Using meta-analysis, this research takes a look at studies included in a meta-analysis by R. Jaeger (1989) that compared the cut score set by one standard setting method with that set by another. This meta-analysis looks beyond Jaeger's studies to select 10 from the research literature. Each compared at least two types of standard setting method. To assess the difference in cut scores produced by each standard setting method, a common metric was used for every cut-score comparison. The standardized magnitude of the difference between two compared cut scores, called the effect size, was calculated, with the Angoff method results treated as the control group in effect size calculations. Depending on the conceptualization chosen, results of this analysis may be interpreted differently. For a fixed effects model approach, some standard setting methods produce significantly different cut scores. This conclusion must be tempered by the amount of heterogeneity in the model. If a random effects model is endorsed, it is recognized that no significant differences between methods are produced. This conclusion also must be taken with caution because of the relationship between the within-method variation and the between-method variation. Although the analysis results do not make it possible to make conclusive statements about systematic differences in effect sizes and cut scores produced by the standard setting methods presented, meta-analysis holds much promise in its ability to answer these questions in the future. (Contains 4 tables, 3 figures, and 26 references.) (SLD)
A Meta-Analytic Assessment Of Empirical Differences
In Standard Setting Procedures

BRIAN D. BONTEMPO
CASIMER M. MARKS
GEORGE KARABATSOS

UNIVERSITY OF CHICAGO

A Meta-Analytic Assessment Of Empirical Differences
In Standard Setting Procedures

In testing, setting performance standards involves identifying cut scores that divide examinees into groups such as pass/fail, master/non-master, or certify/deny certification. Performance standards are used to make very important decisions in education and the job market. Standard-setting methods are also used to classify test takers into multiple levels of performance. A simple example is assigning grades of A, B, C, D, or F to examinees. Standards decide whether people are competent enough to work as teachers, school administrators, nurses, dentists, doctors, or other types of professionals. Standards also determine whether students are proficient enough to graduate, enter educational institutions, or be placed in certain classrooms.

In setting a standard, there are many methods to choose from, all of which have been attacked and defended from both a theoretical and empirical perspective (see Reference section). Many empirical studies claim that different standard setting procedures yield different cut scores. Jaeger (1989) summarized this research by looking at the results of 12 different studies. These twelve studies reported the cut score set by one method with the cut score set by another method. Within these studies, multiple standard setting procedures were conducted on each of 32 different examinations. Jaeger calculated the ratio of the highest/lowest cut score and the highest/lowest expected failure rate for each examination. When analyzed this way, the results indicate that the different methods do produce different cut scores. The median ratio of the cut score was approximately 1½, indicating that one procedure was 1½ times as stringent as another.

Although Jaeger’s findings are interesting, they are not comprehensive. Jaeger acknowledged that there was great deal of variation in the ratios. This may be attributable to the nature of the ratios themselves. In some of the 32 contrasts, the Angoff standard (1971) may have been the most stringent, making it the numerator, while in others, it may have been the least stringent, making it the denominator. Furthermore, the ratios seldom compared the difference in cut score of the same two methods. Sometimes, a Nedelsky's cut score (1954) may have been compared to an Angoff, while other times, an Ebel (1972) cut score may have been compared to a Contrasting Groups (Livingston & Zieky, 1982).
Using meta-analysis, this research takes a deeper look at the studies in Jaeger’s research, by comparing cut scores derived by the Nedelsky (1954), Ebel (1972), Angoff (1971) in all of its modified versions, Jaeger (1982), and the Borderline/Contrasting Groups methods (Livingston & Zieky, 1982). This meta-analysis also looks beyond the articles in the Jaeger study to the entire literature base on standard setting procedures, and infers that different standard-setting procedures do not systematically yield different cut scores. This result is important because it provides validation for choosing a standard setting method less for its statistical & theoretical properties and more for its ease of implementation. Indeed, if the decision to use a certain method can be based on issues of implementation, having assurance that the choice of method will not systematically influence the cut score produced, testing organizations can be more efficient and productive in their test development and maintenance.

