Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research

Michael C. Rodriguez, *University of Minnesota*

Multiple-choice items are a mainstay of achievement testing. The need to adequately cover the content domain to certify achievement proficiency by producing meaningful precise scores requires many high-quality items. More 3-option items can be administered than 4- or 5-option items per testing time while improving content coverage, without detrimental effects on psychometric quality of test scores. Researchers have endorsed 3-option items for over 80 years with empirical evidence—the results of which have been synthesized in an effort to unify this endorsement and encourage its adoption.

Keywords: multiple choice, item writing, item analysis, meta-analysis

> Item-writing is an art. It requires an uncommon combination of special abilities. It is mastered only through extensive and critically supervised practice. It demands, and tends to develop, high standards of quality and a sense of pride in craftsmanship. (Ebel, 1951, p. 185)

Item writing has been, is, and always will be *an art*. However, sophisticated, technically oriented, and computer-generative techniques have been developed to assist the item writer (see Baker, 1989; Bejar, 1993; Haladyna, 2004; Roid & Haladyna, 1982). Nonetheless, the science of item writing is still under development, as argued by each of the researchers whose work is reviewed below. Research on item writing has largely turned from empirical evaluation of the existing item format to evaluating the properties of new item types (Haladyna, 2004).

Measurement specialists have been writing about the construction of multiple-choice items since the early 1900s (e.g., Chapman & Toops, 1919; Wood, 1923; Yerkes, 1919), indeed since the initial large-scale use of the item type. Empirical work on item writing has been conducted since the 1920s (e.g., Ruch & Stoddard, 1925). However, even with this long tradition and attention to item writing, guidelines remain largely anecdotal—many item-writing rules may be nothing more than "item writing niceties" (Mehrens, personal communication, April 21, 1997). The lack of rigorous empirical study on item writing has troubled measurement specialists yet has not sparked enough interest to motivate the field to engage in extensive study. Virtually all of the authors of empirical studies investigating item format effects have expressed discontent with the amount of systematic study of item construction (Rodriguez, 1997).

One item-writing guideline has undergone a relatively substantial amount of empirical research, answering the question: How many options should a multiple-choice item have? The advice as stated by most measurement textbook authors is to write as many options as feasible (Haladyna & Downing, 1989a). After their review of the empirical literature, Haladyna and Downing (1989b) recommended a slight revision: "develop as many functional distractors as are feasible" (p. 59). This guideline has received more attention in the empirical literature on item writing than any other item-writing rule (Haladyna, Downing, & Rodriguez, 2002).

A reviewer pointed out the limited role of multiple-choice items in some contexts and the important role of performance assessment in classrooms. Performance assessments and authentic assessment activities have profound importance for communicating and demonstrating real-life activities in various fields—an important tool for teachers to employ for formative purposes and to add to the depth of important constructs in large-scale assessment. In areas such as Advanced Placement exams, performance tasks (including constructed-response items) play an important role in directing instruction and providing incentive for teachers to develop relevant classroom assessment activities. At the same time, the role of multiple-choice items is important in assessing broad ranges of knowledge and comprehension and, although more difficult, for assessing higher-order thinking skills as well (Haladyna, 1997).

In this study, I reviewed the existing empirical research as well as narrative and theoretical reviews regarding the optimal number of multiple-choice options. I then synthesized the empirical findings using meta-analytic techniques. The results have strong

Michael C. Rodriguez is Assistant Professor of Quantitative Methods in Education, College of Education and Human Development, University of Minnesota, 206 Burton Hall, 178 Pillsbury Drive SE, Minneapolis, MN, 55455; mcrdz@umn.edu. His areas of specialization include item writing, test design and evaluation, meta-analysis, and hierarchical linear modeling.

implications for validity-related arguments supporting the interpretation and use of test scores. As a reviewer pointed out, validity appears to be a central unifying theme in this line of research; the potential improvements in tests through the use of 3-option items enable the test developer and user to strengthen several aspects of validity-related arguments. Such improvements are uncovered through the process and results of this meta-analysis.

Background

Ebel (1951) reported in the first edition of *Educational Measurement* that hisreview of the literature uncovered five research articles on preparing multiple-choice (MC) items. More recently, Haladyna and Downing (1989a) reviewed 46 measurement instructional textbooks (dating back to 1935) to develop a taxonomy of MC item-writing rules. Of the 43 item-writing rules suggested, 10 addressed general item writing, 6 addressed stem development, and 20 addressed option development. This focus on options conveys a general concern about the importance of options to the MC item.

Prior Reviews

Haladyna and Downing (1989a) completed their review to create a taxonomy of rules, assessed the validity of their taxonomy of item-writing rules by reviewing research dating from 1926 (Haladyna & Downing, 1989b), and followed up this work with a review including more recent empirical research (Haladyna, Downing, & Rodriguez, 2002). Regarding the number of options, they found 22 related articles in 1989 and an additional 7 in 2002. They reported that changing the number of options affected the item difficulty but not discrimination. Where differences were found, they involved comparisons with the 2-option format. In addition, based on recommendations from 29 measurement textbook authors as reviewed in 1989, 55% supported this rule as stated above while 45% did not support this rule. However, these percentages changed in 2002 when 70% recommended using as many plausible distractors as possible whereas 26% refrained from making a recommendation and 4% were against the rule as stated, generally recommending the 4-option item.

In their review of theoretical and empirical studies, Haladyna and Downing (1989b) concluded the key was "not the *number* of distractors but the *quality* of distractors" (p. 59). In this work, they weighted effects across studies (with an unreported weighting function). Haladyna, Downing, and Rodriguez (2002) found more recent evidence did not support the standard use of four or five options: "three options are sufficient in most instances" (p. 318). They argued "the effort of developing that fourth option . . . is probably not worth it. If the fourth option is preferred, empirical research has established that it is very unlikely that item writers can write three distractors that have item response patterns consistent with the idea of plausibility" (p. 318). The current study improves upon the original 1989 synthesis by including additional research, a precision-weighted meta-analysis, estimates of standard errors, homogeneity test results, and evaluation of potential moderators.

Theoretical Research

As early as 1944, Lord derived a formula expressing change in reliability due to changes in the number of options per item. He argued that the reliability of the original test, a constant (determined by the percentage of correct responses), and the number of options per item in the original and revised tests, determined the revised reliability. Lord (1977) later argued that this formulation may not have been adequate across the ability range. Employing an IRT framework, he explained "the effect of decreasing the number of choices per item while lengthening the test proportionately is to increase the efficiency of the test for high-level examinees and to decrease its efficiency for low-level examinees" (p. 36). He also demonstrated the superior efficiency of the 3-option item test and suggested that the differential effect across the ability scale may be mediated by adjusting the difficulty of the test.

A mathematical proof was presented by Tversky (1964) to demonstrate how using three options per item maximizes discrimination, power, and information of a test, given a fixed total number of options for a test. "Whenever the amount of time spent on the test is proportional to its total number of alternatives, the use of three alternatives at each choice point will maximize the amount of information obtained per time unit" (p. 390).

Ebel (1969) also derived a predictive reliability formula that was a function of the number of items and the number of choices per item. This was offered as an alternative to the methods of Remmers and colleagues (Denny & Remmers, 1940; Remmers & Adkins, 1942; Remmers & Ewart, 1941; Remmers & House, 1941; Remmers, Karslake, & Gage, 1940; Remmers & Sageser, 1941) who predicted reliability as a function of the Spearman-Brown formula given number of options rather than number of items; however, this required an empirically determined reliability coefficient as a starting point. The predictive power of Ebel's formula was particularly good for tests of 100 items and suggested a trade-off between the number of items and the number of options.

Grier (1975) extended Ebel's formula to estimate optimal reliability by maximizing an approximation to KR-21. In doing so, he supported the theoretical advantages of 3-option items, showing their use maximized the expected reliability of a test when the number of items was increased to compensate for fewer alternatives per item. Grier (1976) later advanced Tversky's (1964) argument by generalizing the goal of optimizing the number of alternatives under a fixed total time. He demonstrated how allowing for "travel time," that is, the time it takes to read a question and the time it takes to consider each option (where Tversky considered no time between items and time as a linear function with the number of options) yielded Tversky's optimum result exactly. All of these methods demonstrated the advantage of 3-option items in comparison to others under the condition that the total number of options on a test remained constant.

Finally, Bruno and Dirkzwager (1995) investigated the optimal number of options for MC items through an information-theoretic perspective. Maximum information was obtained on test items with three options under the condition where each option had an equal probability of being answered (equally plausible) by an uninformed individual. In addition, this held for forced-choice items (where only one option could be selected) and items where the individual was allowed to

respond with personal probabilities for each alternative (accounting for partial knowledge).

This handful of theoretical approaches was consistent with the result: 3-option items maximize value. Value has been defined in terms of improved reliability (Grier, 1975, 1976; Lord, 1944) and information efficiency (Bruno & Dirkzwager, 1995; Lord, 1977; Tversky, 1964).

Empirical Research

Empirical research on the optimal number of options covered a wide range of conditions, subject areas, ages, and testing stakes. Most research was experimental, with forms of varying numbers of options randomly assigned to individuals. However, summarizing this research was difficult because studies investigated different numbers of options, typically some combination of 5-option to 2-option items. Studies will be described in terms of number of trials (independent experiments within a study).

An early evaluation of distractor effectiveness was conducted by Wakefield (1958) who examined 3,752 4-option and 3,294 5-option items from the California State Personnel Board exams for items with difficulties between .20 and .80. He found that 16% of 4-option items functioned like 4-option items where all options were functional; 3% of 5-option items functioned like 5-option items. Wakefield defined a distractor as functioning if more than 5% of the participants selected it.

Haladyna and Downing (1988) similarly examined a high-quality national standardized achievement test for physicians and found that 11 of the 200 5-option items had four functional distractors (49 items had one functional distractor and 13 had none). They defined a functional distractor as one that has (a) a significant negative point-biserial correlation with the total test score, (b) a negatively sloping item characteristic curve, and (c) a frequency of response greater than 5% for the total group. When they later examined a standardized medical education test, the reading and social studies ACT subtests, and a health science state certification exam with similar criteria, Haladyna and Downing (1993) found the number of effectively performing distractors per item to be about one; items with two or three effective distractors were very rare (1–8%). Also, the number of effective distractors was unrelated to item difficulty and positively related to item discrimination. They suggested that three options per item might be a natural limit for item writers in most circumstances. Beyond the empirical evidence suggesting that few 4-option and 5-option items actually have a complete set of effective distractors, a number of theoretical approaches and experimental and quasi-experimental studies have demonstrated the optimality of 3-option items.

Respondents preferred fewer options as well. When Owen and Froman (1987) completed their study of 3- versus 5-option forms, they asked the 114 participants to vote for their preferred form: 111 voted for the 3-option form, 3 had no preference, and none voted for the 5-option form.

The objectives of this study were to (a) formally synthesize the empirical results to estimate the effects of changing the number of options per multiple-choice item on item difficulty, item discrimination, test score reliability, and test validity; (b) explain potential variation in outcomes given study design characteristics, and (c) clarify the research history in this area and the role of multiple distractors. As described below, much of the research in this area has been conducted in K–12 settings, including both classroom-based tests and standardized instruments. Implications of this research are believed to be generalizable to K–12, postsecondary, and professional settings with both locally developed and standardized tests.

Method

Data Collection

The collection of data (studies) began with the reference lists provided by Haladyna and Downing (1989b) and Haladyna, Downing, and Rodriguez (2002). Computer searches of the *PsychLit, ERIC* (Educational Resources Information Center), and *Dissertation Abstracts International* databases were conducted. Studies regarding the number of MC options were obtained and the references from those studies were also reviewed. This process uncovered 48 studies.

Studies were screened for inclusion in the meta-analysis based on two criteria. First, the study must have evaluated the effect of varying the number of options in achievement or aptitude-type items. Second, the study must have reported the number of items in each format, the number of participants, and at least one of the following outcomes: item difficulties, item discriminations, test score reliabilities, or validity evidence for each format. All included studies employed experimental designs randomly assigning forms to participants or pre-post designs with the same participants. No other criteria were used to judge study quality.

Two studies were irretrievable (Charles, 1926; Parker & Somers, 1982). Twelve studies were eliminated because they did not report the statistics for any of the four outcomes included in this meta-analysis (Bruno & Dirkzwager, 1995; Ebel, 1969; Grier, 1975, 1976; Haladyna & Downing, 1988, 1993; Lord, 1944, 1977; Martín Andrés & Luna del Castillo, 1990; Tversky, 1964; Wakefield, 1958; Zimmerman & Humphreys, 1953) and were largely theoretical treatments of the topic. Four studies were eliminated because they did not use achievement- or aptitude-type items; they included attitude inventories (Remmers & Ewart, 1941; Remmers & Sageser, 1941) and auditory exams (Pollack & Ficks, 1954; Sumby, Chambliss, & Pollack, 1958). Finally, three were eliminated because they were based on data reported in other studies included in this synthesis (Cizek, Robinson, & O'Day, 1998; Remmers, Karslake, & Gage, 1940; Swanson, 1976). Twenty-seven studies met the selection criteria; nearly all have been reviewed previously in the literature reviews described above. Studies used in the final meta-analysis are starred (*) on the reference list.

