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

BRIAN D. BONTEMPO
CASIMER M. MARKS
GEORGE KARABATSOS

UNIVERSITY OF CHICAGO

Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 1998.

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 Δ (Glass, McGaw, & Smith 1981).

$$\text{Glass's } \Delta = \frac{C_i - C_A}{S_A} \tag{1}$$

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 Δ was calculated by:

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

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:

$$Q_{BETWEEN} = \sum_{i=1}^{k} W_{i\bullet}(\Delta_{i\bullet} - \Delta_{\bullet\bullet})^2 \quad (4)$$

$$Q_{WITHIN} = \sum_{i=1}^{k} \sum_{j=1}^{mi} W_{ij}(\Delta_{ij} - \Delta_{i\bullet})^2 \quad (5)$$

$$Q = Q_{BETWEEN} + Q_{WITHIN}. \quad (6)$$

In the random effects model the variance component (between-studies variance) was calculated as follows:

$$\hat{\sigma}_{\Theta}^2 = \left[Q - (k-1)\right] \Big/ \sum_{i=1}^{k} w_i - \left[\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i\right] \quad (7)$$

Results

Fixed effects model

The overall standardized mean effect size difference was not significantly different from 0 ($\Delta_{..}$ = -0.02, z = -.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_B = 10.31) and the Jaeger (z_J = 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_n = -14.02) (See Figure 2). As expected, the variance was heterogeneous (Q = 676, df=91), both between ($Q_{BETWEEN}$ = 329, df = 4) and within (Q_{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 ($\Delta_{.small} = -0.034$, $\Delta_{.large} = -0.019$, $\underline{p} > .05$). Furthermore, there was no difference between these two groups ($Q_{BETWEEN}=.01$), but there was a significant amount of variability within groups ($Q_{within\ small}=189$, $Q_{within\ large}=488$).

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 ($Q_{BETWEEN} = 28$, $Q_{WITHIN} = 161$), 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 ($\Delta_{..} = 0.19$, $\underline{z} = 0.17$, $\underline{p} > .05$). The one factor model also revealed that for all but the Jaeger comparison, the mean effect size was not significantly different than 0 ($\underline{z}_{N,B,C} < 1.96$, $\underline{p} > .05$) (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

Andrew, B. J., & Hecht, J. T. (1976). A preliminary Investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

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

Baron, J. P., Rindone, D. A., & Prowda, P. (1981). Will the "Real" Proficiency Standard Please Stand Up? A paper presented at the annual meeting of the New England Educational Research Organization, MA.

Behuniak, P., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. Educational and Psychological Measurement, 42, 247-255.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion referenced tests. Review of Educational research, 56 (1), 137-172.

Beuk, C. H. (1984). A guide to criterion-referenced test construction. Baltimore, MD: Johns Hopkins University Press.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4(2), 219-240.

Cizek, G. J. (1996). An NCME instructional module on setting passing scores. Educational Measurement: Issues and Practice, 15 (2), 20-31.

Cooper, H., & Hedges, L. V. (1994). The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the national teacher examinations. Journal of Educational Measurement, 21(2), 113-129.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Glass, G., McGaw, B, & Smith M. (1981). Meta-Analysis in Social Research. Beverly Hills, CA: Sage.

Haplin, G., & Haplin, G. (1987). An analysis of the reliability and the validity of procedures for setting minimum competency standards. Educational and Psychological Measurement, 47, 977-983.

Haplin, G., Sigmon, G., & Haplin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hofstee, W. (1983). The case for compromise in educational selection and grading. In S.B. Anderson & J.S. Helmick (Eds.), On educational testing (pp. 109-127). San Francisco: Jossey-Bass.

Jaeger, R. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and applications. Educational and Evaluation Policy Analysis, 4, 461-475.

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

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17(3), 167-178.

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

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion referenced tests, Journal of Educational Measurement, 20(3), 283-292.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Plake, B. S. (1998). Setting Performance Standards for Professional Licensure and Certification. Applied Measurement in Education, 11(1), 65-80.

Poggio, J. P, Glasnapp, D. R., & Eros, D. S. (1981). An Empirical Investigation of the Angoff, Ebel, and Nedelsky Standard Setting Methods. A paper presented at the annual meeting of the American Educational Research Association, LA.

Shadish, W. R., Haddock, C. H. (1994) Combing Estimates of Effect Size. In Cooper, H., & Hedges, L. V. (Eds.). The Handbook of Research Synthesis (pp 261-281) New York: Russell Sage Foundation.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. Journal of Educational Measurement, 25(4), 259-274.

Van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19(4), 295-308.

Table 1. Description of Studies

Article	Control Method	Control Judges	Alternative Method	Alternative Judges	Test	# of ES estimates
Behuniak, Archambault, Gable	Angoff	Group of 6 judges	Nedelsky	Different parallel group of 8 judges	Connecticut School System Reading Test	1
Behuniak, Archambault, Gable	Angoff	Group of 7 judges	Nedelsky	Different parallel group of 6 judges	Connecticut School System Math Test	1
Baron, Rindone, Prowda	Angoff	Group of 10 judges	Nedelsky	Different parallel group of 11 judges	Connecticut School System Reading Test	1
Baron, Rindone, Prowda	Angoff	Group of 9 judges	Nedelsky	Different parallel group of 9 judges	Connecticut School System Math Test	1
Baron, Rindone, Prowda	Angoff	Group of 10 judges	Contrasting Groups	1 teacher at 200 different schools	Connecticut School System Reading Test	1
Baron, Rindone, Prowda	Angoff	Group of 9 judges	Contrasting Groups	1 teacher at 200 different schools	Connecticut School System Math Test	1
Baron, Rindone, Prowda	Angoff	Group of 10 judges	Borderline Group	1 teacher at 200 different schools	Connecticut School System Reading Test	1
Baron, Rindone, Prowda	Angoff	Group of 9 judges	Borderline Group	1 teacher at 200 different schools	Connecticut School System Math Test	1
Cross, Impara, Frary, Jaeger	Angoff	Group of 5 Angoff judges	Jaeger	Different parallel group of 5 Jaeger judges	NTE, Math	3
Cross, Impara, Frary, Jaeger	Angoff	Group of 5 Angoff judges	Jaeger	Different parallel group of 5 Jaeger judges	NTE, Elem Ed	3
Cross, Impara, Frary, Jaeger	Angoff	Same Group of 5 Angoff judges	Nedelsky	Different parallel group of 5 Nedelsky judges	NTE, Math	3
Cross, Impara, Frary, Jaeger	Angoff	Same Group of 5 Angoff judges	Nedelsky	Different parallel group of 5 Nedelsky judges	NTE, Elem Ed	3
Brenan & Lockwood	Angoff	Group of 5 judges	Nedelsky	Same group of 5 judges	Health Related	1
Mills	Angoff	Group of 16 judges	Borderline Group	Same group of 16 judges	Louisiana Grade 2 Basic Skills Test Math (6 forms)	6
Mills	Angoff	Group of 15 judges	Borderline Group	Same group of 15 judges	Louisiana Grade 2 Basic Skills Test Reading (6 forms)	6
Mills	Angoff	Group of 16 judges	Contrasting Groups	Same group of 16 judges	Louisiana Grade 2 Basic Skills Test Math (6 forms)	6
Mills	Angoff	Group of 15 judges	Contrasting Groups	Same group of 15 judges	Louisiana Grade 2 Basic Skills Test Reading (6 forms)	6
Livinston & Zieky	Angoff	4 different schools (5 judges at each)	Borderline Group	Same group of 5 judges at each school	Basic Skills Test Elementary Reading	4
Livinston & Zieky	Angoff	4 different schools (5 judges at each)	Borderline Group	Same group of 5 judges at each school	Basic Skills Test Elementary Math	4
Livinston & Zieky	Angoff	Same group of 5 judges at each school	Contrasting Groups	Same group of 5 judges at each school	Basic Skills Test Elementary Reading	4
Livinston & Zieky	Angoff	Same group of 5 judges at each school	Contrasting Groups	Same group of 5 judges at each school	Basic Skills Test Elementary Math	4

Haplin & Haplin	Angoff	3 non parallel groups of 5 judges	Ebel	Same 3X5 judges	Missouri College English Test	3
Haplin & Haplin	Angoff	Same 3X5 judges	Nedelsky	Same 3X5 judges	Missouri College English Test	3
Smith & Smith	Angoff	Group of 15 judges	Nedelsky	Different parallel group of 16 judges	NJ Reading Test	1
Poggio, Glasnapp, Eros	Angoff	10 separate groups of judges (N~30)	Nedelsky	Different parallel groups of judges	Kansas Competency Tests (10 different tests)	10
Poggio, Glasnapp, Eros	Angoff	10 separate groups of judges (N~30)	Ebel	Different parallel groups of judges	Kansas Competency Tests (10 different tests)	10
Mills & Melican	Angoff	Group of 10 judges	Nedelsky	Different parallel groups of 3 judges	NTE, Math	2
Mills & Melican	Angoff	Group of 13 judges	Nedelsky	Different parallel groups of 4 judges	NTE, Elem Ed	2

Table 2. Statistical Results Tables

		Fixed Effects							Random Effects			
Method	k	t. (unwieghted)	t. (weighted)	Var(T.)	SE(t.)	Z	Q[+]	Ci	t.* (weighted)	Var(T.*)	sd(T.*)	Ci
Borderline Groups	22	1.02	1.22	0.014	0.118	10.31	70.3	.99,1.45	1.22	0.80	0.893	-0.53,2.97
Contrasting Groups	22	-0.39	0.0056	0.010	0.100	0.06	69.7	-.19,.20	-0.03	0.56	0.750	-1.50,1.44
Ebel	13	-0.051	0.253	0.006	0.077	3.27	26.3	.097,.41	0.21	0.11	0.339	-0.45,0.88
Jaeger	6	1.43	1.20	0.104	0.322	3.72	3.5	.57,1.83	1.20	0.10	0.323	0.57,1.84
Nedelsky	29	-1.24	-1.33	0.009	0.095	-14.02	177.8	-1.52,-1.15	-1.23	1.47	1.213	-3.61,1.15
Total	92	-0.16	-0.02	0.002	0.047	-0.43	676.6	-.11,.07	0.19	1.33	1.153	-2.07,2.45

[+]All significant except Jaeger

Table 3. Small Studies Analysis

		Fixed Effects					
Size	k	t. (weighted)	Var(T.)	SE(t.)	Z	Q*	Ci
Small[++]	47	-0.034	0.013	0.114	-0.30	188.8	-2.6,.19
Large	45	-0.019	0.003	0.055	-0.35	487.8	-.13,.09

		Fixed Effects					
Method	k	t. (weighted)	Var(T.)	SE(t.)	Z	Q*	Ci
Borderline Groups	10	0.78	0.126	0.355	2.2000	63.1	.08,1.48
Contrasting Groups	10	-0.12	0.075	0.273	-0.4575	49.2	-.66,.42
Ebel	3	-0.85	0.186	0.431	-1.9723	2.9	-1.7,-.01
Jaeger	6	1.20	0.104	0.322	3.73	3.5	.57,1.83
Nedelsky	18	-0.37	0.026	0.162	-2.2673	42.3	-.69,-.05
Total (small)	47	-0.034	0.013	0.114	-0.30	188.8	-2.6,.19

[++]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

	Standard Setting Procedure														
	Ebel			Nedelsky			Jaeger			Contrasting			Borderline		
	Small	Large	Total	Small	Large	Total	Small	Large	Total	Small	Large	Total	Small	Large	Total
Behuniak, Archambault, Gable				2		2									
Baron, Rindone, Prowda				2		2				2		2	2		2
Cross, Impara, Frary, Jaeger				6		6	6		6						
Brenan & Lockwood				1		1									
Mills											12	12		12	12
Livinston & Zieky										8		8	8		8
Haplin & Haplin	3		3	3		3									
Smith & Smith					1	1									
Poggio, Glasnapp, Eros		10	10		10	10									
Mills & Melican				4		4									
Total	3	10	13	18	11	29	6		6	10	12	22	10	12	22

Figure 1.Effect Size Estimates (95% Confidence Interval)

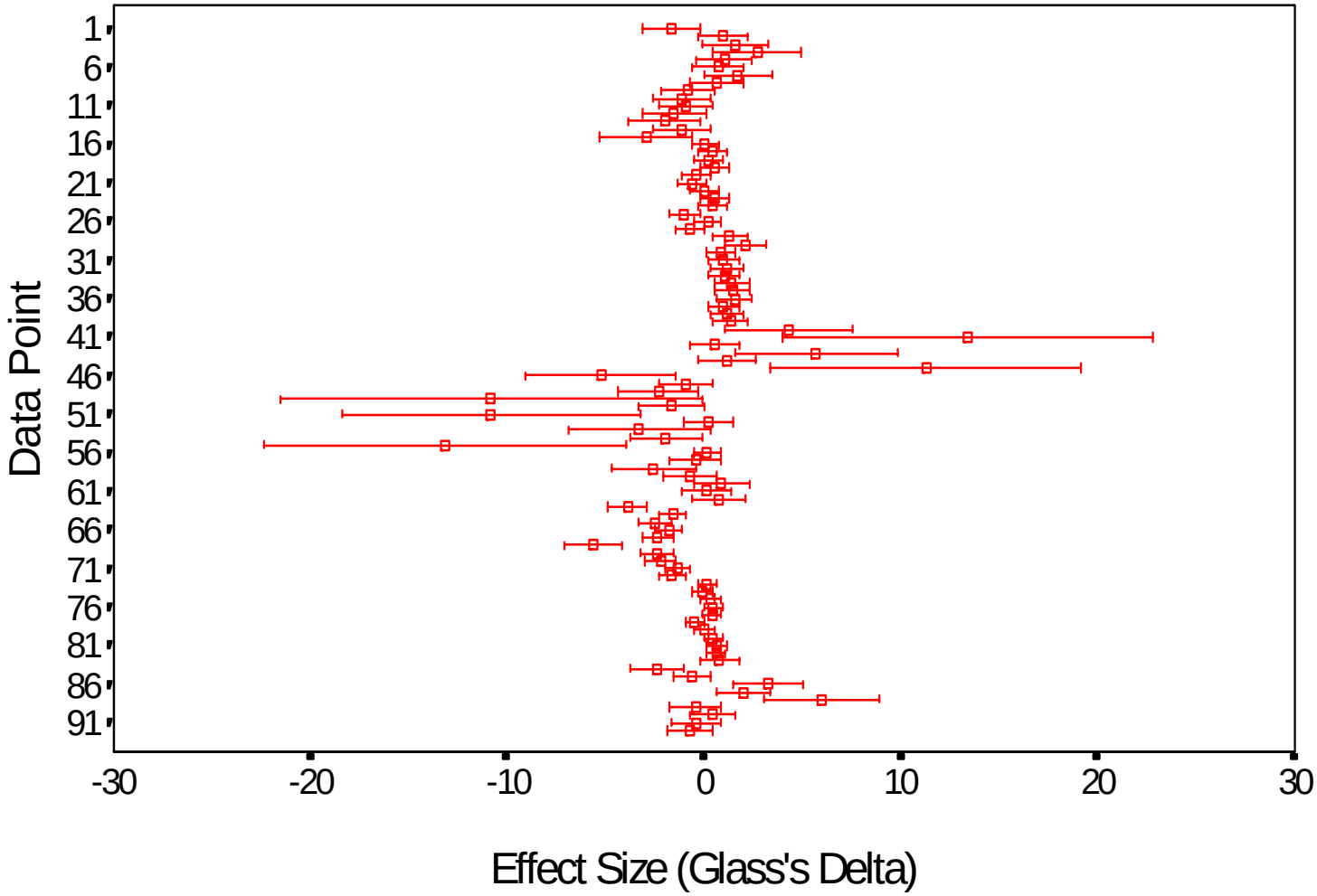

Figure 2.Fixed Effect Size Estimates By Method (95%Confidence Interval)