Method

Data Collection

Studies were collected from many sources within the published professional literature and papers presented at the annual meeting of the American Educational Research Association. Collection methods were designed to be comprehensive enough to represent the current state of empirical research conducted in the area of standard setting. It is noted that there is a high degree of overlap between studies used in this analysis and Jaeger’s (1989). However, the studies included in this analysis were collected from a completely independent literature search. In order for an article to be used, it had to provide a comparison of at least two types of standard setting methods by stating the cut score that each method rendered as well as a measure of the variance or error. The data allowed for over ninety comparisons from ten different articles. All standards were produced for multiple choice tests which varied in content, age of examinees, importance, and length. The exact standard-setting procedures may have differed in how they were executed in each study. The procedures varied in how much judgment was made, what types of normative information were provided to the judges, and how the groups of judges were divided. Nonetheless, each cut score was classified by its theoretical underpinnings, e.g., all of the modified-Angoff procedures were
grouped together. Among the studies collected, the group of modified-Angoff procedures was the most frequently encountered. The following section describes the articles from which data were used.

Behuniak, Archambault, and Gable (1982) compared the standards set by content specialists when they applied the Angoff and Nedelsky procedures to the Connecticut School System Test for reading and mathematics. These specialists were split into 8 parallel groups of 3-4 judges. Each group executed one of the standard setting procedures by making judgments on 30 items of either the reading or math test.

Brennan and Lockwood (1980) used generalizability theory to "characterize and quantify" the expected variance in cut scores resulting from the Nedelsky and Angoff procedures. A group of 5 judges ran through both the Angoff and Nedelsky standard setting procedures for a 126-item test in a "health-related" subject area.

Cross, Impara, Frary, and Jaeger (1984) compared the standards set by the Angoff, Jaeger, and Nedelsky methods for a national teaching examination focusing on mathematics and elementary education. The elementary education examination consisted of 150 items, while the mathematics examination consisted of 120 items. For each test, 15 judges were divided up into 3 panels of 5 judges. Each panel conducted one of the standard setting procedures in three iterative sessions using different portions of the test and different normative feedback information.

In a study involving multiple elementary schools, Livingston and Zieky (1989) compared the standards set on the ETS Basic Skills Assessments tests for reading and math. In eight different middle schools, two groups of judges performed three standard setting procedures, the Contrasting Groups, Borderline Group, and either the Nedelsky or Angoff method. There were 3-5 judges in each group. In each middle school, one group reviewed the math test, while the other group judged reading.

Mills (1983) compared the standards set by the Angoff, Contrasting Groups, and the Borderline Group methods on Louisiana's 2nd grade basic skill tests (having between 30-60 items each). Six different overlapping test forms for both the language arts and math section were reviewed by two groups of judges. Sixteen judges reviewed all 6 forms of the math examination, while 15 judges reviewed all 6 forms of the language arts examination.
Mills and Melican (1988) also compared the standards set for the elementary education and mathematics sections of the National Teachers Examination. In this study, four groups of judges were formed. Each group performed one method, either the Angoff or the Nedelsky for one section of the NTE.

Smith and Smith (1988) compared the standards set by the Angoff and Nedelsky procedures. Working with the 64-item high school reading competency test in New Jersey, 31 judges performed one of the two standard setting procedures. These judges were randomly assigned to a procedure, 16 in one group, 15 in the other.

Three different standard setting procedures, Ebel, Nedelsky, and Angoff were used to set standards on the Missouri College English Test (Halpin and Halpin, 1987; Halpin, Sigmon, and Halpin, 1983). Three non-parallel groups of judges, 5 graduate students, 5 high school teachers, and 5 university faculty executed all three procedures by looking at all 90 items of the test.

Baron, Rindone, and Prowda (1981) also contrasted the cut scores set for Connecticut's basic skill tests for reading and mathematics. The Angoff, Nedelsky, Contrasting Groups, and Borderline Group, methods were employed. In using the first two methods, four groups of approximately 10 judges evaluated one section of the examination using either the Angoff or Nedelsky method. For the latter two methods, teachers at over 200 schools were asked to evaluate a group of 30 students selected by the principal at random.