Calculation of Effect Sizes for the Meta-Analysis

The following methods were used to calculate effect sizes and related statistics used in the meta-analysis, by outcome.

Item Difficulty. The first outcome synthesized in this study was the difference in mean item difficulty due to reduction in the number of options. This was reported as a difference in mean item difficulties for items in each format.

First, $\bar{p}$, the mean item difficulty for items with a given number of options was obtained as reported in primary studies or computed from available

information. Then the difference in mean item difficulties $T_{\text{diff}} = \bar{p}_2 - \bar{p}_1$ was computed, where $\bar{p}_1$ is the mean item difficulty for items in one format and $\bar{p}_2$ is the mean item difficulty of an alternate format with one fewer option. A positive difference indicates an increase in the item difficulty index from reducing the number of options, making the item easier.

One meta-analytic method for combining differences between p_1 and p_2 (as two independent proportions) is $D_i = p_{i1} - p_{i2}$, with a variance of $v_i = \frac{p_{i1}(1-p_{i1})}{n_{i1}} + \frac{p_{i2}(1-p_{i2})}{n_{i2}}$ (Shadish & Haddock, 1994). Using this same conceptualization where $\bar{p}$ is the mean item difficulty for items in a given format, the variance of the mean item difficulty can be estimated as $\frac{\bar{p}(1-\bar{p})}{n}$, where n is the number of items contributing to the mean item difficulty. The conditional variance of T_{diff} was calculated similarly, but with attention to the variance of a composite difference (the sum of the variances minus two times the covariance), $v_i = \frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2} - 2\rho_{12}\sigma_1\sigma_2$.

Item Discrimination. The mean item discrimination was reported in two forms, first as D (16 out of 29 effects), the difference in the proportion of correct responses for the upper-scoring 27% of students versus the lower-scoring 27%. Item discrimination was also reported as a point-biserial correlation between the item and the given format test score (13/29). Various item discrimination indices are highly correlated and discrepancies between them only occur for items at extreme difficulty ranges (Crocker & Algina, 1986). Englehart (1965) reported correlations between D and r_{pb} of .92 and .95 on two forms of a high school 60-item history exam. Oosterhof (1976) reported a correlation of .94 from a 50-item verbal analogy test (Differential Aptitude Test) of 1,000 high school students. In a Monte Carlo study varying the sample size, number of factors in an instrument, and item difficulty, Beuchert and Mendoza (1979) found differences among 10 indices of item discrimination "to be extremely small or nonexistent in situations tending to accentuate those differences" (p. 116). Both discrimination indices were treated as correlations in this synthesis.

Similar problems exist in computing effect sizes and variance estimates for the mean item discrimination as with the mean item difficulty. Are mean item discrimination indices like mean difficulties, since they are means across items, or like correlations, since many are based on point-biserial correlations? The distribution and standard error of the mean item discrimination are not known. At this time, the mean item discrimination will be treated like a correlation, since this is the metric in which effects will be interpreted.

All discrimination values were transformed using Fisher's normalizing and variance stabilizing Z-transformation. For any correlation r, $Z_r = {}^1\!/_2 \log_e\left[\frac{1+r}{1-r}\right]$, with variance $v = \frac{1}{(n-3)}$ where n is the within-study sample size (number of items in this case) for each form (Rosenthal, 1994). The difference between the Zs across formats was calculated as $T_{\text{disc}} = Z_2 - Z_1$, which is typically referred to as Cohen's q with a variance $v_i = \frac{1}{(n_1-3)} + \frac{1}{(n_2-3)}$ (Rosenthal, 1994). Positive differences indicate an increase in item discrimination from reducing the number of options. The conditional variance of T_{disc} was calculated as the variance of a composite, $v_i = \frac{1}{(n_1-3)} + \frac{1}{(n_2-3)} - 2\rho_{12}\sigma_1\sigma_2$.

Test Score Reliability. Although in a more modern conception of psychometric properties of tests reliability is an important component of validation, it is treated separately here because of the specific attention paid to reliability by primary study authors. Reliabilities were reported for each set of items in a given format. Most often, the reliability coefficient was estimated in one of three equivalent forms, KR-20 (23 out of 42 trials), Hoyt's analysis of variance reliability (10 trials), and coefficient, alpha (4 trials). Less frequently, reliabilities were reported as split-half reliability (5 trials). Because of the equivalence of these various estimates of reliability, and the small number of split-half coefficients, all reported coefficients were treated equivalently in this synthesis. Regarding this equivalence, the following argument was employed: Coefficient alpha is the average of all possible split-halves, whereas a split-half estimate of reliability is but one component of coefficient alpha (each possible split half could yield a higher or lower estimate than coefficient alpha). Since each form employed in computing a change in reliability for a given trial was split in the same way, the *difference* between split-half coefficients was assumed to be relatively consistent regardless of whether that particular split yielded a high or low estimate.

The synthesis of reliability coefficients was conducted employing the methods of Rodriguez and Maeda (2002), which rely on the sampling distribution of coefficient alpha as derived by Feldt (1965) and generalized by Hakstian and Whalen (1976). The basis of this analysis is a normalizing transformation of coefficient alpha $T_i = (1 - r_{\alpha i})^{1/3}$ and conditional variance estimate $\frac{18J_i(n_i-1)(1-r_{\alpha i})^{2/3}}{(J_i-1)(9n_i-11)^2}$. The meta-analysis was conducted on the difference in reliability $T_{\text{rel}} = T_2 - T_1$, with a conditional variance of this difference equal to the variance of a composite difference, $v_i = \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2$, to parallel the methods used for T_{diff} and T_{disc}.

Test Score Validity. Only two studies explicitly examined format effects on validity-related interpretations of test scores where the validity evidence was reported in the form of concurrent criterion-related validity coefficients. The results of these two studies were summarized narratively. At the same time, a reviewer pointed out that modern conceptions of validity embody a more holistic or unified concept of validity (Messick, 1989), particularly regarding Kane's (1992) argument-based approach that explicitly recognizes the wealth of information that can be brought to bear on specific interpretations of test scores as well as relevant assumptions. To this extent, reliability and item discrimination can provide useful information regarding the validity of the inferences we draw from test scores regarding content consistency and generalizability. To some degree, the results from analyses of item statistics and reliability provide validity evidence to support, in an argument-building approach (Kane, 1992), study-related inferences for the support of 3-option items.

Comparability of Effects

Because each author employed a different instrument and tested different numbers of options (e.g., some authors only tested changes from 5 to 3 options or 4 to 3 to 2 options) analyses were completed including sets of trials

testing each possible change in numbers of options from five to two (e.g., 5 to 4, 5 to 3, 5 to 2, 4 to 3, 4 to 2, and 3 to 2 options). This allowed the meta-analysis of all reported outcomes while maintaining comparability of effects.

Estimating Covariance

The variances of composite differences described above require an estimate of covariance (based on the correlation between corresponding forms varying the number of items). Sample correlations were obtained between mean statistics for corresponding forms from each of the three outcomes (item difficulty, item discrimination, and test score reliability). Forms with only one option difference typically resulted in very high correlations (based on individual correlations ranging between .84 and .97), whereas forms with two options difference resulted in lower correlations (ranging between .72 and .94) and forms with three options difference resulted in even lower correlations (ranging between .53 and .91). Since each correlation was based on a different number of trials, a median estimate across all parallel trials and the three psychometric statistics was used with a consistent result. The correlation employed for the reduction from 5 to 4, 4 to 3, and 3 to 2 options was .90; for 5 to 3 and 4 to 2 options was .80; and for 5 to 2 options was .70.

Data Analysis Procedures

The conceptual design of the meta-analysis stems from the following considerations. The universe to which I hope to generalize is a hypothetical collection of studies that could be conducted on the effects of altering the number of options. I am treating the studies as a sample from that universe, including both published and unpublished investigations. I assume that sampling error results from variation due to sampling of items within studies and sampling studies from that universe. These are the elements of a random effects model.

In order to account for study precision, the effects are weighted by a function of the variance estimates for each effect as described above (to account for sampling items within a trial) plus the random effects variance component, σ^2 (to account for sampling trials or studies from the universe of studies). The weight is $w_i = \frac{1}{\mathrm{var}(T_i)+\sigma^2}$. The mean weighted effect size then is $\bar{T}_\bullet = \frac{\sum T_i w_i}{\sum w_i}$. The standard error of the weighted mean effect size is $\mathrm{SE}(\bar{T}_\bullet) = \sqrt{\frac{1}{\sum w_i}}$. To allow for inferences regarding the estimated mean effects, I tested the hypothesis for homogeneity of effects, H_0: $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$. To do so, I calculated the Q-statistic, which is distributed as chi square with k-1 degrees of freedom: $Q = \sum_{i=1}^{k} w_i (T_i - \bar{T}_\bullet)^2$. See Shadish and Haddock (1994) for a more thorough explanation of these procedures.

A basic set of procedures for the meta-analysis includes the testing of homogeneity of effects across studies. When effects appear to be homogenous ($\sigma^2 = 0$), a fixed effects model is employed where the common parameter is estimated, uncertainty is quantified by calculating its standard error, and a significance test is performed on the estimate.

When effects appear heterogeneous (a significant Q), the studies can be described using a random effects or mixed model. Moderator analysis can be conducted to evaluate the possibility of explaining variance in effects given coded study characteristics. A significance test for potential moderators is based on the Q-statistic. In this study, the method employed for distractor deletion was used to examine effect heterogeneity, where the test of between-group differences is based on Q_{BETWEEN} ($Q_{\mathrm{TOTAL}} = Q_{\mathrm{BETWEEN}} + Q_{\mathrm{WITHIN}}$), which is exactly the weighted between-groups sum of squares obtained from a weighted ANOVA, distributed as chi square with k-1 degrees of freedom.

When outlier analysis and moderator analysis do not explain the heterogeneity of effects, the best one can do is to then estimate the random variance component, which may result in a different estimate of the average effect due to the inclusion of the random variance component in weighting effects and a larger standard error. When necessary, a maximum likelihood estimate of the random effects variance component was computed with HLM 5.0 (Raudenbush, Bryk, & Congdon, 2000) and added to the conditional variance estimate for computing effect weights. Under fixed effects models, this random effects variance component is set to zero. These were the basic procedures used to obtain the following results.

Remaining Methodological Issues

Ideally, to answer the question regarding the optimal number of options, item-level data from all previous empirical studies could be combined into a mega-study. Rarely do primary study authors report observed data or, in this case, item-level data. Meta-analysis and related techniques provide methods for the combination of summary statistics to achieve the approximation of the mega-study. The effects synthesized here were differences in mean effects based on the mean effects reported in the primary studies and in a couple of instances, mean effects computed from item-level data when reported.

In the Cooper, Hedges, and Olkin camp of meta-analysis (e.g., Cooper & Hedges, 1994; Hedges & Olkin, 1985), effects would be weighted by their conditional variance, a function of the sampling distribution of the effect, which maximizes the chance of correct inferences to a population parameter. Unfortunately, sampling distributions for mean item difficulty or item discrimination indices are unknown. This is what led to the computation of variances as a function of the composite variance of the differences in each effect, an approximation to the variance of the sampling distribution.

Results

The 27 studies yielded 56 independent trials (effects). Among the 27 studies were 25 journal articles and 2 technical reports. Hereinafter, the counts of effects will be in terms of number of trials (56 trials in total). Studies were published between 1925 and 1999 (Table 1). Studies in this area have been conducted in a variety of

Table 1. Publication Dates of Studies Included in the Synthesis

	1920–1939	1940–1959	1960–1979	1980s	1990s
Number of studies	2	4	6	8	7

Table 2. Summary of Study Characteristics for 56 Trials

Variable	Number of Trials
Age level of participants	
Professional	7
Postsecondary	23
Secondary grades (7–12)	14
Primary grades (1–6)	11
Other	1
Test type	
Standardized	26
Teacher-made	20
Researcher-made	10
Subject assignment to form	
Random	34
Existing groups	21
Matched on IQ	1
Distractor deletion method	
Random distractor	26
Ineffective distractor	19
Most attractive distractor	3
Added distractors	1
Not reported	7
Subject area	
Language arts	19
Social science	13
Science	6
Math	5
Mixed subjects	3
Others	10

contexts under a variety of conditions. Several study characteristics were coded for each of the 56 trials (Table 2).

The average number of items per form was 43 ($SD = 25$), ranging from 12 to 144 item forms, with a total of 2,406 items across all trials. The total number of participants across all trials was 12,591, ranging from 22 to 1,657 within trials ($M = 243$, $SD = 405$). The 56 trials included the subject areas of language arts (19 trials), social sciences (13), science (6), math (5), mixed subjects (3), and 10 trials in a variety of other areas including exams in musical acoustics, Air Force instruction, entry-level police officer selection, and health professions.

