The Art & Science of Item-Writing:
A Meta-Analysis of Multiple-Choice Item Format Effects

Michael C. Rodriguez
Michigan State University

Paper presented at the annual meeting of the
American Educational Research Association, Chicago, IL, April 1997.
Revised August 1997

Acknowledgments

I would like to thank Betsy J. Becker, Pam Jakwerth, and Bill Mehrens for their guidance and support on this study. I also thank Kyle Fahrbach for his technical support regarding the derivation of the conditional variance estimation formula for mean item difficulties. Chris Chiu, Christine Demars, Kyle Fahrbach, and Shelly Naud were very helpful with coding studies. ***SynRG***, the Synthesis Research Group at Michigan State University, was very helpful with suggestions for focusing and streamlining the meta-analysis and responding to methodological issues I faced throughout the study. Finally, I must acknowledge the support and ongoing feedback regarding this project from Susan Phillips, Ken Frank, and Irv Lehmann. Several authors were also very helpful in sending me documents and hard-to-retrieve articles, especially Thomas Haladyna (Arizona State University West), and also Donald Hogben (The Flinders University of South Australia), Cynthia Schmeiser (American College Testing), and Richard Smith (Applications of Rasch Measurement).

Table of Contents

Introduction & Background

> 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.
>
> Item writing is essentially creative. Each item as it is being written presents new problems and new opportunities. Just as there can be no set formulas for producing a good story or a good painting, so there can be no set of rules that will *guarantee* the production of good test items. Principles can be established and suggestions offered, but it is the item writer's judgment in the application (and occasional disregard) of these principles and suggestions that determines whether good items or mediocre ones will be produced. (Ebel, 1951, p. 185)

Item writing has been, is, and always will be *an art*. Although sophisticated, technically oriented, and computer generative techniques are being developed to assist the item writer (see Baker, 1989; Bejar, 1993; Haladyna, 1994), there remain considerations that only human review can address. These include simultaneous consideration of many item-writing rules and concepts, word choice, social definitions of words and concepts, and the prevailing consensus among measurement and testing specialists who advise us about item construction.

Item writing is also serious business. The prevalence of multiple-choice items in large-scale tests (standardized tests, state testing programs, employment and vocational tests, certification exams, etc.) requires that items be written well. Item and test analyses provide the tools for empirical review of item quality.

Measurement specialists have been writing about the construction of items since the early 1900s. Empirical work on item writing has been conducted since the 1920s. However, even with this long tradition and attention to item writing, it has remained anecdotal and advice oriented--some even refer to item-writing rules as “item writing niceties” (Mehrens, personal

communication, April 21, 1997). The lack of serious empirical study on item writing has troubled measurement specialists (virtually all of the authors of empirical studies report their discontent with the amount of systematic study of item construction) yet has not sparked enough interest to motivate the field to engage in further study. I will describe the existing research investigations and reviews regarding the use and manipulation of several commonly studied rules. I then synthesize the empirical findings using meta-analytic techniques. Throughout this summary, I rely heavily on the reviews by Haladyna and Downing (1989a, 1989b). I briefly review their work and describe the rules included in this synthesis. For a more thorough review of multiple-choice formats and examples of items in each format, see Haladyna (1994).

A few definitions will clarify terms as I use them in this report. A *multiple-choice item* is a question with a *stem* (the body of the question or prompt) and a corresponding set of *options*, including one correct or best option and one or more *distractors* (incorrect options, also referred to as *foils*).

About the "Rules"

Haladyna and Downing (1989a) classified 43 multiple-choice (MC) item-writing rules into three major categories: (1) general item writing, (2) stem construction, and (3) option development. The seven rules originally selected for this synthesis are those most frequently empirically studied (Haladyna & Downing, 1989b); the wording of the rules is theirs. For a list of all the rules, see Appendix A. The first rule is a general item-writing procedural consideration: *avoid the complex multiple-choice (Type K) format.* Two are stem construction considerations: *state the stem in question form*, and *word the stem positively*. Four are general option-development considerations: *use as many functional distractors as possible; avoid, or*

use sparingly, the phrase "all of the above;" avoid, or use sparingly, the phrase "none of the above;" and *keep the length of options fairly consistent*.

Research Questions

I designed this study to synthesize the empirical research on several multiple-choice item-writing rules. These rules include: (1) avoid, or use sparingly, the option "none-of-the-above;" (2) avoid, or use sparingly, the option "all of the above;" (3) use as many functional distractors as are feasible; (4) avoid using the complex or Type-K multiple-choice format; (5) word the stem positively; (6) use the question format, avoid the completion format; (7) keep the length of options fairly consistent. I ask the following questions about each change in format:

1. What is the mean effect of each format on item difficulty, discrimination, and test reliability and validity?
2. Is each mean format effect different from zero?
3. Are the effects homogenous (similar) across studies?
4. Does the magnitude of the format effect vary by (a) age level of subjects, (b) the subject matter of the tests, or (c) other study level characteristics such as how tests were constructed?
5. Is a fixed-effects model tenable, through which we can explain with just a few predictors why study results differ?

As a secondary objective, I will address the theoretical issues regarding each item-writing rule in light of the results of the meta-analysis.

1. What evidence do we have regarding the validity or importance of each rule in item writing?
2. Does the evidence support or refute each rule?

3. What is an appropriate "model" for the factors contributing to item difficulty, with respect to item format effects? Can we similarly model item discrimination, test reliability, or validity?

Related Reviews

Robert Ebel (1951) reported in the first edition of *Educational Measurement* that his review of the literature uncovered five research articles on preparing multiple-choice (MC) items. More recently, Haladyna and Downing (1989a) reviewed 46 authoritative references (dating back to 1935) to develop a taxonomy of MC item-writing rules, assessing the level of consensus regarding each rule as well. Of the 43 item-writing rules suggested, 10 rules addressed general item-writing, 6 addressed stem development, and 20 addressed option development. Lack of consensus among references was found mainly for empirically-testable rules rather than value-laden rules (for which authors did not question the validity of the rule or demand empirical evidence).

Haladyna and Downing (1989b) assessed the validity of those item-writing rules by reviewing research dating from 1926. Their review of the research uncovered 96 theoretical and empirical studies. The frequency of study for each rule and their assessment of the effects on item statistics are summarized in Table 1. For example, they reported that the use of the NOTA option generally had a negative effect on item and test characteristics. Items including this option were more difficult, less discriminating, and test scores were less reliable.

Table 1

Summary of Rules, Frequency of Study, Author Support, and Effects

Rule	Frequency of Empirical Study	Author Support For	Author Support Against	Effect on Difficulty	Effect on Discrimination
Avoid NOTA	10	26	27	*	*
Avoid AOTA	3	19	15	*	*
Use as many functional distractors as feasible	32	16	13	*	
Avoid complex formats	9	6	8	*	*
Avoid negative phrasing	4	31	4	*	
Use the question or completion format	6	41	0	*	
Keep option lengths similar	9	38	0	*	

Note. Frequency is the number of independent empirical reports on each item-writing rule. Author support is the number of authors who wrote about the rule and either supported it or advised against it.

* Rule alternatives had an effect based on Haladyna & Downing's review of the research.

Meta-analytic techniques were not strictly applied to the synthesis of effects in the Haladyna and Downing reviews. In some cases, they reported weighted effect sizes, but did not describe their weighting factor. Averaging effects (with no assessment of consistency) and vote counting can be problematic (Hedges and Olkin, 1980), can overlook small effects, and do not evaluate the potential effects of study characteristics on reported outcomes. Finally, Haladyna and Downing (1989b) concluded that new research was needed on six rules in particular, including the use of "none of the above" and "all of the above," the complex MC format, negative phrasing in the stem, the number of options, and the use of question format versus the completion format. Most of these rules are included in this meta-analysis.

Meta-analytic techniques have not been rigorously applied to the area of item format effects. However, Knowles and Welch (1992) conducted a meta-analytic review of item

difficulty and discrimination for multiple-choice items using NOTA. They argued that the use of NOTA was controversial because common advice on the rule to avoid the use of NOTA was not based on empirical research. They summarized the results on the use of NOTA "because of the contradictory results of the research" (p. 573), but they did not try to understand possible reasons for contradictory results. They used methods suggested by Hunter and Schmidt (1990) and reported non-significant average effect sizes comparing items with NOTA as an option to items without the option (effect sizes of .01 for discrimination and -.17 for difficulty were reported).

Knowles and Welch's results did not match Haladyna and Downing's summary of effects, and did nothing to explain the controversy about NOTA. The authors simply summarized overall effects, which may be misleading. Furthermore, Knowles and Welch either failed to report enough information to allow the reader to assess the results or failed to conduct a comprehensive analysis of their data. They did not report which study characteristics were coded or how effect sizes were calculated.

A second meta-analysis was conducted by Aamodt and McShane (1991), from the same institution as Knowles and Welch. This study summarized results of research on three formats: the number of options per item, the order of items by difficulty, and the organization of items by content. Aamodt and McShane reported methods in an identical format to that of the Knowles and Welch article, and the review suffered from many of the same omissions.

Nonetheless, Aamodt and McShane reported "neither the number of choices in an item nor the arrangement of items in an exam greatly affect exam scores. Thus, three-choice items can be confidently used with the advantage of less time to both create and take the exam" (p. 156). The effect size (standardized difference in item difficulty) comparing a four-option to a three-option item was 0.09. The effect size comparing easy-to-difficult item order versus

random order of items was 0.11; the contrast of easy-to-difficult versus difficult-to-easy orderings was 0.22. Finally, the effect size for ordering items by content area versus random order was 0.04.

Aamodt and McShane's conclusion, that the number of choices per item does not affect exam scores, appears to be based on the size of the mean effect (.09) rather than their 95% confidence interval (.04 to .13) which does not contain zero. The primary problems are whether or not a standardized difference in item difficulty is the appropriate effect to synthesize and if not, what is the variance of the difference in item difficulties (essentially, what is the distribution of the difference in item difficulties?). These are two of the questions that were unanswered in both previous meta-analyses, yet must be addressed in order to conduct a meta-analysis.

I based the following section on a general review of the literature, including articles used in my synthesis. The method I used to identify and select studies for inclusion in the synthesis is reported in the *Method* section.

Review of the Literature

In the following review, I present the level of agreement and treatment of each rule found among authors reviewed by Haladyna and Downing (1989a). Then, I report the number of articles found for my synthesis, the number of independent outcomes in those articles, and the subject area for each study. In many cases, any given article will contain multiple *studies* and multiple *trials* (replications of a study or multiple administrations of a given exam). Trials and studies are considered the same if the samples for each study across trials are independent (different groups). I summarize the general level and direction of effects found in each study in

tables 2a-2f. The name of each exam used and the method of constructing each exam (manipulation of formats) can be found in Appendix B [these are listed by "Study ID" which is the first letter of each author's last name and the publication year].

General Item-writing: Avoid the complex multiple-choice format (Type K)

#. Multiple-choice items measure higher-order thinking skills better than:

1. Hands-on performance items
2. Matching items ← *initial options which are part of the stem from which to select a response below.*
3. Short answer items
4. Essay items
5. True-false items

A. 1 only.
B. 1 and 4. ← *secondary response options from which to select an answer to complete the stem.*
C. 1, 2, and 3.
D. 2, 3, and 5.
E. None of the above

Haladyna and Downing (1989a) reported that 6 of 14 authors agreed with this rule as stated. However, since only 14 of 46 sources mentioned this rule, it was not as important as others. This rule has not reached consensus among measurement authors.

Seven articles reported thirteen independent outcomes appropriate to use in this synthesis. These tests were given for a real-estate course (six trials), social sciences (three trials), health sciences (two trials), and one each in language arts and mathematics.

Table 2a

Using complex alternatives: Study Results Summaries by Outcomes Reported

Study	Difficulty	Discrimination	Reliability	Validity
Hughes, Trimble (1965)	+	=	.	.
Huntley, Plake (1988)	=	=	.	.
Mueller (1975)	+	=	.	.
Nyquist (1981)	+	—	.	.
Tollefson, Tripp (1983)	+	=	.	.
Tripp, Tollefson (1985)	+	=	—	.
Weiten (1982)	+	—	—	=

+ increase; **=** no change; **–** decrease; **.** not evaluated.

Because of the prevalent use of Type-K items in medical-licensure and certification exams, Albanese and others have studied the potential for clueing (presence of clues to the correct answer in the item) in this type of item by comparing student performance on complex and multiple true-false items (similar to Type-K items except where the test taker responds *true* or *false* to each of the initial options. They have repeatedly found that clueing present in complex multiple-choice items decreases item difficulty compared to the same items in multiple true-false format and also decreases reliability (Albanese, 1982; Albanese, 1993; Albanese, Kent, & Whitney, 1979; Kolstad, Briggs, Bryant, & Kolstad, 1983). As distractors, complex options may have little discriminating power (Rossi, McCrady, & Paolino, 1978).

Stem Construction: State the stem in question form

A basic first step in item writing is deciding whether to write the stem as a complete question or in sentence-completion form where the options complete the stem (the example Type-K item above is in sentence-completion form). Forty-one of the 46 sources reviewed by

Haladyna and Downing (1989a) addressed this issue and all supported the use of either format. However, Haladyna and Downing (1989b) revised the rule which was previously stated "use either the question format or sentence-completion format" to read "use the question format" since the results of six studies suggested that the question format improved the performance of the item.

Ten articles reported seventeen independent outcomes appropriate to use in this synthesis. These tests included social sciences (seven trials), the Army Basic Military Subjects Test (four trials), science (three trials), and language arts (three trials).

Table 2b

Using Complete/Question Stem: Study Results Summaries by Outcomes Reported

Study	Difficulty	Discrimination	Reliability	Validity
Board, Whitney (1972)	—	.	+	+
Crehan, Haladyna (1991)	=	=	.	.
Dudycha, Carpenter (1973)	—	=	.	.
Dunn, Goldstein (1959)	=	.	=	=
Eisley (1990)	+	=	=	.
Schmeiser, Whitney (1973)	—	.	=	=
Schmeiser, Whitney (1975a)	—	=	+	+
	=	=	+	+
Schmeiser, Whitney (1975b)	—	=	=	=
Schrock, Mueller (1982)	=	.	=	.
Statman (1988)	—	.	.	.

+ increase; **=** no change; **–** decrease; **.** not evaluated.

Stem Construction: Word the stem positively

Whether to word the stem positively or negatively was addressed by 35 authors (Haladyna & Downing, 1989a). Thirty-one of these authors suggested avoiding negative stems. Haladyna and Downing (1989b) also reviewed four empirical studies, which did not appear to invalidate this recommendation.

Five articles reported sixteen independent outcomes appropriate to use in this synthesis. These test included science (ten trials), mixed subjects (four trials), and one each in social science and health science.

Table 2c

Using negatively worded stems: Study Results Summaries by Outcomes Reported

Study	Difficulty	Discrimination	Reliability	Validity
Dudycha, Carpenter (1973)	+	.	.	.
Harasym, et al. (1992)	—	.	+	.
Nyquist (1981)	=	—	.	.
Tamir (1993)	=	.	.	.
Terranova (1969)	+	.	=	.
	=	.	=	.

+ increase; **=** no change; **–** decrease; **.** not evaluated.

Tamir (1991) found that negative items were more difficult when they included items of a higher cognitive level. Cassels & Johnstone (1984) similarly argued that questions with a negative stem may require at least one additional thinking stage than the same question worded positively, since such items were more difficult. Casler (1983) examined the practice of emphasizing the negative term in the stem by underlining it or capitalizing all letters (e.g., <u>not</u>, NOT). By underlining the negative word, the item became less difficult for high ability students

and more difficult for lower ability students; item discrimination was greater for the emphasized questions. By capitalizing all letters in the negative word, the items were consistently less difficult with no effect on discrimination.

Chang (1995) took a slightly different approach to the issue of using negative wording and argued that the appropriate comparison is regarding the consistency of the connotation among items, where connotatively consistent items agree with the connotations of a majority of other items. In a study where college students responded to the "Life Orientation Test," confirmatory factor analysis showed connotatively consistent and connotatively inconsistent items were correlated, but measured distinct traits. Chang's position was that use of connotatively inconsistent items may alter the operational definition of the underlying construct (e.g., "feeling happy" is not the same as "not feeling sad"). Whether this applies to achievement-type items is unclear. None of the research examined here evaluated the factor structure of a test composed of positively and negatively worded items. This seems like an appropriate area for further study.

Option Development: Use as many functional distractors as possible

Prior to their review of the empirical research, Haladyna and Downing (1989a) found that the rule "use as many options as feasible" was supported by sixteen of the twenty-nine authors addressing this issue. They carefully reviewed theoretical and empirical studies and concluded that the key is not the *number* of alternatives but the *quality* of the alternatives. They also mentioned that the evidence did not support the standard use of four to five options, and "in most testing circumstances, three options per MC item suffices" (Haladyna & Downing, 1989b, p. 58). Based on their review, they revised the rule: "use as many functional distractors as possible."

Twenty-five articles reported 51 independent outcomes appropriate to use in this synthesis. Results for this rule are relatively more difficult to summarize, since studies examined various combinations of numbers-of-options, from two to five options.

Table 2d

Increasing the number of options: Study Results Summaries by Outcomes Reported

Study	Difficulty	Discrimination	Reliability	Validity
Asmus (1981)	—	.	=	.
Budescu, Nevo (1985)	+	+	+	.
Catts (1978)	+	.	=	.
Costin (1970)	+	—	.	.
Costin (1972)	=	=	=	.
Costin (1976)	=	.	=	.
Crehan, Haladyna, Brewer (1993)	—	=	.	.
Denney, Remmers (1940)	.	.	+	.
Duncan (1983)	=	=	=	.
Hodson (1984)	+	=	=	.
Hogben (part 1, 1973)	—	—	—	.
Hogben (part 2, 1973)	+	+	+	.
Kolstad, Briggs, Kolstad (1986)	=	.	.	.
Kolstad, Kolstad, Wagner (1986)	+	.	.	.
Owen, Froman (1987)	=	=	.	=
Ramos, Stern (1973)	.	+	+	.
Remmers, Adkins (1942)	.	.	+	.
Remmers, Ewart (1941)	.	.	+	.
Remmers, House (1941)	.	.	+	.
Remmers, Sageser (1941)	.	.	+	.
Ruch, Charles (1928)	.	.	+	.
Ruch, Stoddard (1925)	+	.	+	.
Straton, Catts (1980)	—	+	—	.
Trevisan, Sax, Michael (1991)	+	.	=	+
Trevisan, Sax, Michael (1994)	+	.	—	.
Williams, Ebel (1957)	+	+	+	.

+ increase; **=** no change; **–** decrease; **.** not evaluated.

These studies, listed in Table 2d, included tests covering language arts (nineteen trials), math (four trials), science (six trials), social science (twelve trials), mixed subjects (three trials), an Air Force instructor exam (four trials), and a musical acoustics exam (three trials).

One potentially important study characteristic is the method used to delete options across forms. In twenty-six trials, options were randomly deleted to create items with varying numbers of options. In fourteen trials, the most ineffective distractors were deleted, and the most attractive option was deleted in three trials. Trevisan, Sax, & Michael (1994) added options to two-option items using the Haladyna and Downing (1989a) taxonomy of rules as a guideline.

Several researchers have demonstrated the power of deleting ineffective options for improving the efficiency of a test (Haladyna & Downing, 1988; Haladyna & Downing, 1993; Zimmerman & Humphreys, 1953). However, the number of functional distractors (and the number of options) may have more of an effect on lower ability students (Haladyna & Downing, 1988). Apparently, very few 5-option and 4-option items actually have a complete set of functional distractors. Wakefield (1958) found that 16% of 4-option items functioned like 4-option items; 3% of 5-option items functioned like 5-option items.

Haladyna and Downing (1988) 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). When they later examined a standardized medical-education test, the reading and social-studies ACT subtests, and a health-science state certification exam, 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 (1.1% to 8.4% respectively). Also, the number of effective distractors was

unrelated to item difficulty and positively related to item discrimination. They suggested that three options per item may be a natural limit for item writers in most circumstances.

The optimal characteristics of three options were also illustrated in studies on tonal information transmission (Sumby, Chamblis, & Pollack, 1958; Pollack & Ficks, 1954). Having three options at each decision point (where the subject must identify letters which were encoded and represented by tonal patterns) resulted in more effective identification of letters than having two- or four-option items per decision point (binary-coded or quinary-coded networks).

Several others have examined the issue of optimal number of options on a theoretical, mathematical level. 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 (r_{tt}), a constant (f, determined by the percentage of correct responses), and the number of options per item in the original and revised tests, determined the revised reliability: $r'_{tt} = \frac{Nfr_{tt}}{N + (N-1)(f-1)r_{tt}}$. However, Lord (1977) also explained that "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). Tversky (1964) demonstrated how using three options per item maximizes discrimination, power, and information of a test, given a fixed total number of options for a test.

Ebel (1969) also derived a predictive formula which was a function of the number of items (k) and the number of choices per item (N), $r = \frac{k}{k-1}\left[1 - \frac{9(N+1)}{k(N-1)}\right]$. This was offered as an alternative to the methods of Remmers and others who predicted reliability as a function of the Spearman-Brown formula, which required an empirically determined reliability coefficient as a starting point. The predictive power of Ebel's formula was good for tests of at least 100 items,

and suggested a trade-off between the number of items and the number of options. Finally, Grier (1975) extended Ebel's formula to estimate optimal reliability. Results extended support for the theoretical advantages of 3-option items, showing their use maximizes expected reliability of a test when the number of items is increased to compensate for fewer alternatives per item.

Option Development: Avoid, or use sparingly, the phrase "all of the above"

Haladyna and Downing (1989a) found this rule to be one of the most controversial. Nineteen authors favored the rule and 15 suggested that "all of the above" (AOTA) could be used effectively. They reported that the use of AOTA increased item difficulty in three studies and decreased discrimination in two of the three studies.

Mueller (1975) was the only author who had examined the effects of inclusive alternatives (both none-of-the-above and all-of-the-above) and reported independent outcomes for AOTA. The others confounded AOTA with other formats. Because Mueller's study is the only independent examination of AOTA, it will be described more in depth here, but this rule is not included in the meta-analysis below. The use of AOTA will not be discussed after this section because this was the only study that examined it. Item difficulty and discrimination were studied using results from a real estate exam including six independent testing periods over two years, with 4,642 examinees. One important problem with Mueller's design was that items were **not** stem equivalent across formats. Examinees only took one form that contained items in all formats. In order to generalize these findings, we must assume that item difficulties were randomly distributed across all formats used ("none of the above," "all of the above," complex alternatives, and standard specified alternatives). Since the average number of items in the

AOTA format was 13 while the average number of standard specified items was 15 across the six independent tests, this may be a plausible assumption.

Mueller reported that items with AOTA were the least difficult among the formats examined (including NOTA and Type-K items) and where AOTA was the correct response, the item was very easy. The weighted (by number of items and subjects) mean item difficulty for standard items was 0.788 (s.d. = .045) and for items with AOTA it was 0.767 (.109). The use of AOTA increased item difficulty on average by 0.021, a very small difference. However, AOTA items were much less difficult than complex items (0.64) and items with "none of the above" (0.74). Item discrimination was slightly effected, dropping from 0.30 in standard items to 0.26 in items with AOTA. All of the alternative distractors (incorrect options) examined appeared to function better than the substantive distractors. As a distractor, AOTA functioned better than most substantive distractors.

Based on the results of this one study, the use of AOTA may have no effect on item difficulty and may very slightly decrease item discrimination. However, AOTA may actually function well as a distractor. Haladyna and Downing (1989b) recommended that this rule be examined more, particularly in light of the disagreement among item-writing authors.

Option Development: Avoid, or use sparingly, the phrase "none of the above"

Twenty-six out of 33 authors recommended that "none of the above" (NOTA) should be avoided (Haladyna & Downing, 1989a). This was one of the most controversial rules found. Upon reviewing ten empirical studies investigating the use of NOTA, Haladyna and Downing (1989b) found that using NOTA generally increased item difficulty and lowered item discrimination and test reliability. They found no advantage to using the NOTA option.

Seventeen articles reported fifty-six independent outcomes appropriate to use in this synthesis. Two of these studies combined the effects of AOTA with the option NOTA so that independent effects for each option type were unrecoverable (Dudycha & Carpenter, 1973; Hughes & Trimble, 1965). These two studies are included in this analysis for NOTA.

Table 2e

Using "None of the above": Study Results Summaries by Outcomes Reported

Study	Difficulty	Discrimination	Reliability	Validity
Crehan, Haladyna (1991)	+	=	.	.
Crehan, Haladyna, Brewer (1993)	+	=	.	.
Dudycha, Carpenter (1973)	+	—	.	.
Forsyth, Spratt (1980)	+	.	.	.
Frary (1991)	+	=	.	.
Hughes, Trimble (1965)	+	=	.	.
Kolstad, Kolstad (1991)	+	.	=	.
Mueller (1975)	+	—	.	.
Oosterhof, Coats (1984)	+	.	—	.
Rich, Johanson (1990)	+	+	=	.
Rimland (1960)	=	.	.	.
Schmeiser, Whitney (1975a)	+	—	—	—
	=	=	—	—
Tollefson (1987)	+	=	=	.
Tollefson, Chen (1986)	+	—	.	.
Tollefson, Tripp (1983)	+	=	.	.
Wesman, Bennett (1946)	+	—	.	—
	=	.	.	+
Williamson, Hopkins (1967)	+	.	=	=

+ increase; **=** no change; – decrease; **.** not evaluated.

The reasons for using NOTA vary a great deal and in part appear to dependent on the subject matter to which it is applied. Boynton (1950), for example, used NOTA as an option in spelling items for a Civil Service exam as a means to increase the number of alternatives and reduce the chance of a correct guess. He found that items that did not contain the correct

spelling were much more difficult. Appropriate uses of NOTA will be discussed further in the *Discussion* section.

Gross (1994) made an argument for logical rather than empirical guidelines for item writing. He suggested *"any stem or option format that by design diminishes an item's ability to distinguish between candidates with full versus misinformation, should not be used*" (p. 125). This, at face value, seems to be a sensible position. He also suggested "unless a new perspective is raised, we know enough about the NOTA format to take a firm, logical position--it should not be used" (p. 126). He illustrated his argument with the following example:

Suppose you are faced with the following multiple-choice question.

> Which of the following cities is the capital of Texas?
>
> a. Dallas
> b. El Paso
> c. Houston
> d. Lubbock
> e. None of the above

Did you answer the question correctly? "The correct answer is none of the above, because, as everyone knows, the capital city of Texas is--*San Antonio*. *What! The correct answer is not San Antonio, but Austin?*" (p.124). Gross argued that it doesn't matter; neither city is listed. He suggested that the correct answer could be obtained with misinformation. "This is the real, unnoticed flaw of the NOTA option" (p. 124).

Gross argued that NOTA items reward students with serious knowledge deficiencies or misinformation. He cited a footnote from the Frary (1991) report suggesting that "all of the above" should not be used because of its logical deficiency. He also stated that the study by

Frary "is mechanically sound and well focused" (p. 124). However, Frary's study is problematic because the item stems were *not* equivalent; therefore, Frary's suggestions regarding the use of complex alternatives should be considered only in light of this flaw. Frary reported that the results of his study on the use of none of the above in college level exams "suggest that the general advice to avoid NOTA items is not justified" (p. 122).

The logical problem is not a new critique of NOTA. In 1951, Ebel addressed the potential dangers in using NOTA. One danger that exists in using NOTA is that

> the examinee who chooses 'none of these' as the correct response may be given credit for a wrong answer. [This] danger can be avoided by sparing use of 'none of these' *as the correct answer* after the beginning of the test, and by limiting its use as the answer to items in which the possible incorrect responses are relatively few. (Ebel, p. 236)

When the number of possible incorrect responses to an item is limited, they should all be included as distractors, so that only examinees who solve the problem correctly will be likely to choose NOTA. When it is impossible to anticipate all of the possible errors students might make in responding to an item, it would be undesirable to use NOTA as the correct answer with only a few of the possible incorrect responses listed; in such a case, "it is far more appropriate to use it as a distractor" (Ebel, 1951, pp. 237).

Finally, Lehmann (personal communication, October, 1995) suggested that Gross' argument was not complete. If a student truly did not know the answer, for example, to the capital-city question above, then his or her chance of getting the item correct by guessing would be one out of five options or 20 percent. If the options used as distractors are plausible, perhaps based on responses from students during class discussions or previous uses of the item, then they should serve their purpose well. If the student is able to at least eliminate the distractors, then

the information they use to select NOTA may be irrelevant. The 20 percent chance of guessing correctly is tempered by the plausibility of the distractors.

In other cases, for example computational problems, we may want to use NOTA to eliminate the chance that a student would continue to recompute the problem until one of the options is obtained. The presence of NOTA may limit the number of times a student recalculates the problem, improving the item's ability to distinguish between students with full information versus misinformation.

Option Development: Keep the length of options fairly consistent

All thirty-eight authors reviewed by Haladyna & Downing (1989a) agreed that the length of options should be consistent. This rule is mentioned because of a tendency of some item writers to be wordier in writing the correct option; item writers may be slightly more descriptive for the correct option than for the distractors. This is only the case in some exams--where options are more than single-word alternatives or a set of numbers solving a math problem. Seven retrieved articles reported seventeen independent outcomes appropriate to use in this synthesis. These studies used tests in the health sciences, social sciences, and the Army Basic Military Subjects Test.

Table 2f

Correct option is different length: Study Results Summaries by Outcomes Reported

Study	Difficulty	Discrimination	Reliability	Validity
Board, Whitney (1972)	—	.	—	—
Dunn, Goldstein (1959)	—	.	=	=
Evans (1984)	—	.	.	.
McMorris, et al. (1972)	—	.	=	=
Schmeiser, Whitney (1973)	+	.	—	—
Strang (1977)	—	.	.	.
Weiten (1984)	—	—	=	=

+ increase; **=** no change; – decrease; **.** not evaluated.

As mentioned earlier, this rule is substantively different from the others in that differences in option length are more often viewed as an item-writing fault than a conscious decision made in formatting items. A skilled item writer should not intentionally decide to make the correct option longer or shorter than all other options. This rule is included in this synthesis because of its frequency of study.

Chase (1964) reported that longer alternatives resulted in higher response rates, particularly when following difficult items. When difficult longer-alternative items followed very easy items where a shorter alternative was the correct option, there was no tendency to select the longer alternative (i.e., the long-alternative response set disappeared). The existence of a response set (choosing long alternatives in difficult items) depended on the nature of the items surrounding the difficult ones.

Carter (1986) reviewed teacher-made tests from 78 teachers in four states. She found at least one item in 86% of the tests had a longer correct-option. The teachers said that they needed to make sure the correct answer was worded so that no one would argue about its correctness. Teachers were mostly unaware of this item-writing principle. She also cited several researchers

who identified "concern with length of the correct answer" as an element of test-wiseness (Millman, Bishop, & Ebel, 1965, as cited by Carter). Mentzer (1982, as cited by Carter) also found this tendency in 35 item files of college psychology texts.

Summary

These studies illustrate the importance of promoting item-writing rules, of educating item writers about basic item-writing principles, and in understanding the impact of violating the rules on item and test statistics. The prevalence of item-writing faults and disagreement among authors on the importance and "validity" of many item-writing rules only adds to our need to understand the often-contradictory findings throughout the literature. Meta-analysis of multiple-choice item-writing rules provides a sensible set of tools to address these issues. However, there is no tradition in classical test theory or meta-analysis for the quantitative synthesis of changes in item and test statistics as outcomes. Without tested models for the synthesis of these outcomes across studies, we can only approximate a model at this time. The following analytic methods are proposed to begin this process. They are derived from the work of many meta-analysts and particularly, the methods presented in *The Handbook of Research Synthesis* (edited by Cooper & Hedges, 1994).

Method

Data Collection

The collection of data (studies) began with the reference lists provided by Haladyna and Downing (1989b), Aamodt and McShane (1991), and Knowles and Welch (1992). The empirical studies investigating each item-writing rule were obtained and the references from

those studies were also reviewed. Computer searches of the *PsychLit, ERIC* (Educational Resources Information Center), and *Dissertation Abstracts International* databases were conducted. Finally, the most recent four years of the major measurement journals were searched manually, including *Applied Measurement in Education, Applied Psychological Measurement, Educational and Psychological Measurement,* and *Journal of Educational Measurement*. Correspondence with several authors via electronic mail also led to several studies not found in the above ways; these included theoretical articles concerning measures of item quality and effectiveness. All authors contacted except one assisted me in my search or request for unpublished or hard-to-find documents.

Studies were screened for inclusion in the meta-analysis based on the following criteria:

1. The study evaluated the effect of at least one of the following multiple-choice item-writing rules on item difficulty: (a) using the "none of the above" option, (b) using the "all of the above" option, (3) varying the number of option, (4) using a complex format, (5) using negative phrasing in the stem, (6) using the question versus completion format, or (7) varying option length consistency.
2. The study reported the number of items in each format, the number of subjects, and at least one of the following outcomes: item difficulties, item discriminations, test reliabilities, test validities for each format.

This process uncovered 109 studies. Five were irretrievable: Knowles & Welch (1990); Charles (1926); Wolkow (1978); Parker & Somers (1983); and Swanson (1976) whose data were later reported in Duncan (1983). Sixteen studies were eliminated because they did not report the statistics for any of the four outcomes included in this meta-analysis (Boynton, 1950; Carter, 1986; Casler, 1983; Chase, 1964; Cassels & Johnstone, 1984; Ebel, 1969; Grier, 1975; Haladyna & Downing, 1993; Harasym, Doran, Brant, & Lorscheider, 1993; Jones & Kaufman, 1975; Millman, 1978; Rossi, McCrady, & Paolino, 1978; Tamir, 1991; Tollefson & Tripp, 1986; Wakefield, 1958; Zimmerman & Humphreys, 1953). Six studies were eliminated because they

did not use achievement or aptitude-type items; they included attitude inventories (Chang, 1995; Remmers & Ewart, 1941; Remmers, Karslake, & Gage, 1940; Remmers & Sageser, 1941) and auditory exams (Pollack & Ficks, 1954; Sumby, Chambliss, & Pollack, 1958).

Twenty-five articles were theoretical treatments of at least one item-writing rule and did not include quantitative results. These will be used when appropriate for discussion purposes but were not included in the empirical synthesis. **Fifty-seven** studies met the above criteria--studies used in the final meta-analysis are starred (*) on the reference list.

The distribution of publication or print date of these studies in each decade from the 1920's to 1995 is illustrated in Table 3.

Table 3

Publication Date of Studies Included in the Synthesis

	1920-29	1930-39	1940-49	1950-59	1960-69	1970-79	1980-89	1990-95
Number of Studies	2	0	4	2	4	14	21	10

Coded Study Characteristics

Independent variables. The following study characteristics were coded as independent variables (see Appendix C for the code-book & coding form). Most studies involved random assignment of subjects to forms when multiple forms were used; eight studies used previous

assignments of students (e.g., by course sections) or some matching criteria (e.g., IQ scores). Table 4 presents counts for several coded variables.

Table 4

Summary of Coded Variables

Variable	Count
Document Type	
journal	46
conference paper	5
dissertation	3
technical reports	2
unpublished manuscript	1
Author Affiliation	
university	47
testing agency	5
government agency	5
Age Level of Subjects	
professional	7
post-secondary	31
secondary grades (7-12)	14
primary grades (1-6)	4
other	1
Test Type	
researcher-made	24
teacher-made	20
standardized	12

The total number of subjects across all studies is 46,628, ranging from 44 to 4,642 within studies (one study reported a sample size of 11,002 for 20 independent exams). The number of test instruments used ranges from 1 to 21 within studies, with a total of 254 instruments across studies. Forty-six studies (81%) used items that were stem-equivalent across formats. The

number of independent effects or studies are reported in Table 5 according to the subject area assessed.

Table 5

Number of Independent Studies (Effects) by Rule and Subject Area

Rule	Language Arts	Math	Science	Social Science	Mixed	Health Sciences	Other	Total # of Studies
NOTA	4	15	5	23	-	2	8	57
Negation	-	-	12	1	4	1	-	18
Open/ Closed	3	-	3	7	-	-	4	17
Option Lengths	-	-	-	5	-	8	4	17
Complex	1	1	-	3	-	2	6	13
# of Options	19	4	6	12	3	-	7	51
Total	27	20	26	51	7	13	29	173

Note. Subject Area "Other" included: musical acoustics, Army Basic Military Subjects Test, real estate course exam, and Air Force Instructor Exam.

Dependent variables. The following study outcomes and statistics were coded to compute effects on the dependent variables of mean item difficulty, mean item discrimination, test reliability and validity:

(a) the reported mean item difficulty for each format or as calculated from mean test form scores;

(b) the reported mean item discriminations for each format and the type of discrimination index reported;

(c) standard deviations for mean item difficulties and discriminations as reported or calculated from t-test results;

(d) reliabilities for each format and the type of reliability reported;

(e) validities for each format (all validities were concurrent criterion-related);

(f) the number of subjects completing items in each format;

(g) the number of items in each format.

Summary statistics for all study results (dependent variables) as coded for this synthesis are reported in Tables 6a through 6f. These tables are summaries of the data coded from primary studies that are used in the synthesis. Results of the synthesis are reported in the following section.

Table 6a

None of the Above: Summary Statistics of Primary Coded Variables and Outcomes

Variable	*Mean*	*s.d.*	Minimum	Maximum	*n*	Label
Specified Items						
Difficulty difficulty	0.638	0.167	0.150	0.893	57	mean item
Discrimination discrim	0.358	0.100	0.163	0.630	47	mean item
Reliability	0.648	0.152	0.260	0.830	21	form reliability
Validity	0.658	0.149	0.410	0.860	11	form validity
Subjects	374	412	17	1800	57	
Items	29	21	7	123	57	
NOTA Items						
Difficulty difficulty	0.586	0.154	0.170	0.820	57	mean item
Discrimination discrim	0.339	0.092	0.187	0.600	47	mean item
Reliability	0.629	0.173	0.170	0.835	21	form reliability
Validity	0.642	0.218	0.220	0.890	11	form validity
Subjects	374	412	17	1800	57	
Items	21	19	4	123	57	

Table 6b

Stem Negation: Summary Statistics of Primary Coded Variables and Outcomes

Variable	*Mean*	*s.d.*	Minimum	Maximum	*n*	Label
Negative Stem						
Difficulty difficulty	0.581	0.100	0.315	0.712	18	mean item
Discrimination discrim	0.260	0.010	0.250	0.260	2	mean item
Reliability	0.808	0.107	0.670	0.923	4	
Validity	.	.	.	.	0	
Subjects	102	119	30	562	18	
Items	26	33	8	143	18	
Positive Stem						
Difficulty difficulty	0.576	0.117	0.387	0.739	18	mean item
Discrimination discrim	0.316	0.078	0.260	0.371	2	mean item
Reliability	0.830	0.123	0.693	0.973	4	
Validity	.	.	.	.	0	
Subjects	103	119	30	562	18	
Items	26	33	8	143	18	

Table 6c

Open/Closed Stem: Summary Statistics of Primary Coded Variables and Outcomes

Variable	*Mean*	*s.d.*	Minimum	Maximum	*n*	Label
Open Stem						
Difficulty difficulty	0.555	0.115	0.307	0.722	17	mean item
Discrimination discrim	0.326	0.053	0.260	0.400	6	mean item
Reliability	0.590	0.215	0.130	0.830	10	
Validity	0.518	0.149	0.350	0.700	4	
Subjects	171	134	25	562	17	
Items	26	13	15	64	17	
Closed Stem						
Difficulty difficulty	0.583	0.130	0.307	0.228	17	mean item
Discrimination discrim	0.332	0.055	0.260	0.386	6	mean item
Reliability	0.608	0.173	0.260	0.770	10	
Validity	0.543	0.115	0.410	0.690	4	
Subjects	180	139	25	562	17	
Items	26	13	15	64	17	

Table 6d

Option Length: Summary Statistics of Primary Coded Variables and Outcomes

Variable	*Mean*	*s.d.*	Minimum	Maximum	*n*	Label
Options Same Length						
Difficulty difficulty	0.510	0.132	0.235	0.702	17	mean item
Discrimination discrim	0.440	.	.	.	1	single item
Reliability	0.587	0.165	0.420	0.750	3	
Validity	0.538	0.157	0.370	0.690	4	
Subjects	106	89	32	249	17	
Items	17	16	4	52	17	
Options Different Length						
Difficulty difficulty	0.591	0.119	0.36	0.783	17	mean item
Discrimination discrim	0.370	.	.	.	1	single item
Reliability	0.437	0.135	0.30	0.570	3	
Validity	0.425	0.058	0.34	0.470	4	
Subjects	107	89	32	249	17	
Items	12	9	4	25	17	

Table 6e

Using Type-K Items: Summary Statistics of Primary Coded Variables and Outcomes

Variable	*Mean*	*s.d.*	Minimum	Maximum	*n*	Label
Standard Options						
Difficulty difficulty	0.689	0.137	0.480	0.850	13	mean item
Discrimination discrim	0.334	0.067	0.250	0.500	10	mean item
Reliability	0.601	0.061	0.520	0.650	4	
Validity	0.540	.			1	
Subjects	399	404	21	1026	13	
Items	32	31	7	124	13	
Type-K Options						
Difficulty difficulty	0.559	0.102	0.374	0.66	13	mean item
Discrimination discrim	0.267	0.102	0.185	0.53	10	mean item
Reliability	0.561	0.172	0.410	0.71	4	
Validity	0.540	.	.	.	1	
Subjects	396	407	21	1026	13	
Items	27	34	1	124	13	

Table 6f

Number of Options: Summary Statistics of Primary Coded Variables and Outcomes

Variable	*Mean*	*s.d.*	Minimum	Maximum	*n*	Label
2-Option Items						
Difficulty difficulty	0.693	0.079	0.577	0.828	12	mean item
Discrimination discrim	0.260	0.089	0.110	0.412	8	mean item
Reliability	0.664	0.191	0.280	0.929	15	
Validity	.	.	.	.	0	
3-Option Items						
Difficulty difficulty	0.693	0.101	0.452	0.895	44	mean item
Discrimination discrim	0.356	0.070	0.200	0.476	33	mean item
Reliability	0.750	0.110	0.520	0.941	43	
Validity	0.240	.	0.240	0.240	1	
4-Option Items						
Difficulty difficulty	0.602	0.111	0.344	0.766	37	mean item
Discrimination discrim	0.349	0.092	0.190	0.580	32	mean item
Reliability	0.738	0.120	0.500	0.945	40	
Validity	0.470	.	0.470	0.470	1	
5-Option Items						
Difficulty difficulty	0.595	0.148	0.362	0.894	31	mean item
Discrimination discrim	0.362	0.091	0.230	0.600	23	mean item
Reliability	0.797	0.095	0.580	0.940	30	
Validity	0.420	.	0.420	0.420	1	
Subjects	185	302	22	1566	51	
Items	42	22	12	100	51	

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

Item Difficulty. One outcome synthesized in this study was the difference in mean item difficulty due to format change. This was reported as a difference in mean item difficulties for items in each format or mean test score for each form or format total score. Previous meta-analysts (Knowles & Welch, 1992; Aamodt & McShane, 1991) most likely calculated the standardized mean difference of item difficulties as the effect size (the difference between mean item difficulties in two formats divided by a pooled standard deviation of item difficulties). It is unclear, however, since they did not report any formulas.

The effect size synthesized was calculated as $T_{diff} = \bar{p}_2 - \bar{p}_1$, 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 alternative format. The conditional variance of T_{diff} was calculated as $\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}$, as described above.

It may be inappropriate to conceptualize the difference between mean item difficulties in terms of a standardized mean difference effect size. Several study design characteristics complicate this analysis and make the use of standardized mean differences inappropriate, including the issue that items may be stem-equivalent in each format or not and that each subject may or may not take items in both formats. Depending on the design, there may be covariation among the mean item difficulties for the two formats, which would have to be accounted for in

calculating the variance of the effect size. An estimate of this covariance is typically not available from the studies involved in this synthesis.

A secondary consideration is whether the overall magnitude of the difficulties is of interest; this is lost when computing the mean difference. That is, is a difference of 0.10 in item difficulty more significant (of practical significance) if the item difficulty is 0.85 than if it is 0.5? This is not easy to evaluate. Standardized mean differences can be difficult to interpret in some contexts. In the case where item difficulties are reported in a similar metric (proportion answering an item correctly), it may be more appropriate to simply report the difference in raw item difficulties of items in two formats.

It remains unclear as to what would be the best way to compute effect sizes for mean item difficulties and their corresponding conditional variance (usually a function of within-study sample size). For my synthesis, the focus is on the difference in item difficulty of items in various formats. The sampling distribution of this *difference* and the standard error of the item difficulty *difference* is unknown. In classical test theory, which virtually all of the studies use to report outcomes, the variance of an item is $p(1\text{-}p)$ where p is the item difficulty. Using this same conceptualization where $\bar{p}$ is the mean item difficulty for items in a given format, we could estimate the variance of the mean item difficulty as $\frac{\bar{p}(1-\bar{p})}{n}$, where n is the number of items contributing to the mean item difficulty, $\bar{p}$. The variance of the difference would then be the sum of the variances for the two given formats.[1]

[1] By not subtracting a factor for the covariation, I am computing a conservative estimate of the conditional variance, making it larger than it actually might be and thus suggesting that the estimate of the mean difference may be less precise.

Item Discrimination. The mean item discrimination was reported in three forms, most commonly as a point-biserial correlation between the item and the given format test score (ranging from -1.00 to 1.00, but usually positive). Less frequently, item discrimination was reported as *D*, the difference in proportion of correct responses for the upper-scoring 27% of students versus the lower-scoring 27% (D also ranges from -1.00 to 1.00). In a couple of instances, discrimination was reported as a tetrachoric correlation. It has been found that various item discrimination indices are highly correlated and discrepancies between them only occur for items at extreme difficulty ranges (Crocker & Algina, 1986; Englehart, 1965). All reported discrimination indices are treated as equivalent in this synthesis; however, when possible, differences between the three types of indices reported will be examined.

Similar problems exist in computing effect sizes and variance estimates for the mean item discrimination as with the mean item difficulty. Do we consider mean item discrimination indices to be like mean difficulties (since they are means across items), or like correlations, since most 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--the average change in correlation between the items and total score for items due to change in format.

All discrimination index values were transformed using Fisher's normalizing and variance stabilizing *Z*-transformation. As described by Rosenthal (1994), for any correlation *r*, $Z_r = \frac{1}{2}\log_e\left[\frac{1+r}{1-r}\right]$. The difference between the *Z*s across formats was calculated as $T_{disc} = Z_2 - Z_1$, where Z_1 is the value for the mean item discrimination (transformed) for items in one format and Z_2 is for the alternative format. The conditional variance of T_{disc} was calculated as

$\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$, the standard formula for the variance of a composite, as described below. The conditional variance of Z_i (σ_i^2 in the conditional variance formula for T_{disc}) was calculated as $v_i = \frac{1}{(n_i - 3)}$, where n_i is the within-study sample size (number of items) of the *i*th study (Shadish & Haddock, 1994).

It is unclear as to what the best estimate of the covariance would be--if in fact the mean item discrimination values for items in two formats covary. The covariance (unknown) between discrimination coefficients was estimated by computing the correlation (ρ) between all reported pairs of discrimination coefficients. For 46 pairs of coefficients in the NOTA studies, r = .886; for 6 pairs in Open/Closed stem studies, r = .978; for 10 pairs in Type-K studies, r = .731; and for various Number-of-Options studies, r = .892 (5- vs. 4-options, n=19 pairs), r = .844 (4- vs. 3-options, n=29), r = .887 (3- vs. 2-options, n=8); no correlations were calculated for the 2 pairs in Negation studies or for the single pair in Option-Length studies. Since these correlations are high, and the resulting conditional variance estimate would be reduced (because we would subtract twice the covariance from the sum of the variances) thus providing a smaller variance estimate, a lower value of 0.60 was used as the estimate of this covariance for all rules. This reduced the probability of underestimating the conditional variance of the effect size and suggested that the effect size estimate was less precise. The result would be a more conservative estimate, although not as conservative as zero covariance that was assumed in computing the conditional variance for the case of mean item difficulty.

Test Reliability. Reliabilities were reported for each set of items in a given format. Most often, the reported reliability coefficient was the KR20[2]. Less frequently, reliabilities were reported as KR21[3]. In a couple instances, Hoyt's[4] analysis of variance analogue to reliability was reported, as were coefficient alpha[5] and split-half reliability coefficients[6]. The coefficient alpha (and equivalents) is a lower bound estimate of the theoretical reliability coefficient. Finally, coefficient alpha can be thought of as the expected value of the split-half correlation, the mean of all possible split-half coefficients (Crocker & Algina, 1986). Because of the moderate equivalence of these various estimates of reliability, all reported coefficients were treated equivalently in this synthesis. When possible, resulting format effect sizes were examined in terms of the type of coefficient reported to examine possible "coefficient-type" effects.

In addition, the reliability of a test may be related to the difficulty and discrimination of a test. Reliability may be slightly higher among tests with homogenous item difficulties (Myers, 1962). There is also a relationship between the reliability of a test r_{tt}, the average item intercorrelation $\bar{r}_{ij}$, and the squared average item-test correlation r_{it}^2 to the effect that $r_{tt} = \frac{\bar{r}_{ij}}{\bar{r}_{it}^2}$ (Silverstein, 1980). This suggests that where studies report multiple outcomes, the appropriate analytic methods should be conducted in a multivariate space, since these outcomes are likely to be correlated. However, since meta-analytic methods are not well developed to deal with large amounts of missing data in multivariate analyses (the case here since many studies do not report on all four outcomes), univariate analyses are used.

---

[2] Kuder and Richardson's formula is one method of computing coefficient alpha, a set of methods based on item covariances, and is appropriate for dichotomously scored items.

[3] Kuder and Richardson's simpler formula assumes items of equal difficulty, otherwise resulting in a lower estimate than the KR20.

[4] Hoyt's reliability treats persons and items as sources of variation and is equivalent to KR20.

All reported reliability coefficients were transformed to Fisher's *Z*s. The difference between *Z*s in each format was calculated as $T_R = Z_2 - Z_1$ and the variance was as for item discrimination effects.

The correlation between reliability coefficients (ρ, used to estimate the covariance between reliability coefficients for items in each format) was estimated by computing the correlation between all reported pairs of reliability coefficients. For 21 pairs of coefficients in the NOTA studies, $r = .905$; for 10 pairs in Open/Closed stem studies, $r = .93$; for 4 pairs in Negation studies, $r = .802$; for 4 pairs in Type-K studies, $r = .367$; for 3 pairs in Option-Length studies, $r = .466$; and for various Number-of-Options studies, $r = .88$ (5- vs. 4-options, n=24 pairs), $r = .84$ (4- vs. 3-options, n=37), $r = .813$ (3- vs. 2-options, n=15).

Since these correlations range from 0.367 to 0.905, a moderate value of 0.60 was used to estimate the covariance of format reliabilities for all rules. This would reduce the probability of underestimating the conditional variance of the effect size which would suggest that the effect size estimate was less precise. The result was a more conservative estimate, as for the case of mean item discrimination.

Test Validity. Few studies actually examined format effects on test validity and in all cases, the validity coefficients reported were concurrent criterion-related validity coefficients: the correlation between a given form and a related criterion measure administered simultaneously (or very close in time). Again, there may be a relationship between test validity and other outcomes described above. The relationship between validity and item difficulty has been found to be negligible (Myers, 1962). However, the relationship between validity and

[5] Cronbach's estimate of internal consistency is based on item variances and covariances.

reliability has been theoretically specified as $\rho_{xy} \leq \sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}$, where the validity is less than or equal to the product of the square roots of the reliabilities of the two forms. Validity coefficients are synthesized without correction for unreliability of either the predictor or criterion.

All validity coefficients are transformed using Fisher's *Z* transformation. The same procedures described above are used with the validity coefficients, where $T_V = Z_2 - Z_1$. In order to compute the variance of the effect size, the validities for each pair of formats are correlated to estimate the covariance. For 11 pairs of coefficients in the NOTA studies, $r = .670$; for 4 pairs in Open/Closed Stem studies, $r = .561$; for 4 pairs in Option-Length studies, $r = .563$. Only one study reported validities for Type-K and Number-of-Options formats, none reported validities for use of Negative stems. A moderate correlation of 0.50 was used to account for possible covariance among test validities for items in different formats. Since it is not clear that such covariation exists, this will results in a conservative estimate of the conditional variance of the effect size, resulting in less precise effect estimates.

Effect size and conditional variance computational formulas are summarized in Table 7. The "1" and "2" subscripts denote a given format (1) and its alternative (2).

---

[6] This is the correlation between two halves of a single test form corrected for length.

Table 7

Effect Size and Conditional Variance Formulas

Dependent Variable	Effect Size	Conditional Variance
Mean Item Difficulty	$T_{diff} = \bar{p}_2 - \bar{p}_1$	$\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}$
Mean Item Discrimination	$T_{disc} = Z_2 - Z_1$	$\frac{1}{(n_1-3)} + \frac{1}{(n_2-3)} - 2\sigma_{12}$
Test Reliability	$T_R = Z_2 - Z_1$	*same as for discrimination*
Test Validity	$T_V = Z_2 - Z_1$	*same as for discrimination*

Data Analysis Procedures

The conceptual design of the meta-analysis stems from the following considerations. The universe to which we hope to generalize is a hypothetical collection of studies that could be conducted on the effects of violating any given multiple-choice item-writing rule. 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 and people within studies. For most analyses, the items will be the unit of analysis with respect to sampling uncertainty. These are the elements of a fixed-effect model. Where this model appears untenable, due to significant heterogeneity among effects across studies, a random-effects model will be used to estimate the average effect, the variance due to sampling of items within a study, and the random effects variance due to the sampling of studies.

In order to account for study precision, the effects are weighted by a function of the variance estimates for each effect as described in Table 7. The weight is w_i and is calculated as $w_i = \frac{1}{\text{var}(T_i)}$. The mean weighted effect size is then $\overline{T}. = \frac{\sum T_i w_i}{\sum w_i}$ (Shadish & Haddock, 1994). The standard error of the weighted mean effect size is $SE(\overline{T}.) = \sqrt{\frac{1}{\sum w_i}}$. To allow for inferences regarding the estimated mean effects, I tested the hypothesis for homogeneity of effects, H_o: $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$. To do so, I calculated the *Q*-statistic which is distributed as a Chi-square with k-1 degrees of freedom: $Q = \sum \frac{(T_i - \overline{T}.)^2}{v_i}$.

A basic set of procedures for meta-analysis includes the testing of homogeneity of effects across studies. When effects appear to be homogenous, we can estimate the common parameter, quantify uncertainty by calculating its standard error, and perform a significance test on the estimate. When effects appear heterogeneous, we can describe the studies using a fixed-effects, random-effects, or mixed model. For the fixed-effects model, we can estimate the population variance, conduct outlier analysis (which may result in the decision to drop certain effects that for some reason are not like the others), conduct moderator analysis based on study characteristics to explain between study differences, and then estimate the common parameter for each homogenous group of studies. When outlier analysis and moderator analysis do not explain the heterogeneity of effects, the best we can do is to then estimate the random variance component. The addition of the random variance component may result in a slightly different estimate of the average effect due to the inclusion of the random variance component in

weighting effects and also results in a larger standard error for the average effect estimate. These were the basic procedures used to obtain the following results.

These effects are all calculated based on a fixed-effects model, where effects variance is a result of sampling of subjects within studies. This assumes that there is one population effect (a common parameter) and that all variation can be accounted for by a couple predictors (moderator variables). For those effects that are found to be heterogeneous, the next steps include an analysis of outliers, analysis of moderators to investigate between study differences, and possible estimation of a common parameter for each homogenous subgroup, if identifiable based on moderator analysis.

For those effects found to be homogenous, the following is offered as a preliminary set of results. The following results are stated in terms of *violating* each rule as stated by Haladyna and Downing (1989b). The effects are in terms of "*what happens to the item and test statistics when a given rule is violated*." Only fixed-effects results are reported at this time.

Considering Effects on Difficulty and Discrimination

Item difficulty and discrimination are frequently used as a means to assess item quality. However, whether or not a specific magnitude of difficulty or discrimination is good or bad depends on the intent of the test designer. Items should be selected, based on their difficulty and discrimination, to optimize the measurement characteristics of a given test for the purposes that it is intended. Knowing what general effect a given formatting characteristic may have on an item's difficulty or discrimination may provide useful information to the item writer, given that the item writer is targeting items at a specific difficulty or discrimination or trying to obtain a wide range of these item statistics.

Kolstad, Briggs, and Kolstad (1984) have suggested, as have many measurement authors, that teachers should use the measurement of instructional objectives as the primary criteria for the content and format of an item in classroom tests. Expectations for the items should be described in advance and item analysis should be interpreted accordingly. Simply knowing if a change in format increases an item's difficulty is not enough to determine whether or not the format is appropriate. For some testing purposes, increasing an item's difficulty may be inappropriate.

In terms of an optimal difficulty level, p_i=0 or p_i=1 provide no information about differences among examinees because everyone either got the item wrong or right. An item provides the maximum amount of information about examinee differences when p_i=.5, where the variance in items scores is maximized at $p_i(1-p_i)$. Lord (1953) suggested that for multiple-choice items, the optimal level of difficulty is just less than halfway between 1.0 and the chance success level. For a five-option item, the chance success level is about 0.20 and the optimal level of difficulty would be just less than 0.60. ***specific #'s from Lord (Mehrens & Lehmann)

Similarly for discrimination, the effect of a given format on discrimination is only part of what the test developer must consider. In terms of the work of Guttman (1950, as cited by Masters, 1988), the most highly discriminating items should be selected for inclusion in a test. Item discrimination is also used in some IRT models as a weighting factor in estimation of abilities. The assumption here is that higher discrimination is better. Masters (1988) has argued that more discrimination is not necessarily better. He demonstrated how an item might be sensitive to individual differences on a second, undesired trait that is correlated with the construct intended to be measured. Possible secondary influences of this type include

opportunity to learn, opportunity to answer, tendency to become fatigued or frustrated, and test wiseness. Since these characteristics are likely correlated with the construct being measured, they may work to make an item unusually discriminating.

> A problem arises … when some students are given an item that tests material that they have not covered. If these students are already low-achieving students, then their failure on that item will make it unusually discriminating, which, in turn, may be interpreted as evidence that the item should be given special weight. In this way, low-achieving students could be doubly penalized. (p. 28)

Considering Effects on Reliability and Validity

Test reliability and validity provide much more direct assessments of the value of item formats, since we usually want tests of greater reliability and validity unconditionally. Unfortunately, the least amount of study has occurred investigating the effects of item formats on these test-level statistics. Classically, if the reliability of a test is high, the observed scores are highly correlated with the true scores. As the reliability of a test increases, the error-score variance becomes smaller (Allen & Yen, 1979), resulting in observed scores that are very close to true scores. This is a widely accepted goal in test development.

The case for the use of validity as an item format quality indicator is more complicated because of the inconsistencies in what qualifies as validity evidence throughout the measurement literature. Messick (1994) has repeatedly argued for a unified view of validity, where "validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the *adequacy* and *appropriateness* of *inferences* and *actions based on test scores*" (p. 33). Since the impetus is on the interpretation and use of test scores, the test user is in an important position to evaluate the meaning of individual scores under local circumstances. The test developer also has responsibility to demonstrate the validity of test scores under intended interpretations and suggest appropriate uses or inferences. With the revision of the *Standards for Educational*

Psychological Tests (AERA/APA/ NCME, 1985), many expect to see a call to bring multiple pieces of validity related evidence to bear on test score use.

For our purposes, the only type of validity-related evidence reported in item format studies is concurrent criterion-related validity, a correlation between a given format and a criterion measure given simultaneously. This measure is problematic for several reasons, including the fact that it is attenuated due to unreliability of the form under investigation and unreliability of the criterion. There may also be restriction of range problems. This further complicates analysis of the *change* in validity coefficient due to format change, when the format change may simultaneously result in a change in reliability that confounds the format effect on validity. Generally, we consider an increase in criterion-related validity to be a positive outcome.

Finally, Haladyna (personal communication, April 24, 1997) suggested that very small changes in item and test statistics could actually lead to dramatic effects in large-scale testing programs. A change in an item difficulty of 0.04 could affect hundreds if not thousands of examinees when the number tested is large.

Results

An overall summary of results is provided in Table 8. The weighted average effect on each outcome is reported with the standard error of the effect in parentheses under the fixed-effect model. Each effect is in the same metric as the outcome; the difficulty effect is in terms of difference in the item difficulty index; the discrimination, reliability, and validity effects are in terms of their original metric. The number of effects (differences) involved is in the lower right hand corner of each cell.

Table 8

Summary of Average Effects and Standard Errors for Violating Each Rule

Rule Violation	Difficulty Index	Discrimination Index	Reliability Coefficient	Validity Coefficient
Using "none of the above"	-0.035 (0.005) *n*=57	-0.027* (0.035) *n*=47	-0.001* (0.039) *n*=21	0.073 (0.051) *n*=11
Stating the stem negatively	-0.032 (0.010) *n*=18		-0.166 (0.082) *n*=4	
Using an open, completion-type stem	0.016* (0.009) *n*=17	-0.003* (0.076) *n*=6	0.031* (0.069) *n*=10	0.042* (0.123) *n*=4
Making the correct option longer	0.057* (0.014) *n*=17			-0.259* (0.163) *n*=4
Using complex (Type-K) format	-0.122* (0.011) *n*=13	-0.145* (0.063) *n*=10	-0.007* (0.083) *n*=4	

* Homogenous effects (consistent across studies), based on *Q*-test statistic.

Using Complex (Type-K) Format

Using the complex or Type-K format decreases the difficulty index, making items more difficult, by an average of 0.122 (*SE* = 0.011, *Q*=16.6, *p*=.17). The complex format is also less discriminating than the standard format, where the discrimination index is decreased by 0.145 on average (*SE* = 0.063, *Q*=1.65, *p*=.98). Finally, there is virtually no effect on reliability when converting items to the complex format, with only four studies reporting.

The evidence found in this synthesis suggests that the complex format should not be used because of its negative effect on discrimination. The increased difficulty of complex formatting may or may not be a positive result, depending on the desired level of difficulty that the item writer is trying to achieve. However, the evidence is not overwhelming and the format may be used appropriately through careful design.

Using an Open Completion-Type Stem

Using an open or completion-type stem had a negligible effect on difficulty ($\overline{T}_{diff}$ =0.015, *SE*=0.009, *Q*=18.3, *p*=.31), no effect on discrimination ($\overline{T}_{disc}$ =-0.003, *SE*=0.076, *Q*=0.02, *p*>.99), reliability ($\overline{T}_{R}$ =0.0312, *SE*=0.076, *Q*=4.4, *p*=.88), or validity ($\overline{T}_{V}$ =0.042, *SE*=0.124, *Q*=1.9, *p*=.60). These findings were consistent across studies. There appears to be no evidence in this synthesis to argue against the use of either format (closed-question stem versus open-completion stem). This finding supports the consensus among authors reviewed by Haladyna and Downing (1989a), but does not agree with Haladyna and Downing's review of the empirical research (1989b) where they concluded that it was best to use the complete question-type stem.

Stating the Stem Negatively

Stating the stem negatively slightly decreased the item difficulty index inconsistently across studies, increasing the difficulty of the item by 0.032 on average (*SE* = 0.010, *Q*=63, *p*<.001). Including the random-effects variance estimate significantly reduced the effect of negative stems to 0.003 (*SE*=0.02).

Negatively worded stems also decreased reliability inconsistently across studies, on average by 0.166 (*SE* = 0.082, *Q*=28, *p*<.001). Using a random-effects model, the average effect remains 0.168 (*SE*=0.217) with a much wider 95% confidence interval due to the additional random-error variance (-0.256, 0.593).

Based on the limited results of this synthesis, the rule to state the stem positively is supported. However, additional analyses should be done to investigate potential factors that might explain some of the variance or inconsistent findings between studies. One study had a strong negative impact on the homogeneity of effects for difficulty (Harasym, Price, Brant, Violato, & Lorscheider, 1992) and one on the homogeneity of effects for validity (Terranova, 1969). When we account for the random-effects variance, the effects on difficulty and reliability are not significant ($\overline{T}_{diff}$ = 0.003, *SE* = 0.02; $\overline{T}_{R}$ = 0.168, *SE* = 0.216).

Number of Options per Item

The effect of increasing the number of options from 2 to 3, 4, and 5 is not consistent across all types for difficulty, discrimination, and reliability. Table 9 summarizes the effects.

Table 9

Change in Item & Test Statistics per Number of Options

	Difficulty	Discrimination	Reliability
2-Options	.698 (.019)	.309 (.056)	.755 (.036) *.718 (.103)*
3-Options	.698 (.011) *.697 (.015)*	.370 (.029)	.804 (.024) *.783 (.045)*
4-Options	.610 (.012) *.604 (.019)*	.364 (.029)	.800 (.025) *.774 (.053)*
5-Options	.591 (.014) *.594 (.027)*	.386 (.037)	.840 (.029) *.824 (.055)*

Note. Parentheses contain standard errors. The second row of values for some outcomes (in italics) represent those estimates under a random-effects model, where the original fixed-effects estimates were non-homogenous across studies.

Briefly, there is no change in item difficulty from 2- to 3-options, a slight increase in difficulty from 3- to 4-options, and no further change from 4- to 5-options per item. There is a slight gain in discrimination from 2- to 3-options, and no further change from 3- to 4- to 5-options per item. Finally, there is a slight gain in reliability from 2- to 3-options, with no further change in reliability from 3- to 4- to 5-options per item. Overall, items become slightly more difficult going from 3- to 4-options and slightly more discriminating with slightly more reliability going from 2- to 3-options per item. This can be seen in Figure 1.

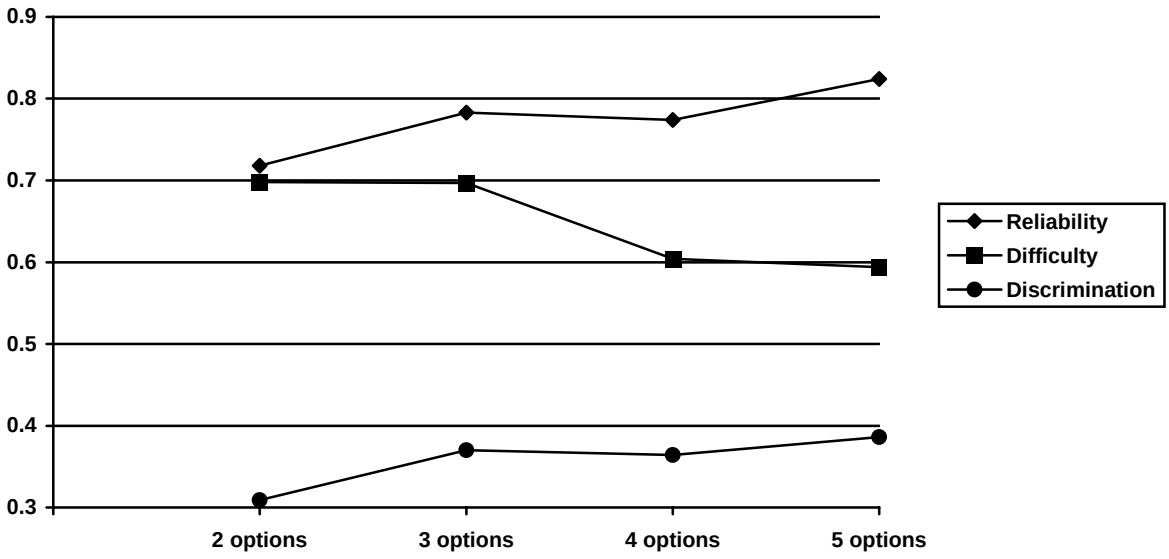

Figure 1. *Change in Item and Test Statistics per Number of Options.*

Based on these results, and based on all but one of the recommendations of the primary study authors, the evidence suggests that the 3-option item may be the most effective format for multiple-choice items. Several considerations make this a reasonable conclusion. First, the amount of time and effort it takes to construct 3-option items is much less than developing the 4- or 5-option item. It seems that most item-writers are able to construct items with one correct answer and two plausible distractors. In some cases, and some fields in particular, the identification of three or four plausible distractors is very difficult. Second, the amount of time it takes to answer 3-option items is reduced because students do not need to read and consider additional options. Finally, the reduction in overall test time may allow for the inclusion of additional items--possibly increasing the content-related validity of the test as well as reliability. Validity was not examined in these studies but would improve our understanding of the effect of changing the number of options per item. This is an area of study that would be fruitful.

Using "None of the Above"

The effect of using the option "none of the above" on item difficulty is small and not consistent across studies. On average, it decreases the item difficulty index, increasing the difficulty of the item by 0.035 (*SE* = 0.005, *Q*=236, *p*<.001). Remember that the estimate of the difference is -0.035 and that a decrease in the difficulty index means the item is more difficult (fewer students answer it correctly). We can account for the random-effects variance by using a random-effects model to weight the average effect. Using the random effects model, the increase in item difficulty is 0.045, the standard error is increased to 0.099 leading to a wider confidence interval for the estimate (-0.065, -0.026). This change is due to the additional random-effects variance.

Using the NOTA option decreases the item discrimination insignificantly but consistently across studies by 0.027 (*SE* = 0.035, *Q*=1.83, *p*>.99) under a fixed-effects model. There is no effect on test reliability--a decrease of 0.001 (*SE*=0.039), which is consistent across studies (*Q*=13.8, *p*=.84).

Finally, there was an insignificant and inconsistent improvement in test validity with an increase of 0.073 on average (*SE* = 0.051, *Q*=40.4, *p*<.001). Using a random effects model, the average increase in validity was reduced to 0.025 (*SE*=0.103), and is nonsignificant.

The evidence for or against the use of NOTA is inconclusive at this point. There is no significant impact on item discrimination, test reliability or validity. Further moderator analysis could be helpful in understanding the inconsistency of effects between studies. One study in particular has a significant negative impact on the homogeneity of effects for reliability and validity (Williamson & Hopkins, 1967). Nothing appears unique about this study at this time.

Making the Correct Option a Different Length

Making the correct option a different length (most often longer than the other options) increased the difficulty index, thereby making the items less difficult on average by 0.057 (*SE* = 0.014, *Q*=18.5, *p*=.30) consistently across studies. This item-writing fault also decreased validity consistently across studies by 0.259 on average (*SE* = 0.163, *Q*=0.76, *p*=.86).

Based on the results of this synthesis, the rule is supported as stated. Options should be written to be similar in length. Again, this rule is addressing an item-writing *fault* more than an item format. However, the results of the synthesis demonstrate its detrimental effects. Clueing most likely accounts for the decrease in the difficulty of the item and loss of validity.

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.

One-fifth of the studies were unpublished and numerous studies reported non-significant findings (ten of the twelve reporting non-significant findings were published articles). Of the 57 studies included in the meta-analysis, five were unpublished conference papers, 3 were dissertations, 2 were testing agency technical reports, and one was an unpublished manuscript. Looking across the rules, several studies found no significant effects on any of the outcomes, including one on complex alternatives, three on open versus closed-stems, two on negatively worded stems, five on increasing the number of options, and one on the use of NOTA. There is no compelling evidence suggesting the presence of publication bias among item-format effects.

Discussion

Based on the evidence synthesized in this meta-analysis, the rules can be restated or revised. Table 10 lists the rules as supported, where starred rules indicate those that were revised from the original wording provided by Haladyna and Downing (1989b).

Table 10

Summary of Multiple-Choice Item-Writing Rules as Supported through the Synthesis

1. * Use "none of the above" when it fits the format and purpose of the exam.
2. State the stem positively.
3. * Use either the open (completion-type) or closed (complete question) stem.
4. Make all options similar in length.
5. Avoid the complex (Type-K) format.
6. * Use three options per item.

* Indicates the rule has been revised as a result of the evidence from the synthesis.

A secondary objective of this study was to address theoretical issues regarding the item-writing rules. Three questions guide this discussion regarding (1) the evidence to support or refute each rule, (2) the evidence regarding the appropriateness of each rule or item-format in item-writing, and (3) the specification of a model of item difficulty (and perhaps item discrimination, test reliability, and test validity).

Does the evidence support or refute each rule?

Three rules have been refuted or altered slightly based on this synthesis. The other three were supported. Each will be briefly discussed, beginning with those that were altered.

The present findings regarding the use of NOTA conflict with those of two previous reviews. The Knowles and Welch (1991) meta-analysis found no significant difference in item difficulty when using NOTA as an option versus specified options. A direct comparison with this meta-analysis was not possible because Knowles and Welch did not specify how their effect sizes were calculated. They based their analyses on 20 effect sizes for item difficulty and 11 for item discrimination. This meta-analysis was based on 57 effect sizes for difficulty and 47 for discrimination. In addition, Knowles and Welch did not evaluate effects on test reliability or validity. Second, these new findings on NOTA also conflict with the summary by Haladyna and Downing (1989b). They summarized the findings of 10 studies and concluded that the use of NOTA generally had a negative effect on item characteristics, making items about 4.5% more difficult. Although the small effect on item difficulty was also found in this synthesis, the decision to recommend against use of NOTA is not supported.

The use of the open-stem (completion-type) or the closed-stem (complete question) also resulted in conflicting results. Haladyna and Downing (1989b) suggested that due to a slight drop in reliability (based on four studies) and a slight drop in validity (based on two studies), the rule should be revised to read: “Use the question format; avoid the completion format” (p. 64). Based on this synthesis (ten effects for reliability and four effects for validity), both reliability and validity increased slightly and consistently across studies when using the completion-type format. The availability of a larger set of studies may have helped clarify some uncertainty and support the consensus among item-writing authors who support the use of either format.

The most studied rule and one of the most controversial is regarding the number of options-per-item. Aamodt and McShane (1992) reported similar findings on this rule, although they had only evaluated the effect of 3-option versus 4-option items on exam scores and item discrimination. The findings do not conflict with the Haladyna and Downing (1989b) recommendation that "the key in distractor development is not the *number* of distractors but the *quality* of distractors" (p. 59). However, based on the results of this synthesis, the rule could be more direct and promote the use of three options per item.

The vast majority of authors who have studied this rule recommended using three-option items (Asmus, 1981; Catts, 1978; Costin, 1970, 1972, 1976; Crehan, Haladyna, & Brewer, 1993; Duncan, 1983; Hogben, 1973; Kolstad, Briggs, & Kolstad, 1985; Kolstad, Kolstad, Wagner, 1986; Owen & Froman, 1987; Straton & Catts, 1980; Trevisan, Sax, & Michael, 1991, 1994). Others who have limited their investigation to four-option versus five-option items recommend using four-option items (Hodson, 1984; Ramos & Stern, 1973). Only one author recommended against using only three options per item. Budescu and Nevo (1985) investigated the assumption of proportionality that suggests that the total testing time is proportional to the number of items and the number of options per item. They found a strong and consistent negative relationship between rate of performance and the number of options for tests of fixed number of items. The number of options accounted for over 50% of the variance on the three tests.

The remaining three rules (state the stem positively, make all options similar in length, and avoid the complex format) remain as stated by Haladyna and Downing (1989b). Stating the item negatively has a negative impact on test reliability. Making the correct option longer makes items easier and negatively impacts validity. Finally, using complex Type-K items makes items substantially more difficult and negatively impacts item discrimination.

What evidence do we have regarding the role of each rule or format in item-writing?

One of the riches of this body of literature is the set of recommendations from authors concerning appropriate uses for each of the formats investigated.

The role of none-of-the-above. Several guidelines commonly recommended for the use of NOTA make it an appropriate option. NOTA can be useful for increasing item difficulty when the options are not approximations (Dudycha & Carpenter, 1973; Mehrens & Lehmann, 1991). NOTA will reduce the magnitude of chance variance reflected in test scores (when the examinee does not possess the relevant knowledge) and subsequently increase test reliability and validity (Williamson & Hopkins, 1967). NOTA should be used as a correct option about 1/c times the number of items in which it appears, where c is the number of options per item. It should also appear as the correct response early in the test in a relatively easy item to assure the examinee that it could be a correct response (Mehrens & Lehmann, 1991).

Several authors, including those whose studies are used in this meta-analysis, have suggested appropriate uses of NOTA. The first is with respect to NOTA serving as the correct answer (Dudycha & Carpenter, 1973; Frary, 1991; Hughes & Trimble, 1965; Mehrens & Lehmann, 1991; Tollefson & Chen, 1986; Wesman & Bennett, 1946).

I. When NOTA is used as the correct response to a MC item, it may prevent simple recognition of the answer for those students who would otherwise not be able to produce it in a completion-type item. Students may recognize the answer if it were an option but not be able to recall it if it were not an option. Examinees must know that all of the distractors are incorrect responses in order to correctly

answer the item with confidence. Recognizing NOTA as the correct response may reflect greater understanding of the material (measuring knowledge of what is wrong) than recognizing one answer is correct.

A second appropriate use of NOTA is closely related to the first. The NOTA option may motivate examinees to consider each option more carefully (Frary, 1991; Oosterhof & Coats, 1984; Wesman & Bennett, 1946).

II. Regardless of whether NOTA is the answer or not, its presence as an option may motivate examinees who are uncertain to consider the correctness of all the remaining options, rather than simply select the option that seems most appropriate. NOTA can extend the range of possible responses to the entire domain.

A third appropriate use pertains to mathematics exams, which often use NOTA as an option. NOTA may encourage more accurate calculation and discourage repeated attempts to find the correct answer (Forsyth & Spratt, 1980; Frary, 1991; Haladyna, 1994; Oosterhof & Coats, 1984; Rimland, 1960; Tollefson & Tripp, 1983).

III. When the answer is calculated or determined by algebraic manipulation, the presence of NOTA may prevent informing the examinee that an incorrect solution is indeed incorrect when they do not find their solution among the options. Without the NOTA option, examinees may rework a problem until a solution consistent with an option is obtained. NOTA may also discourage working backward from the options, because the correct answer may not be present. Finally, NOTA will discourage examinees from choosing approximate answers or simply computing one or two digits of the answer.

Each of the three appropriate uses above can be tested. The results of this meta-analysis provide some empirical evidence to support these postulates. One expectation is that the use of NOTA in each instance above should make the item more difficult, demanding greater cognitive skill of the examinee than items where all options are specified. The results of this study, finding a significant NOTA effect on item difficulty, support each of the three postulates in that expectation. It is interesting to note that while discrimination was slightly reduced, validity was increased through the use of NOTA. Neither of these results were significant; however, suggesting that the use of NOTA in some cases could be very beneficial in terms of increasing the validity of test score interpretation.

Finally, there are at least two cases where NOTA is an *inappropriate* option to be used in MC items. The first is when "best answer" type items are being used. When each option contains some correct content, NOTA should not be used. The correct answer must be exactly correct or NOTA may also be correct. The second case is where the other options exhaust all logical possibilities. Consider the following options: (a) less than 5, (b) exactly 5, (c) more than 5. In such situations, NOTA would not be plausible.

The role of complex alternatives or Type-K items. Avoid the complex, Type-K format, based on the results of this synthesis. However, several areas continue to use the Type-K item, particularly in health sciences and medical licensure and certification exams. Although several researchers have found evidence of clueing in Type-K items, there is some support for the continued use of this format.

Several authors, including those whose studies are used in this meta-analysis, have made suggestions and cautions regarding the uses of Type-K items. First, Type-K items often include

the best possible answer in the correct response plus additional responses that are correct in some cases or less often. Respondents must make a finer distinction to ascertain the *best* answer. This type of decision-making may or may not be appropriate given the level of difficulty of the exam (Nyquist, 1981).

I. When using the best-answer direction, the correct response must really be the best answer. Respondents must have had access to all the information required to select the best answer; outside information or information based on the experience of the item writer must be compared to the information made available to the respondent.

Words such as “often” or “may be” create serious problems when respondents attempt to select the best answer. Nyquist (1981) demonstrated this with two stems: (1) “The clinical symptoms of hepatitis are in some ways similar to and <u>often</u> confused with:” (p. 78), and (2) “In a patient with major motor seizures, status epilepticus <u>may be</u> precipitated by:” (p. 79). We could ask how <u>often</u> is often enough and does <u>may be</u> also mean may be not? At any rate, ambiguous terms distract from accurate measurement, particularly in Type-K items where the correct answer may include the best response as well as other less important, but correct options.

II. Careful unambiguous wording of the stem is always a goal, but even more critical when writing Type-K items.

One difficulty in Type-K item development is determining which set of options to include as a Type-K distractor. It is possible that different kinds of combinations of complex alternatives may have different degrees of effectiveness (Mueller, 1975). A common long-standing belief is that Type-K items are more discriminating and complex, thus they require higher order thinking. However, characteristics that make a conventional MC item more discriminating and require higher order thinking to correctly solve the item are most likely the

same that affect Type-K items. Tripp and Tollefson (1985) found that the content and intended cognitive level of the item make it effective in terms of discrimination and higher order thinking, not the fact that it is Type-K alone.

III. The content and intended cognitive level of the item determine the effectiveness of Type-K items to measure higher order thinking and provide for greater discrimination among examinees than conventional MC items.

The role of incomplete stems. Items in sentence-completion form, where options complete the stem, are endorsed equally with items that have complete-question stems. This synthesis supported such a dual endorsement. Some measurement specialists suggested early on that when the stem is a complete question, it reduces ambiguity in the item (Ebel, 1951; Wesman, 1971). More recently, Eisley (1990) found that the problem scope determined the degree of effect of changing this stem format. He suggested that the stem be constructed such that it defines a single restricted problem (one which could be solved without options) and that the stem be in sentence-completion form. When sentence-completion seems inappropriate for a specific stem, the question format could be used as long as the problem scope is restricted. The problem score is the more important consideration for Eisley.

Schmeiser and Whitney (1972, 1973, 1975a, 1975b) have characterized the sentence-completion form as an item-writing flaw. They also have suggested that the results of changing the stem may be confounded with the characteristics of the individual items.

Schrock and Mueller (1982) suggested that subjects required more time when the sentence-completion stem was truncated more severely, needing more time to determine what the item was really asking. Where the stem is truncated severely, the item writer has reduced the

possibility for determining the correct meaning of the item. Although this was not experimentally manipulated in the Schrock and Mueller study, their contention is based on the degree to which stems were truncated. They argued that sentence-completion and complete-question stems would result in the same item characteristics when the sentence-completion stem contains all of the information of the complete question.

Based on the discussions from the empirical studies of this item format effect, it is difficult to ascertain a set of principals regarding the use of either format. General item-writing rules prevail and this formatting issue may be inconsequential, an idea supported by my synthesis.

I. Items written in either sentence-completion or complete-question format should contain a single problem or restricted problem scope.

II. The item stem, in either format, should contain all necessary information to select the best answer.

The role of negatively worded stems. Based on this synthesis, stems should be stated positively. However, several authors continue to suggest appropriate uses of negatively worded stems. Most clearly, attention should be brought to the negative term to avoid construct-irrelevant error. In addition, directions should be stated with respect to selecting the "best answer" rather than the correct answer, since the correct answer for a negatively worded item is actually the incorrect option.

I. Identify all negative terms by bolding, underlining, using all capital letters, or using some other format for those terms.

Negatively worded stems may require examinees to shift their mental set from the positive to the negative, a task some may fail to do. This is particularly a problem when very few items contain negatively worded stems. Harasym et al. suggested that most medical examinations contain about 11% negatively worded stems. (Although I would suggest avoiding them altogether.) In addition, Terranova (1969) found that the effect of using negatively worded stems depends on the frequency of switching from positive to negative items.

II. Use a number of negative items that is comparable to the number of positive items; e.g., at least 10% but less than 50% of the total test should be negatively worded items--if using negative items.

Harasym et al. (1992) suggested that item content and response format (single versus complex responses) affect scores more than wording of the stem. Harasym et al. (1993) also suggested that knowledge as measured by what is <u>incorrect</u> is not equivalent to knowledge as measured by what is correct. This is similar to the issues raised by Chang (1995) in the use of "connotatively inconsistent" test items, as discussed above. Nyquist (1981) also found that, in most medical school examinations, negatively worded stems were poorly focused to begin with and followed typical patterns, such as (1) "All of the following statements concerning . . . are true except: ," (2) "All of the following are associated with . . . except: ," and (3) "All of the following are characteristics of . . . except: ."

III. Negation be restricted to the case when it is critical for an examinee to know what to avoid or what is <u>not</u> true. As usual, the stem should clearly identify the problem or issue.

Tamir (1993) recounted an interesting story of a visit he made to Australia. While there, he encountered the biology matriculation exam used in the State of Victoria, where the exam was comprised completely of negatively worded items. Their rational for this was that it was better

for students to be exposed to correct rather than incorrect information – responding to the test is a learning experience and by having only one incorrect option per item minimized their exposure to incorrect information.

The role of many distractors. It has been suggested that we use as many options as feasible (Haladyna and Downing, 1989a). This is based on a fair review of the literature. I would support this advice by contributing the concern that in most cases, only two are feasible. Based on this synthesis, MC items should consist of three options, one correct option and two plausible distractors. Anything more is either not feasible or not plausible. Mathematical proof was derived by Tversky (1964) that the use of three-option items maximizes the discrimination, power, and information of a test. Other theoretical work has been done to suggest the advantages of three-option items (Ebel, 1969; Grier, 1975; Lord, 1944, 1977).

I. Less time in needed to prepare two plausible distractors rather than three or four distractors.

II. Less time is required by examinees to read and consider three options rather than four or five options.

III. More 3-option items can be administered within a given time period than 4- or 5-option items, potentially improving content coverage.

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 they eliminate the least plausible distractors, 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 (1985) 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. When Owen and Froman (1987) completed their study of 3- versus 5-option forms, they asked the 114 subjects to vote for their preferred form: 111 voted for the 3-option form, three had no preference, and none voted for the 5-option form.

The role for options of different lengths. Basically, there is none. If any of the rules in this synthesis are widely considered "flaws," this is clearly at the top of the list. This synthesis supports the rule as generally stated: make all options similar in length. The clueing evidenced in correct options of longer length results in faulty item statistics. The tendency for some item writers to make correct options longer to clearly articulate their correctness is problematic and more widespread than we would hope. However, it is usually a flaw found in classroom tests rather than large-scale high-stakes tests.

Nothing more needs to be said: There is no role for options of different lengths.

What is an appropriate model for item difficulty, with respect to item-format effects?

Based on the results of this study, the following are suggested as models for item difficulty, using a 95% confidence interval adjustment:

Item Difficulty = p - 0.035 (NOTA) ± .01

Item Difficulty = p - 0.032 (Negative Stem) ± .02

Item Difficulty = p + 0.016 (Completion Stem) ± .018

Item Difficulty = p + 0.057 (Longer Correct Option) ± .028 (*Not that we would plan to do this*!)

Item Difficulty = p - 0.122 (Type-K) ± .022

When we place the results in this form, it is more striking to notice which effects are "reliable" or within a reasonable confidence interval (which may or may not include zero). But the basic idea is that we could add or subtract the effect from the original item difficulty if the format of the item was to take on one of the alterations investigated in this synthesis. Similar equations could be written based on effects for item discrimination, test reliability, and test validity. Of course we would not tolerate the use of equations for item difficulty if they simultaneously resulted in tests of lower reliability or score interpretations with less validity.

Future Directions

Item analysis (gathering empirical data on item performance, usually in the form of item difficulty and item discrimination) is a critical step in test development (Allen & Yen, 1979). 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.

Mehrens and Lehmann (1991) suggested careful use of item-analysis data. Item-analysis data are tentative and based on a given sample. In addition, item selection should not be based solely on item-analysis data. Items should have item-analysis results that are appropriate for the intended purposes of the test. Finally, content sampling of the test must be maintained. This may result in the selection of a few very easy or very difficult items. Making test items more difficult could result in improved differentiation among examinees if using a test of moderate difficulty.

Downing and Haladyna (1997) outlined eleven item development activities and proposed types of validity evidence needed to assure the quality of test item development. One area of

validity evidence directly addresses the use of item-writing principles. To secure validity evidence in this area, Downing and Haladyna suggest that the 43 rules in their taxonomy (Haladyna & Downing, 1989a) could serve as a literal checklist. They argued that there should be documented evidence of adherence to the relevant item-writing principles and a rationale should be documented regarding the use of particular item formats. For example, test developers (item-writers) should provide their rationale for the number of options used in multiple-choice items.

The question remains as to whether other format characteristics not included in this synthesis have a significant effect on item and test statistics. Future meta-analyses of item format effects would improve our understanding of these effects appreciably if they included several formats simultaneously, where analysis is conducted in a multivariate space, and where the distributions of effects are more completely derived. A more complete model of item format effects on item difficulty could then be developed. For example, in one study included in this analysis, NOTA was included as an additional option, increasing the number of options in those items by one. Knowing the simultaneous effect of an additional option, when that option is NOTA, on item difficulty would help us understand the "purified" effect of NOTA, and vice versa. A simple model might look like this:

$$\text{Item Difficulty} = \mu + \alpha_1(\text{NOTA}) + \alpha_2(\text{\# Options}) + \alpha_3(\text{Negation}) + \alpha_4(\text{Closed stem}) + \alpha_5(\text{Type-K format}) + \alpha_6(\text{Consistent option length}) + \varepsilon.$$

The effects of using "none of the above" as an option (NOTA), increasing the number of options (# Options), using negation or negative terms in the stem (Negation), and using a closed stem (Closed stem), among other format characteristics, could all be specified and tested in such

a model. Each multiple-choice item would either have or not have each or any number of these format characteristics. Based on a small review of studies manipulating more than one format characteristic at a time and also testing for interaction effects, item-format effects did not interact (Crehan & Haladyna, 1991; Crehan, Haladyna, & Brewer, 1993; Dudycha & Carpenter, 1973). These results suggest that item-format effects act independently. In addition, several studies reported no interaction between item format effect and achievement level of examinees.

One area of inquiry that has been absent from this line of research is that regarding potential effects of individual differences. Rarely did the studies included in this meta-analysis examine potential gender, race, or age differences in the effects of item formats. A second would be to more clearly describe the cognitive requirements of each item. In several cases, research has demonstrated the important influence of the specific content and cognitive requirements of each item in terms of interpreting the effect of changing the item format.

Another area that has been limited is examining the effect of NOTA when it is used as the correct response compared to its use as a distractor. Studies conducted by Tollefson et al. (1983; Tripp & Tollefson, 1985, 1986, 1987), consistently reported larger effects on item difficulty when NOTA was used as the correct response, making items substantially more difficult. Haladyna (1994) suggested that there are three good reasons to expand distractor analysis: (1) each distractor should be useful; (2) useful distractors contribute to more effective scoring with polytomous scoring, improving test sore reliabilities; and (3) information about distractors could potentially provide misconception information to instructors and students.

Similarly, clarifying the role of the methods used to delete or create options in the studies varying the number of options-per-item will help specify potential effects. Budescu and Nevo (1985) found that the impact of changing the number of options-per-item depended on the

method used to delete options in creating multiple forms. This might allow us to more clearly articulate Haladyna and Downing's (1989b) recommendation to use as many plausible distractors as possible.

Information that increases our understanding of multiple-choice items and tests will improve our ability to measure student achievement and other constructs. Improved information may lead to improved item-writing, improved test-design, better measures of student achievement, more appropriate score interpretation, some combination of the above, or hopefully all of the above.

References

*Note: * indicates those references which were included in the quantitative synthesis.*

Aamodt, M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management, 21*(2), 151-160.

Ace, M. C., & Dawis, R. V. (1973). Item structure as a determinant of item difficulty in verbal analogies. *Educational and Psychological Measurement, 33,* 143-149.

Albanese, M. (1982). Multiple-choice items with combinations of correct responses. *Evaluation and the Health Professions, 5*(2), 218-228.

Albanese, M. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practices, 12*(1), 28-33.

Albanese, M. A., Kent, T. H. & Whitney, D. R. (1979). Cluing in multiple-choice test items with combinations of correct responses. *Journal of Medical Education, 54*(12), 948-950.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

American Psychological Association, American Educational Research Association, & National Council of Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

*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.

Becker, B. J., & Fahrbach, K. (1995). Unpublished course packet for CEP 936, College of Education, Michigan State University.

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* (323-357). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving.* New York: Springer-Verlag New York Inc.

*Board, C., & Whitney, D. R. (1972). The effect of selected poor item-writing practices on test difficulty, reliability and validity. *Journal of Educational Measurement, 9*(3), 225-233.

Boynton, M. (1950). Inclusion of none of these makes spelling items more difficult. *Educational and Psychological Measurement, 10*, 431-432.

Brennan, R. L. (1997, March). Measurement (mis)conceptions at the intersection of theory and practice. Presidential Address at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

*Budescu, D. V., & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement, 22*(3), 183-196.

Carter, K. (1986). Test wiseness for teachers and students. *Educational Measurement: Issues and Practices, 5*(4), 20-23.

Casler, L. (1983). Emphasizing the negative: A note on the not in multiple-choice questions. *Teaching of Psychology, 10*(1), 51.

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple-choice tests in chemistry. *Journal of Chemical Education, 61*, 613-615.

*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.

Chang, L. (1995). Connotatively inconsistent test items. *Applied Measurement in Education, 8*(3), 199-209.

Chase, C. (1964). Relative length of option and response set in multiple choice items. *Educational and Psychological Measurement, 24*(4), 861-866.

*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*(3), 144-145.

*Crehan, K. D., & Haladyna, T. M. (1991). The validity of two item-writing rules. *The Journal of Experimental Education*, *59*(2), 183-192.

*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*(1), 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, in press). Test item development: Validity evidence from qualitative assurance procedures. *Applied Measurement in Education.*

*Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item formats on item discrimination and difficulty. *Journal of Applied Psychology, 58*, 116-121.

*Duncan, R. E. (1983). An appropriate number of multiple-choice item alternatives: A difference of opinion. *Measurement and Evaluation in Guidance, 15*(3), 283-292.

*Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and reliability as a function of a selected multiple-choice item construction principles. *Educational and Psychological Measurement, 19*(2), 171-179.

Ebel, R. L. (1951). Writing the test item. In E. F. Linquist (Ed.), *Educational measurement* (1st ed., 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.

*Eisley, M. E. (1990). *The effect of sentence form and problem scope in multiple-choice item stems on indices of test and item quality*. Unpublished doctoral dissertation, Brigham Young University.

*Evans, W. (1984). Test wiseness: An examination of cue-using strategies. *Journal of Experimental Education, 52*(3), 141-144.

*Forsyth, R. A., & Spratt, K. F. (1980). Measuring problem solving ability in mathematics with multiple-choice items: The effect of item format on selected item and test characteristics. *Journal of Educational Measurement, 17*(1), 31-43.

*Frary (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education, 4*(2), 115-124.

Gardner, P. L. (1970). Test length and the standard error of measurement. *Journal of Educational Measurement, 7*(4), 271-273.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: John Wiley and Sons, Inc.

Green, K. E. (1981). *Identification and investigation of determinants of multiple-choice test item difficulty*. Unpublished doctoral dissertation, University of Washington.

Green, K. E. (1983). Subjective judgment of multiple-choice item characteristics. *Educational and Psychological Measurement, 43*(2), 563-570.

Green, K. E. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement, 44*(3), 551-561.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*(2), 109-112.

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items. *Evaluation and the Health Professions, 17*(1), 123-126.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Haladyna, T. M., & Downing, S. M. (1985, April). *A quantitative review of research on multiple-choice item writing*. Paper presented at the annual meeting of the AERA, Chicago, IL.

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*(1), 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 2*(1), 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.

Harasym, P. H., Doran, M. L., Brant, R., & Lorscheider, F. L. (1993). Negation in stems of single-response multiple-choice items. *Evaluation and the Health Professions, 16*(3), 342-357.

*Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation and the Health Professions, 15*(2), 198-220.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88*, 359-369.

*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*(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*(1), 63-68.

*Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple-choice items. *Educational and Psychological Measurement, 25*(1), 117-126.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

*Huntley, R. M., & Plake, B. S. (1988). *An investigation of multiple-response-option multiple-choice items: Item performance and processing demands*. Unpublished manuscript. (ERIC Document No. ED 306 236)

Knowles, S. L., & Welch, C. A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using none-of-the-above. *Educational and Psychological Measurement, 52*, 571-577.

Kolstad, R. K. Briggs, L. D. Bryant, B. B. & Kolstad, R. A. (1983). Complex multiple-choice items fail to measure achievement. *Journal of Research and Development in Education, 17*(1), 7-11.

Kolstad, R. K., Briggs, L. D., & Kolstad, R. A. (1984). The application of item analysis to classroom achievement tests. *Education, 105*(1), 70-72.

*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. (1991). The option "none of these" improves multiple-choice test items. *Journal of Dental Education, 55*(2), 161-163.

*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*(2), 4-8.

Kramer, G. A., & Smith, R. M. (1993, April). *A confirmatory analysis of factors influencing the difficulty of form-development items*. Paper presented at the annual meeting of the AERA, Atlanta, GA.

Kromrey, J. D., & Bacon, T. P. (1992, April). *Item analysis of achievement tests based on small numbers of examinees*. Paper presented at the annual meeting of the AERA, San Francisco, CA.

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. (1953). An examination of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*, 57-76.

Lord, F. M. (1977). Optimal number of choices per item -- A comparison of four approaches. *Journal of Educational Measurement, 14*(1), 33-38.

Marzano, R. J., & Jesse, D. M. (1987). *A study of general cognitive operations in two achievement test batteries and their relationship to item difficulty*. Report of the Mid-continent Regional Educational Laboratory, Aurora, CO.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15-29.

*McMorris, R. F., Brown, J. A., Snyder, G. W., & Pruzek, R. M. (1972). Effects of violating item construction principles. *Journal of Educational Measurement, 9*(4), 287-295.

Mehrens, W. A. (1997, April). *The consequences of consequential validity*. Presentation at the Measurement and Quantitative Methods Colloquium, Michigan State University, East Lansing, MI.

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

Messick, S. (1994). The once and future issues of validity: Assessing meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-48). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Millman, J. (1978). *Determinants of item difficulty: A preliminary investigation* (Center for the Study of Evaluation, Report No. 114). Los Angeles: UCLA Graduate School of Education.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.335-366). Washington, DC: American Council on Education.

*Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational and Psychological Measurement, 35*, 135-141.

Myers, C. T. (1962). The relationship between item difficulty and test validity and reliability. *Educational and Psychological Measurement, 22*(3), 565-571.

*Nyquist, J. G. (1981). *A comparison of the quality of selected multiple-choice item types within medical school examinations*. Unpublished doctoral dissertation, Michigan State University.

*Oosterhof, A. C., & Coats, P. K. (1984). Comparison of difficulties and reliability of quantitative word problems in completion and multiple-choice item formats. *Applied Psychological Measurement, 8*(3), 287-294.

Osterlind, S. J. (1989). *Constructing test items*. Boston: Kluwer Academic Publishers.

*Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement, 47*(2), 513-522.

Pollack,I., & Ficks, L. (1954). Information of elementary multidimensional auditory displays. *Journal of the Acoustical Society of America, 26*(2), 155-158.

Pyrczak, F. (1972). *Objective evaluation of the quality of multiple-choice test items* (Report of the National Center for Educational Research and Development). Washington, DC: U. S. Department of Health, Education, and Welfare.

Pyrczak, F. (1973). Validity of the discrimination index as a measure of item quality. *Journal of Educational Measurement, 10*(3), 227-231.

*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*(4), 305-310.

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., & 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., & 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.

*Rich, C. E., & Johanson, G. A. (1990, April). *An item-level analysis of "none of the above."* Paper presented at the annual meeting of the AERA, Boston, MA.

*Rimland, B. (1960). The effects of varying time limits and of using Right answer not give in experimental forms of the U.S. Navy Arithmetic Test. *Educational and Psychological Measurement, 20*(3), 533-539.

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.

Rossi, J. S., McCrady, B. S., & Paolino Jr., T. J. (1978). A and B but not C: Discriminating power of grouped alternatives. *Psychological Reports, 42*(2), 1346.

*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.

*Schmeiser, C. B., & Whitney, D. R. (1973). *Effect of selected poor item-writing practices on test difficulty, reliability and validity: A replication*. Unpublished manuscript. (ERIC Document No. ED 075 498)

*Schmeiser, C. B., & Whitney, D. R. (1975a, April). *The effect of incomplete stems and "none of the above" foils on test and item characteristics*. