, w_K\}$ and the nodes $\mathbf{y} = \{y_1, \ldots, y_K\}$:

$$Z_e \approx \frac{2}{\sqrt{\pi}}\sum_{i=1}^{K} w_i \Phi(\sqrt{2}\alpha_{e1}y_i) \times \prod_{j=2}^{J_e}\frac{2}{\omega_{ej}}\phi\left(\frac{\sqrt{2}\omega_{e1}y_i + \xi_{e1} - \xi_{ej}}{\omega_{ej}}\right)\Phi\left(\alpha_{ej}\frac{\sqrt{2}\omega_{e1}y_i + \xi_{e1} - \xi_{ej}}{\omega_{ej}}\right). \quad (58)$$

This integral has to be computed only once for each expert; therefore, we can use a very large number of nodes to obtain an accurate approximation. In the empirical example we used $K =$ 20, 000.

Received September 18, 2015
Revision received June 1, 2016
Accepted November 1, 2016 ■