One potentially important study characteristic is the method used to delete options across forms. In 26 trials, distractors were randomly deleted to create items with varying numbers of options. In 19 trials, the most ineffective distractors were deleted, and the most attractive distractor was deleted in 3 trials. Only Trevisan, Sax, and Michael (1994) added (rather than deleted) options to 2-option items, using the Haladyna and Downing (1989a) taxonomy of rules as a guideline.

Table 3. Mean Effects, Standard Errors, and Effect Homogeneity Tests for Change in Item Difficulty

	Effects Summary			Homogeneity Tests	
Change in Number of Options	Difference in $\bar{T}_\bullet$	SE(*T*)	*N*	*Q*	Chi-square *p*-value
5 to 4*	.021	.008	22	13.0	.91
5 to 3*	.070	.009	30	35.0	.21
5 to 2*	.231	.019	9	13.4	.10
4 to 3*	.044	.005	36	57.9	.01
4 to 2*	.188	.014	10	11.3	.26
3 to 2*	.099	.009	12	14.5	.21

*Effect is significantly different from zero at $p < .05$.

Item Difficulty

All reductions in the number of options resulted in significant changes in mean item difficulty (Table 3, Figure 1). Reducing 4 options to 3 yielded a small change in difficulty (.04), while those reductions going to 2 options (5 to 2, 4 to 2, and 3 to 2) yielded the largest differences, all increasing the difficulty index (making the items easier). Nearly all effects were homogeneous across studies, where the random effects variance components were estimated at zero (see the results of the *Q*-test of homogeneity in Table 3). Since only one effect was heterogeneous (reducing 4-option items to 3-option items), a random effects model was not employed. These were fixed effects results.

Item Discrimination

Nearly all reductions in the number of options resulted in significant changes in discrimination. As can be seen in Figure 2 and Table 4, only one of the 95% confidence intervals included zero (reducing 5-option items to 3-option items). In most cases, reducing the number of options reduced item discrimination, except when reducing the number of options from 4 to 3, where a slight increase in item discrimination was observed. The largest changes include those involving reductions in the number of options to 2. The results for all effects were homogeneous across studies; random effects variance components were zero, echoed by the *Q*-test of homogeneity. These were fixed effects results.

Test Score Reliability

Most studies included analysis of test score reliability. Differences in reliability varied significantly across trials.

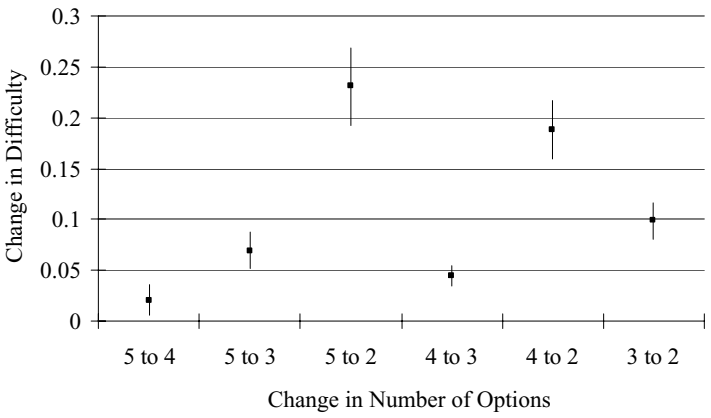

FIGURE 1. Difference in item difficulty with 95% confidence intervals for reductions in number of options.

Table 4. Mean Effects, Standard Errors, and Effect Homogeneity Tests for Change in Item Discrimination

	Effects Summary			Homogeneity Tests	
Change in Number of Options	Difference in $\bar{T}_\bullet$	SE(*T*)	*N*	*Q*	Chi-square *p*-value
5 to 4*	−.040	.018	20	6.5	.99
5 to 3	−.004	.025	20	3.4	.99
5 to 2*	−.111	.055	6	0.3	.99
4 to 3*	.031	.014	30	10.0	.99
4 to 2*	−.093	.035	8	1.5	.98
3 to 2*	−.089	.025	8	2.5	.93

* Effect is significantly different from zero at $p < .05$.

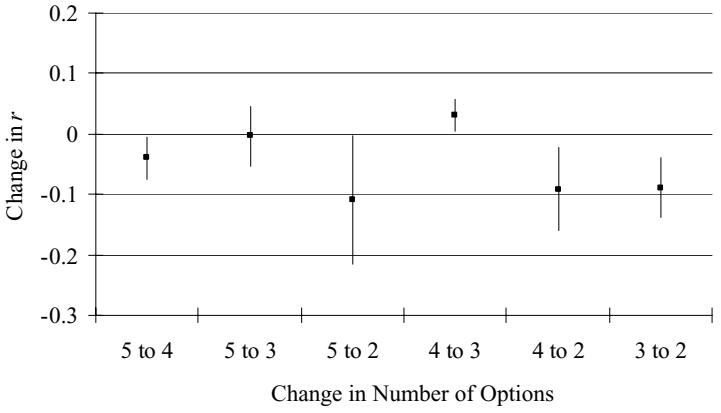

FIGURE 2. Difference in item discrimination with 95% confidence intervals for reductions in number of options.

Random effects variance components were significant (ranging from .0005 to .002). In each case except one (reducing the number of options from 5 to 3), changes in test score reliability were significant. In most cases, reduction in the numbers of options decreased reliability, except when reducing the number of options from 4 to 3, where a slight increase in reliability was observed (.02). As can be seen in Figure 3 and Table 5, the largest decrease in reliability occurred from 5 to 2 options (.11) and 4 to 2 options (.09). These are random effects results.

Validity Evidence

Two studies provided traditional test score validity-related evidence in the form of criterion-related validity correlations with criterion instruments. Owen and Froman (1987) altered final exams for 114 undergraduates in an educational psychology course. Two 100-item parallel forms were administered 10 days apart (parallel in terms of content and item difficulties). Form A consisted of two forms, one which was administered to half of the students where the first 50 items had three options and the second 50 items had five options, with a reversed design on the second Form A for the other half of the students. The two distractors with the least discriminating power were deleted to form the 3-option items. Scores from form A were correlated with form B (the second final exam with all 5-option items). Correlations between form A and form B were .75 with the 3-option test and .73 with the 5-option test. Owen and Froman also found that the 3-option form took 17% less time and suggested that an additional eight or nine items (keeping testing time constant) would improve both content-related validity and reliability. These correlations were presented as validity-related evidence, but could also be conceived of as evidence of parallel forms reliability (perhaps even test-retest reliability evidence). Nonetheless, the 3-option form correlated just as well with the parallel form as did the 5-option form demonstrating consistent agreement regardless of format.

The second study involved the 45-item Washington Pre-College Admissions Test Battery (University of Washington) that was administered to 282 high school students to evaluate the effects of reducing the 5-option items to 4- and 3-option items (Trevisan, Sax, & Michael, 1991). The least discriminating distractors were sequentially removed to create the alternate forms that were randomly assigned to students (approximately 90 students per form). Resulting scores were correlated with self-reported GPA (not a strong criterion), which resulted in correlations of .42 for the 5-option form, .47 for the 4-option form, and .24 for the 3-option form. These correlations were not statistically significantly different, based on 95% confidence intervals from conversion to Fisher's *Z*.

Both studies resulted in a statistically negligible change in criterion-related validity evidence when reducing the number of options from five to three and from five to four to three.

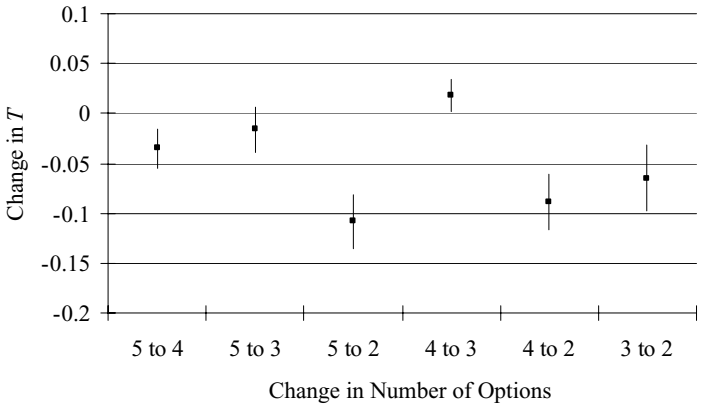

FIGURE 3. Difference in test score reliability (in *T* metric) with 95% confidence intervals for reductions in the number of options.

Table 5. Mean Effects, Standard Errors, and Effect Homogeneity Tests for Change in Score Reliability

Change in Number of Options	Effects Summary: Difference in $\bar{T}_\bullet$	Effects Summary: SE(T)	Effects Summary: N	Homogeneity Tests: Q	Homogeneity Tests: Chi-square p-value
5 to 4*	−.035	.010	25	24.7	.42
5 to 3	−.016	.012	29	28.1	.46
5 to 2*	−.108	.014	12	11.1	.44
4 to 3*	.019	.008	38	37.7	.44
4 to 2*	−.089	.014	13	13.5	.34
3 to 2*	−.065	.017	15	14.1	.44

*Effect is significantly different from zero at $p < .05$.

Both studies involved static forms—the number of options was varied but the number of items was held constant. This small sample of validity-related evidence is not overwhelmingly supportive of an argument for psychometric improvement (or nondestructive impact) due to a reduction in the number of options. Unfortunately, the argument made by so many researchers regarding the increase in validity (both content-related and criterion-related validity evidence, and improvement in reliability) due to the opportunity to increase the number of items because of the reduction in options has not been empirically tested.

Option Deletion Method

Finally, the possible role of the distractor deletion method was investigated. Recall that four methods were employed in this set of studies (random deletion [$n = 26$], deletion of ineffective distractors [$n = 19$], deletion of the most effective distractor [$n = 3$], and adding distractors [$n = 1$]). This analysis compared outcomes vis-à-vis random deletion of distractors versus deletion of ineffective distractors. Given the arguments made by prior researchers, if a reduction in the number of options deteriorates the quality of items, it should be a function of poor distractors; random deletion of distractors should be more destructive (because of the potential of deleting effective distractors) whereas deleting ineffective distractors should not be destructive to item quality. There was no relation between option deletion method and change in item difficulty or item discrimination. Based on the strong evidence of effect homogeneity suggesting little to no variation in change in difficulty or discrimination across studies, the result of no relation to deletion method was expected.

Regarding changes in test score reliability, the method of distractor deletion explained significant variation in results in two cases. There was an average reduction in reliability of .016 when reducing the number of options from 5 to 3 including all four methods of deletion (not significantly different than zero). However, this reduction was a function of the distractor deletion method: for the trials deleting ineffective distractors, no change in reliability was observed (.006); whereas trials randomly deleting distractors yielded an average reduction in reliability of .059 (a statistically significant difference: $Q_{\text{BETWEEN}} = 108$, much larger than the $\chi^2[1] = 3.8$ critical value for $p < .05$). Similarly, there was an average reduction in the reliability of .089 when reducing the number of options from 4 to 2; however, this reduction was less than half that for trials deleting ineffective distractors (.040) as compared to trials deleting distractors randomly ($Q_{\text{BETWEEN}} = 37$). In both cases, random distractor deletion was significantly more damaging than deleting ineffective distractors.

Possible Publication Bias

There is no certain way to assess the presence or impact of possible publication bias. Publication bias is a result of a tendency for published articles to be more likely to report significant effects and for studies not finding significant effects to be less likely to be published.

Two of the 27 studies were unpublished and five studies reported nonsignificant findings on each outcome studied, while many reported mixed results. In fact, in this area, nonsignificant findings are just as important as significant findings, as all findings inform the practice of item writing. There was no compelling evidence suggesting the presence of publication bias.

Discussion

Based on the evidence synthesized in this meta-analysis, the item-writing rule can be revised: Three options are optimal for MC items in most settings. Moving from 5-option items to 4-option items reduces item difficulty by .02, reduces item discrimination by .04, and reduces reliability by .035 on average. Moving from 5- to 3-option items reduces item difficulty by .07, does not affect item discrimination, and does not affect reliability on average. However, when eliminating random distractors to create 3-option items, reliability drops .06 on average, with no change if ineffective distractors are deleted. Moving from 4- to 3-option items reduces item difficulty by .04, increases item discrimination by .03, and increases reliability slightly by .02. More notably, when moving from 5- or 4- to 2-option items, items become significantly more easy (by .23 and .19), less discriminating (by .11 and .09), and scores are less reliable (by .11 and .09). When moving from 4- to 2-option items by deleting random distractors, the reduction in reliability is more than twice that compared to deleting ineffective distractors.

The findings do not necessarily conflict with the Haladyna, Downing, and Rodriguez (2002) recommendation to write as many plausible distractors as possible, particularly in light of the more negative results when randomly deleting distractors as compared to deleting ineffective distractors (in some cases). However, based on the results of this synthesis, the rule could be more direct and promote the use of 3-option items.

The vast majority of authors who studied this rule recommended using 3-option items. Others who limited their investigation to 4-option versus 5-option items recommended using 4-option items (Hodson, 1984; Ramos & Stern, 1973). Only one research team recommended against using three options per item. Budescu and Nevo

(1985) investigated the assumption of proportionality, which suggests that the total testing time is proportional to the number of items and the number of options per item. They found a strong negative relationship between rate of performance and the number of options for tests of fixed number of items; the assumption of proportionality did not hold. They argued, based on a generalized form of the proportionality assumption (introduced by Grier, 1976), that testing time is a function of the number of items, the number of options, and a function of the item's complexity, making change in response time not a simple function of the number of options. They did not offer an answer to the optimal number of options but argued that three would generally be insufficient.

The Role of Many Distractors

It has been suggested that we use as many plausible distractors as feasible (Haladyna, Downing, & Rodriguez, 2002). This is based on a fair review of the literature. I would support this advice by contributing the concern that in most cases, only three are feasible. Based on this synthesis, MC items should consist of three options, one correct option and two plausible distractors. Using more options does little to improve item and test score statistics and typically results in implausible distractors. The role of distractor deletion method makes the argument stronger.

Beyond the evidence, practical arguments continue to be persuasive.

1. Less time is needed to prepare two plausible distractors than three or four distractors.
2. More 3-option items can be administered per unit of time than 4- or 5-option items, potentially improving content coverage.
3. The inclusion of additional high-quality items per unit of time should improve test score reliability, providing additional validity-related evidence regarding the consistency of scores and score meaningfulness and usability.
4. More options result in exposing additional aspects of the domain to students, possibly increasing the provision of context clues to other questions (particularly if the additional distractors are plausible).

The threat of guessing and having a greater chance of a correct guess with 3-option items than with 4- or 5-option items has also not prevailed. Examinees are unlikely to engage in blind guessing, but rather educated guessing where the least plausible distractors are eliminated, essentially reducing the 4- or 5-option item to a 3- or 2-option item (Costin, 1972, 1976; Kolstad, Briggs, & Kolstad, 1985). Kolstad, Briggs, and Kolstad recommended using "no more choices than required for the effective suppression of guessing" (p. 431). They argued that the quality of the distractors guards against awarding undeserved credit, not the number of distractors. However, a reviewer pointed out that this argument likely works well for unspeeded tests, whereas we can expect more frequent blind guessing from lower-ability students in highly speeded tests when they simply run out of time. This would have a detrimental effect on validity.

In some contexts, distractors can provide diagnostic information where distractors are coded to map to common misconceptions. In such cases, more distractors may be needed. However, a constant tension remains between obtaining misconception-related diagnostic information from individual items with more options versus obtaining reliable content-related diagnostic information from smaller sets of items measuring a particular strand or content objective.

Future Directions

Item analysis is a critical step in test development. Item analysis is useful in judging the worth or quality of items and the test. It helps in subsequent revisions of tests and in building item databases for future tests. Classical item-analysis data should be used carefully (Mehrens & Lehmann, 1991), as item-analysis data are tentative and sample specific. On this issue, a reviewer rightly pointed out that, from an IRT perspective, the elimination of distractors makes items easier is irrelevant. Of course, the ability to include more items remains a strong benefit. In a standard setting context, item difficulty plays a significant role in popular methods, including Angoff-related methods and item-mapping or bookmark-related procedures. Finally, in a classroom context, grading standards may need adjustment if tests become easier due to the use of fewer options; however, this could be tempered by including more items–the real benefit here is the potential improvement in content coverage. In addition, item difficulty is a function of more than just the number of distractors and significant shifts in item (test) difficulty by using 3-option items are not apparent.

Validity evidence should be gathered to ensure the quality of test item development, including the use of item-writing principles (Downing & Haladyna, 1997). To secure validity evidence in this area, item-writing guidelines could serve as a checklist (e.g., Haladyna, Downing, & Rodriguez, 2002), providing documentation regarding the use of particular item formats. For example, test developers should provide a rationale for the number of options used in MC items. This is rarely done explicitly and is unfortunately sometimes legislated. And of course, improvements in content-related validity evidence due to the inclusion of additional items suggested by so many researchers should be documented.

Evidence from this meta-analysis and one primary study (Budescu & Nevo, 1985) suggests that in some cases, the impact of changing the number of options per item depends on the method used to delete options. Research on such characteristics might allow us to more clearly articulate Haladyna, Downing, and Rodriguez' (2002) recommendation to use as many *plausible* distractors as possible. Clearly, work needs to be done on the role of more effective plausible distractors. Until then, three appears to be the optimal number of options.

Information that increases our understanding of multiple-choice items and tests will improve our ability to measure student achievement and other constructs. Improved information will lead to improved item writing, improved test design, better measures of achievement and skill level, and more appropriate score interpretation and decision making.

References

*Note: * indicates those references that were included in the meta-analysis.*

*Asmus, E. J., Jr. (1981). The effect of altering the number of choices per item on test statistics: Is three better than five?

Bulletin of the Council for Research in Music Education, 54, 948–950.

Baker, F. B. (1989). Computer technology in test construction and processing. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 409–428). New York: American Council on Education and Macmillan.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Rederiksen, R. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–357). Hillsdale, NJ: Erlbaum.

Beuchert, A. K., & Mendoza, J. L. (1979). A Monte Carlo comparison of ten item discrimination indices. *Journal of Educational Measurement, 16,* 109–118.

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55,* 959–966.

*Budescu, D. V., & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement, 22,* 183–196.

*Catts, R. (1978). *How many options should a multiple-choice question have?* (At-a-glance research report.) Sydney, Australia: New South Wales Department of Education.

Charles, J. W. (1926). *A comparison of five types of objective tests in elementary psychology*. Unpublished doctoral dissertation, State University of Iowa, Iowa City.

Chapman, C. J., & Toops, H. A. (1919). A written trade test: Multiple choice method. *Journal of Applied Psychology, 3,* 358–365.

*Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement, 54,* 861–872.

Cizek, G. J., Robinson, K. L., & O'Day, D. M. (1998). Nonfunctioning options: A closer look. *Educational and Psychological Measurement, 58,* 605–611.

Cooper, H., & Hedges, L. V. (Eds). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

*Costin, F. (1970). The optimal number of alternatives in multiple choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement, 30,* 353–358.

*Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement, 32,* 1035–1038.

*Costin, F. (1976). Difficulty and homogeneity of three-choice versus four-choice objective test items when matched for content of stem. *Teaching of Psychology, 3,* 144–145.

*Crehan, K. D., Haladyna, T. M., & Brewer (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement, 53,* 241–247.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

*Denny, H. R., & Remmers, H. H. (1940). Reliability of multiple-choice as a function of the Spearman-Brown prophecy formula, II. *Journal of Educational Psychology, 31,* 699–704.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61–82.

*Duncan, R. E. (1983). An appropriate number of multiple-choice item alternatives: A difference of opinion. *Measurement and Evaluation in Guidance, 15,* 283–292.

Ebel, R. L. (1951). Writing the test item. In E. F. Linquist (Ed.), *Educational measurement* (pp. 185–249). Washington, DC: American Council on Education.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement, 29,* 565–570.

Englehart, M. D. (1965). A comparison of several item discrimination indices. *Journal of Educational Measurement, 2,* 69–76.

Feldt, R. L. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357–370.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12,* 109–112.

Grier, B. (1976). The optimal number of alternatives at a choice point with travel time considered. *Journal of Mathematical Psychology, 14,* 91–97.

Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219–231.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn and Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Erlbaum

Haladyna, T. M., & Downing, S. M. (1988, April). *Functional distractors: Implications for test-item writing and test design*. Paper presented at the annual meeting of the AERA, New Orleans, LA.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 2,* 37–50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 2,* 51–78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice item? *Educational and Psychological Measurement, 53,* 999–1010.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309–334.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

*Hodson, D. (1984). Some effects of changes in question structure and sequence on performance in a multiple choice chemistry test. *Research in Science & Technological Education, 2,* 177–185.

*Hogben, D. (1973). The reliability, discrimination and difficulty of word-knowledge tests employing multiple choice items containing three, four or five alternatives. *Australian Journal of Education, 17,* 63–68.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

*Kolstad, R. K., Briggs, L. D., & Kolstad, R. A. (1985). Multiple-choice classroom achievement tests: Performance on items with five vs. three choices. *College Student Journal, 19,* 427–431.

*Kolstad, R. K., Kolstad, R. A., & Wagner, M. J. (1986). Performance on 3-choice versus 5-choice MC items that measure different skills. *Educational Research Quarterly, 10,* 4–8.

*Landrum, R. E., Cashin, J. R., & Theis, K. S., (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement, 53,* 771–778.

Lord, F. M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. *Journal of Educational Psychology, 35,* 175–180.

Lord, F. M. (1977). Optimal number of choices per item—a comparison of four approaches. *Journal of Educational Measurement, 14,* 33–38.

Martín Andrés, A., & Luna del Castillo, J. D. (1990). Multiple choice tests: Power, length and optimal number of choices per item. *British Journal of Mathematical and Statistical Psychology, 43,* 57–71.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Orlando, FL: Harcourt Brace Jovanovich.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 3–104). New York: American Council on Education.

Oosterhof, A. C. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement, 13,* 145–150.

*Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement, 47,* 513–522.

Parker, C. S., & Somers, J. E. (1982). *A comparison of the difficulty and reliability of Type K and one-best-response test items*. Paper presented at the meeting of the Iowa Evaluation and Research Association, Des Moines, Iowa.

Pollack, I., & Ficks, L. (1954). Information of elementary multidimensional auditory

displays. *Journal of the Acoustical Society of America, 26,* 155–158.

*Ramos, R. A., & Stern, J. (1973). Item behavior associated with changes in the number of alternatives in multiple choice items. *Journal of Educational Measurement, 10,* 305–310.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2000). *HLM* (Version 5.0) [Computer software]. Chicago, IL: Scientific Software International.

*Remmers, H. H., & Adkins, R. M. (1942). Reliability of multiple-choice measuring instruments, as a function of the Spearman-Brown prophecy formula, VI. *Journal of Educational Psychology, 33,* 385–390.

Remmers, H. H., & Ewart, E. (1941). Reliability of multiple-choice measuring instruments, as a function of the Spearman-Brown prophecy formula, III. *Journal of Educational Psychology, 32,* 61–66.

*Remmers, H. H., & House, J. M. (1941). Reliability of multiple-choice measuring instruments, as a function of the Spearman-Brown prophecy formula, IV. *Journal of Educational Psychology, 32,* 372–376.

Remmers, H. H., Karslake, R., & Gage, N. L. (1940). Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, I. *Journal of Educational Psychology, 31,* 583–590.

Remmers, H. H., & Sageser, H. W. (1941). Reliability of multiple-choice measuring instruments, as a function of the Spearman-Brown prophecy formula, V. *Journal of Educational Psychology, 32,* 445–451.

Rodriguez, M. C. (1997). *The art & science of item writing: A meta-analysis of multiple-choice item format effects*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rodriguez, M. C., & Maeda, Y. (2002). *Statistical issues of reliability generalization and an application to achievement data*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

*Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency results. *Educational and Psychological Measurement, 59,* 234–247.

Roid, H. R., & Haladyna, T. M. (1982). *A technology for test-item writing*. Orlando, FL: Academic Press.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

*Ruch, G. M., & Charles, J. W. (1928). A comparison of five types of objective tests in elementary psychology. *Journal of Applied Psychology, 12,* 398–403.

*Ruch, G. M., & Stoddard, G. D. (1925). Comparative reliabilities of objective examinations. *Journal of Educational Psychology, 16,* 89–103.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

*Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology, 47,* 829–835.

*Straton, R. G., & Catts, R. M. (1980). A comparison of two, three, and four-choice item tests given a fixed total number of choices. *Educational and Psychological Measurement, 40,* 357–365.

Sumby, W. H., Chambliss, D., & Pollack, I. (1958). Information transmission with elementary auditory displays. *Journal of the Acoustical Society of America, 30,* 425–429.

Swanson, R. G. (1976). *Multiple choice tests: How many alternatives?* Maxwell AFB, AL: Academic Instructor School.

*Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement, 51,* 829–837.

*Trevisan, M. S., Sax, G., & Michael, W. B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement, 54,* 86–91.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1,* 386–391.

Wakefield, J. A. (1958). Does the fifth choice strengthen a test item? *Public Personnel Review, 19,* 44–48.

*Williams, B. J., & Ebel, R. L. (1957). The effects of varying the number of alternatives per item on multiple-choice vocabulary test items. In E.M. Huddleston (Ed.), *The fourteenth yearbook of the National Council on Measurements Used in Education* (pp. 63–65). New York: NCMUE.

Wood, B. D. (1923). *Measurement in higher education*. New York: Harcourt, Brace, and World.

Yerkes, R. M. (1919). Report of the Psychology Committee of the National Research Council. *Psychological Review, 26,* 83–149.

Zimmerman, W. S., & Humphreys, L. G. (1953). *Item reliability as a function of the omission of misleads*. Paper presented at the annual meeting of the American Psychological Association.