In evaluating the Kansas Competency Tests, Poggio, Glasnapp, and Eros (1981) employed the Ebel, Angoff, and Nedelsky methods. In this study, cut scores were produced for ten different examinations, five reading and five math, for grades 2, 4, 6, 8, and 11. For each test, three parallel groups of approximately 25 judges evaluated the examination using one of the standards setting methods. For a synopsis of all standard setting procedures, number of judges involved, test content and number of effect sizes estimates obtained from each study see Table 1.

Computation of Effect Size

In order to assess the difference in cut scores produced by each standard setting method, a common metric was employed for every cut-score comparison in the data. The standardized magnitude of the difference between two compared cut scores, called the effect size, was calculated. Due to the
dominant use of the Angoff procedure (Cizek, 1996, Plake 1998), this method was treated as the control
group in effect size calculations. In viewing this study as a comparison of the Angoff procedure with the
other procedures, it is appropriate to calculate effect sizes using Glass’s $\Delta$ (Glass, McGaw, & Smith 1981).

$$\text{Glas's } \Delta = \frac{C_i - C_A}{S_A}$$

(1)

where, $C_A$ = cut score set by a modified-Angoff procedure, $C_i$ = cut score set by an alternative procedure,
and $S_A$ = standard deviation of the modified-Angoff cut score. The variance for Glass’s $\Delta$ was calculated
by:

$$\text{Var (}\Delta\text{)} = \frac{n_i + n_A}{n_i n_A} + \frac{\Delta^2}{2(n_A - 1)}$$

(2)

where, $n_A$ = number of judges who set cut scores using a modified-Angoff procedure, and $n_i$ = number of
judges who set cut scores via an alternative method.

Statistical Analyses of Effect Sizes

The mean effect size was used to determine if the group of cut-score comparisons was
significantly different from zero. In order to ascertain if there was a significant difference in the effect size
measures across methods, the effect sizes were analyzed using fixed and random effects one-way ANOVA
models.

Effect sizes were grouped to produce a one-factor model of five different comparisons:
Borderline/modified-Angoff, Contrasting Groups/modified-Angoff, Ebel/modified-Angoff,
Jaeger/modified-Angoff, and Nedelsky/modified-Angoff. The rational for this separation was to see if the
five non-modified-Angoff methods produced an effect when separated from the rest.

The Q statistic (Hedges, 1994) was used to assess the model assumption of homogeneity of
variance. In the one-factor model, the Q statistic takes the following form:
In the random effects model the variance component (between-studies variance) was calculated as follows:

\[
Q_{\text{BETWEEN}} = \sum_{i=1}^{k} W_i (\Delta_i - \Delta_*)^2
\]  

(4)

\[
Q_{\text{WITHIN}} = \sum_{i=1}^{k} \sum_{j=1}^{m_i} W_{ij} (\Delta_{ij} - \Delta_*)^2
\]  

(5)

\[
Q = Q_{\text{BETWEEN}} + Q_{\text{WITHIN}}.
\]  

(6)

\[
\hat{\sigma}_0^2 = \left[ Q - (k - 1) \right] / \sum_{i=1}^{k} w_i - \left[ \sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i \right]
\]  

(7)

Results

Fixed effects model

The overall standardized mean effect size difference was not significantly different from 0 \((\Delta_\text{z} = -0.02, z = -0.45, p > .05)\). For a graphical representation of all effect size estimates and effect size estimates by method for the fixed effects model see Figures 1. The fixed effects model indicated that the Borderline \((z_\text{z} = 10.31)\) and the Jaeger \((z_\text{z} = 3.72)\) methods produced significantly higher cut scores than the modified-Angoff methods \((p < .05)\) while the Nedelsky method produced significantly lower cut scores \((Z_\text{n} = -14.02)\) (See Figure 2). As expected, the variance was heterogeneous \((Q = 676, df = 91)\), both between \((Q_{\text{BETWEEN}} = 329, df = 4)\) and within \((Q_{\text{WITHIN}} = 347, df = 87)\) the five groups \((p < .05)\). These results show that there may be differences in the cut scores set from different types of standard setting procedures, however, due to the heterogeneity of the variance within groups caution should be exercised in believing the results derived from this model without further exploration.
One method used to control for the heterogeneity of effect sizes is to use a criterion to partition the data. For this analysis the number of judges used in the standard setting procedure was used as the criterion to control for heterogeneity. Standard Setting often involves a small group of judges. The studies gathered for this research were typical in this respect. The average Angoff group had 14 members, and the standard deviation was 10. The nature of these small sample sizes and their great variability had a large impact on the results. The effect size estimates from those studies with larger panels had less variance and were more influential in the results (see Table 3).

Therefore, the studies were separated into two groups, those studies with Angoff panels of less than 15 judges and those with 15 or more judges. When arranged this way, the mean effect sizes were not significantly different from zero ($\Delta_{\text{small}} = -0.034$, $\Delta_{\text{large}} = -0.019$, $p > .05$). Furthermore, there was no difference between these two groups ($Q_{\text{between}} = .01$), but there was a significant amount of variability within groups ($Q_{\text{within\ small}} = 189$, $Q_{\text{within\ large}} = 488$).

Further attempts at reduction of within group variation, such as a two-factor model of size by method, were not computed because there were too many empty cells in the matrix. A simpler approach would have been to compare the large judge panel group to the small judge panel group within method, but this was impossible due to the confounding "study" effects (See Table 4). We did, however, run an analysis using only the small panels of judges. In this analysis ($Q_{\text{between}} = 28$, $Q_{\text{within}} = 161$), the within cell variance was greatly reduced. Despite the reduction of heterogeneity achieved through the identification and separation of studies based on the number of judges, the amount of within group variance was higher than recommended for continued use of a fixed effects model.

Random effects model

Although some of the heterogeneity found in the fixed effects may have been due to randomness resulting from sampling variability, it was also most likely due to some uncertainty involved in the standard setting process. Regardless of the conceptual position chosen, the number of potential moderator variables were too numerous to identify and account for with the small number of studies available for this analysis. For this reason it was reasonable to apply a random effects model to the data. When all data points were weighed equally, the mean effect size was not significant ($\Delta = 0.19$, $z = 0.17$, $p > .05$).
one factor model also revealed that for all but the Jaeger comparison, the mean effect size was not significantly different than 0 ($z_{\text{Jaeger}} < 1.96, p > .05$) (See Table 2 for all random effects statistics). It should be noted that the variance component (between-studies variance) of the Jaeger comparisons was not calculated because the Q statistic was not significantly different than 0 (Shadish and Haddock, 1994).

In assessing the adequacy of the random effects model, it was necessary to investigate the pattern of variance for each of the effect size estimates. As seen in Figure 3, the introduction of the additional variance component changed the relationship between the effect size estimates. All five of the confidence intervals overlapped, whereas for the fixed effects model, only two overlapped. This pattern indicated that for the effect size estimates there was no significant difference between standard setting methods, however, the variance of these estimates indicated that there was at least as much variation within the standard setting methods analyzed as there was between the methods.

Conclusion

Before drawing further conclusions from the results of these analyses, the limitations of the study must be identified. Most notably, this meta-analysis was conducted on a small number of studies. Using this small of a sample is problematic because it limits the stability and generalizability of the findings. Specifically, the variance component for the random effects model is assumed to be known although it is estimated from the data. With such a small sample, this estimate is subject to a high degree of error. One other limitation of the meta-analysis is the presence of “study” effects. The ninety comparisons came from only ten studies where many of the effect sizes were correlated distorting the results and conclusions. Finally, the statistical power of these analyses was low. To alleviate these problems, more data must be identified and collected.

The strength of what can be concluded from these analyses depends on the conception of the problem. If one believes that the data presented for analysis are “true” effect sizes, then the conclusion that there is some difference in the cut scores produced by different standard setting methods should be maintained. If this approach is endorsed, more data must be gathered to allow for further multi-factor analyses to control for the heterogeneity of variance encountered in this study. In fact, the field already
recognizes this in some respects. Many studies have attempted to explain the heterogeneity in cut scores by introducing modifications to the established standard setting procedures. Approaches in the literature have included; changing the number of judges involved in the standard setting process, providing the judges with normative feedback, and allowing discussion amongst the judges (Brennan & Lockwood, 1980; Halpin & Halpin, 1987; Koffler, 1980). A future meta-analysis should be conducted to investigate the effects of these modifications.

In the absence of a way to control for these identified sources of variance, other theoretical approaches are justified. In another conceptualization of the research question, the data are seen as randomly varying effect size estimates. In this approach, effect size variation is inflated due to the addition of a randomness component. The effect size estimates in this model are seen as having been drawn randomly from a "universe" of effect size estimates rather than the "true" values of these differences.

Depending on the conceptualization chosen, the results of this study may be interpreted differently. For a fixed effects model approach, some standard setting approaches produce significantly different cut scores. This interpretation, however, must be tempered by the amount of heterogeneity in the model. If a random effects model is endorsed, it is recognized that no significant differences between methods are produced. Again, this conclusion must be taken with caution because of the relationship between the within method variation and the between method variation. Moving away from these extremes in interpretation, the most important conclusion to be drawn is that the variability within standard setting method is at least as large as any difference between standard setting methods. In other words, the variability within identifiable and recognized standard setting procedures is too great to be able to make definitive statements about the relative differential effect between standards setting methods.

Although the final analysis is unable to make conclusive statements about systematic differences in effect sizes and cut scores produced by the standard setting methods presented, meta-analysis holds much promise in its ability to answer these questions in the future. Systematically investigating these different approaches and the cut scores they produce would benefit testing and certification organizations, providing empirical evidence and justification of the use of a particular standard setting method.
REFERENCES


<table>
<thead>
<tr>
<th>Article</th>
<th>Control Method</th>
<th>Control Judges</th>
<th>Alternative Method</th>
<th>Alternative Judges</th>
<th>Test</th>
<th># of ES estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behuniak, Archambault, Gable</td>
<td>Angoff</td>
<td>Group of 6 judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 8 judges</td>
<td>Connecticut School System Reading Test</td>
<td>1</td>
</tr>
<tr>
<td>Behuniak, Archambault, Gable</td>
<td>Angoff</td>
<td>Group of 7 judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 6 judges</td>
<td>Connecticut School System Math Test</td>
<td>1</td>
</tr>
<tr>
<td>Baron, Rindone, Prowda</td>
<td>Angoff</td>
<td>Group of 10 judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 11 judges</td>
<td>Connecticut School System Reading Test</td>
<td>1</td>
</tr>
<tr>
<td>Baron, Rindone, Prowda</td>
<td>Angoff</td>
<td>Group of 9 judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 9 judges</td>
<td>Connecticut School System Math Test</td>
<td>1</td>
</tr>
<tr>
<td>Baron, Rindone, Prowda</td>
<td>Angoff</td>
<td>Group of 10 judges</td>
<td>Contrasting Groups</td>
<td>1 teacher at 200 different schools</td>
<td>Connecticut School System Reading Test</td>
<td>1</td>
</tr>
<tr>
<td>Baron, Rindone, Prowda</td>
<td>Angoff</td>
<td>Group of 9 judges</td>
<td>Contrasting Groups</td>
<td>1 teacher at 200 different schools</td>
<td>Connecticut School System Math Test</td>
<td>1</td>
</tr>
<tr>
<td>Cross, Impara, Frary, Jaeger</td>
<td>Angoff</td>
<td>Group of 5 Angoff judges</td>
<td>Jaeger</td>
<td>Different parallel group of 5 Jaeger judges</td>
<td>NTE, Math</td>
<td>3</td>
</tr>
<tr>
<td>Cross, Impara, Frary, Jaeger</td>
<td>Angoff</td>
<td>Group of 5 Angoff judges</td>
<td>Jaeger</td>
<td>Different parallel group of 5 Jaeger judges</td>
<td>NTE, Elem Ed</td>
<td>3</td>
</tr>
<tr>
<td>Cross, Impara, Frary, Jaeger</td>
<td>Angoff</td>
<td>Same Group of 5 Angoff judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 5 Nedelsky judges</td>
<td>NTE, Math</td>
<td>3</td>
</tr>
<tr>
<td>Cross, Impara, Frary, Jaeger</td>
<td>Angoff</td>
<td>Same Group of 5 Angoff judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 5 Nedelsky judges</td>
<td>NTE, Elem Ed</td>
<td>3</td>
</tr>
<tr>
<td>Brenan &amp; Lockwood</td>
<td>Angoff</td>
<td>Group of 5 judges</td>
<td>Nedelsky</td>
<td>Same group of 5 judges</td>
<td>Health Related</td>
<td>1</td>
</tr>
<tr>
<td>Mills</td>
<td>Angoff</td>
<td>Group of 16 judges</td>
<td>Borderline Group</td>
<td>Same group of 16 judges</td>
<td>Louisiana Grade 2 Basic Skills Test Math (6 forms)</td>
<td>6</td>
</tr>
<tr>
<td>Mills</td>
<td>Angoff</td>
<td>Group of 15 judges</td>
<td>Borderline Group</td>
<td>Same group of 15 judges</td>
<td>Louisiana Grade 2 Basic Skills Test Reading (6 forms)</td>
<td>6</td>
</tr>
<tr>
<td>Mills</td>
<td>Angoff</td>
<td>Group of 16 judges</td>
<td>Contrasting Groups</td>
<td>Same group of 16 judges</td>
<td>Louisiana Grade 2 Basic Skills Test Math (6 forms)</td>
<td>6</td>
</tr>
<tr>
<td>Mills</td>
<td>Angoff</td>
<td>Group of 15 judges</td>
<td>Contrasting Groups</td>
<td>Same group of 15 judges</td>
<td>Louisiana Grade 2 Basic Skills Test Reading (6 forms)</td>
<td>6</td>
</tr>
<tr>
<td>Livingston &amp; Zieky</td>
<td>Angoff</td>
<td>4 different schools (5 judges at each)</td>
<td>Borderline Group</td>
<td>Same group of 5 judges at each school</td>
<td>Basic Skills Test Elementary Reading</td>
<td>4</td>
</tr>
<tr>
<td>Livingston &amp; Zieky</td>
<td>Angoff</td>
<td>4 different schools (5 judges at each)</td>
<td>Borderline Group</td>
<td>Same group of 5 judges at each school</td>
<td>Basic Skills Test Elementary Math</td>
<td>4</td>
</tr>
<tr>
<td>Livingston &amp; Zieky</td>
<td>Angoff</td>
<td>Same group of 5 judges at each school</td>
<td>Contrasting Groups</td>
<td>Same group of 5 judges at each school</td>
<td>Basic Skills Test Elementary Reading</td>
<td>4</td>
</tr>
<tr>
<td>Livingston &amp; Zieky</td>
<td>Angoff</td>
<td>Same group of 5 judges at each school</td>
<td>Contrasting Groups</td>
<td>Same group of 5 judges at each school</td>
<td>Basic Skills Test Elementary Math</td>
<td>4</td>
</tr>
<tr>
<td>Name</td>
<td>Method</td>
<td>Judges Grouping</td>
<td>Judges</td>
<td>Test</td>
<td>Count</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
<td>----------------------------------</td>
<td>--------------</td>
<td>-----------------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Haplin &amp; Haplin</td>
<td>Angoff</td>
<td>3 non parallel groups of 5 judges</td>
<td>Ebel</td>
<td>Same 3x5 judges</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Haplin &amp; Haplin</td>
<td>Angoff</td>
<td>Same 3x5 judges</td>
<td>Nedelsky</td>
<td>Same 3x5 judges</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Smith &amp; Smith</td>
<td>Angoff</td>
<td>Group of 15 judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 16 judges</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Poggio, Glasnapp, Eros</td>
<td>Angoff</td>
<td>10 separate groups of judges (N=30)</td>
<td>Nedelsky</td>
<td>Different parallel groups of judges</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Poggio, Glasnapp, Eros</td>
<td>Angoff</td>
<td>10 separate groups of judges (N=30)</td>
<td>Ebel</td>
<td>Different parallel groups of judges</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Mills &amp; Melican</td>
<td>Angoff</td>
<td>Group of 10 judges</td>
<td>Nedelsky</td>
<td>Different parallel groups of 3 judges</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mills &amp; Melican</td>
<td>Angoff</td>
<td>Group of 13 judges</td>
<td>Nedelsky</td>
<td>Different parallel groups of 4 judges</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Smith &amp; Smith</td>
<td>Angoff</td>
<td>Group of 15 judges</td>
<td>Nedelsky</td>
<td>Different parallel group of 16 judges</td>
<td>NJ Reading Test</td>
<td></td>
</tr>
<tr>
<td>Poggio, Glasnapp, Eros</td>
<td>Angoff</td>
<td>10 separate groups of judges (N=30)</td>
<td>Nedelsky</td>
<td>Different parallel groups of judges</td>
<td>Kansas Competency Tests (10 different tests)</td>
<td></td>
</tr>
<tr>
<td>Poggio, Glasnapp, Eros</td>
<td>Angoff</td>
<td>10 separate groups of judges (N=30)</td>
<td>Ebel</td>
<td>Different parallel groups of judges</td>
<td>Kansas Competency Tests (10 different tests)</td>
<td></td>
</tr>
<tr>
<td>Mills &amp; Melican</td>
<td>Angoff</td>
<td>Group of 10 judges</td>
<td>Nedelsky</td>
<td>Different parallel groups of 3 judges</td>
<td>NTE, Math</td>
<td></td>
</tr>
<tr>
<td>Mills &amp; Melican</td>
<td>Angoff</td>
<td>Group of 13 judges</td>
<td>Nedelsky</td>
<td>Different parallel groups of 4 judges</td>
<td>NTE, Elem Ed</td>
<td></td>
</tr>
</tbody>
</table>
### Table 2. Statistical Results Tables

<table>
<thead>
<tr>
<th>Method</th>
<th>k</th>
<th>t. (unweighted)</th>
<th>t. (weighted)</th>
<th>Var(T.)</th>
<th>SE(T.)</th>
<th>Z</th>
<th>Q*</th>
<th>Ci</th>
<th>t.* (weighted)</th>
<th>Var(T.*)</th>
<th>sd(T.*)</th>
<th>Ci</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Borderline Groups</strong></td>
<td>22</td>
<td>1.02</td>
<td>1.22</td>
<td>0.014</td>
<td>0.118</td>
<td>10.3</td>
<td>70.3</td>
<td>.99,1.45</td>
<td>1.22</td>
<td>0.80</td>
<td>0.893</td>
<td>-0.53,2.97</td>
</tr>
<tr>
<td><strong>Contrasting Groups</strong></td>
<td>22</td>
<td>-0.39</td>
<td>0.0056</td>
<td>0.010</td>
<td>0.100</td>
<td>0.06</td>
<td>69.7</td>
<td>-.19,20</td>
<td>-0.03</td>
<td>0.56</td>
<td>0.750</td>
<td>-1.50,1.44</td>
</tr>
<tr>
<td>Ebel</td>
<td>13</td>
<td>-0.051</td>
<td>0.253</td>
<td>0.006</td>
<td>0.077</td>
<td>3.27</td>
<td>26.3</td>
<td>.097,41</td>
<td>0.21</td>
<td>0.11</td>
<td>0.339</td>
<td>-0.45,0.88</td>
</tr>
<tr>
<td>Jaeger</td>
<td>6</td>
<td>1.43</td>
<td>1.20</td>
<td>0.104</td>
<td>0.322</td>
<td>3.72</td>
<td>3.5</td>
<td>.57,1.83</td>
<td>1.20</td>
<td>0.10</td>
<td>0.323</td>
<td>0.57,1.84</td>
</tr>
<tr>
<td>Nedelsky</td>
<td>29</td>
<td>-1.24</td>
<td>-1.33</td>
<td>0.009</td>
<td>0.095</td>
<td>-14.02</td>
<td>177.8</td>
<td>-1.52,1.15</td>
<td>-1.23</td>
<td>1.47</td>
<td>1.213</td>
<td>-3.61,1.15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>92</td>
<td>-0.16</td>
<td>-0.02</td>
<td>0.002</td>
<td>0.047</td>
<td>-0.43</td>
<td>676.6</td>
<td>-11,07</td>
<td>0.19</td>
<td>1.33</td>
<td>1.153</td>
<td>-2.07,2.45</td>
</tr>
</tbody>
</table>

*All significant except Jaeger

### Table 3. Small Studies Analysis

<table>
<thead>
<tr>
<th>Size</th>
<th>k</th>
<th>t. (weighted)</th>
<th>Var(T.)</th>
<th>SE(T.)</th>
<th>Z</th>
<th>Q*</th>
<th>Ci</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small</strong></td>
<td>47</td>
<td>-0.034</td>
<td>0.013</td>
<td>0.114</td>
<td>-0.30</td>
<td>188.8</td>
<td>-2.6,19</td>
</tr>
<tr>
<td><strong>Large</strong></td>
<td>45</td>
<td>-0.019</td>
<td>0.003</td>
<td>0.055</td>
<td>-0.35</td>
<td>487.8</td>
<td>-13.09</td>
</tr>
</tbody>
</table>

**As defined by having less than 15 judges in the Angoff group**

95% confidence intervals (Ci) -
### Table 4. Number of Comparisons By Study, Size and Procedure

<table>
<thead>
<tr>
<th>Standard Setting Procedure</th>
<th>Ebel</th>
<th>Nedelsky</th>
<th>Jaeger</th>
<th>Contrasting</th>
<th>Borderline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
<td>Large</td>
<td>Total</td>
<td>Small</td>
<td>Large</td>
</tr>
<tr>
<td>Behuniak, Archambault, Gable</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baron, Rindone, Prowda</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross, Impara, Frary, Jaeger</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Brenan &amp; Lockwood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Livingston &amp; Zieky</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haplin &amp; Haplin</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Smith &amp; Smith</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poggio, Glasnapp, Eros</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Mills &amp; Melican</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>10</td>
<td>13</td>
<td>18</td>
<td>11</td>
</tr>
</tbody>
</table>
Figure 1. Effect Size Estimates (95% Confidence Interval)
Figure 2: Fixed Effect Size Estimates By Method (95% Confidence Interval)

- Borderline vs Angoff
- Contrast vs Angoff
- Ebel vs Angoff
- Jaeger vs Angoff
- Nedelsky vs Angoff
- TOTAL

Effect Size (Glass's Delta)
Figure 3. Random Effect Size Estimates By Method (95% Confidence Interval)

- Borderline vs Angoff
- Contrast vs Angoff
- Ebel vs Angoff
- Jaeger vs Angoff
- Nedelsky vs Angoff
- TOTAL

Effect Size (Glass's Delta)
I. DOCUMENT IDENTIFICATION:

Title: A META-Analytic Assessment of Empirical Differences in Standard Setting Procedures

Author(s): Brian D. Bontempo Casimer M. Marks George Karabatsos

Corporate Source: Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

________________________

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

________________________

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

________________________

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: ____________________________

Printed Name/Position/Title: Casimer Marks/Ass't Prodnanager

Organization/Address: National Council State Boards of Nursing

Telephone: 312-787-6555 FAX 312-787-3888

E-Mail Address: cmarks@nabsfn.org Date: 5/11/93

676 N. St Clair St. Chicago, IL 60611

(over)
March 20, 1998

Dear AERA Presenter,

Congratulations on being a presenter at AERA\(^1\). The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae.net.

Please sign the Reproduction Release Form on the back of this letter and include it with two copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the ERIC booth (424) or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 1998/ERIC Acquisitions

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

This year ERIC/AE is making a Searchable Conference Program available on the AERA web page (http://aera.net). Check it out!

Sincerely,

[Signature]

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

\(^1\)If you are an AERA chair or discussant, please save this form for future use.