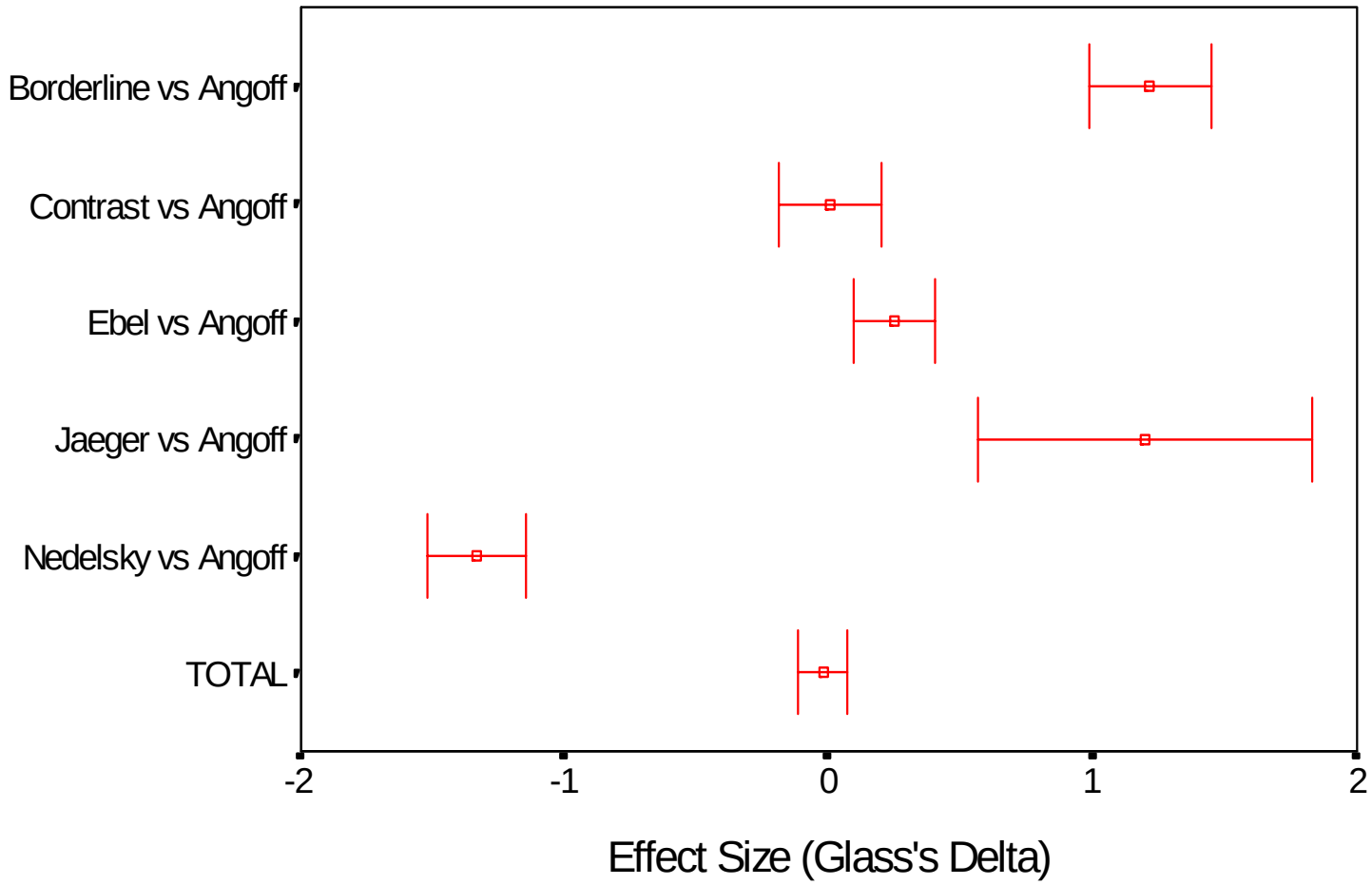

Figure 3. Random Effect Size Estimates By Method (95% Confidence Interval)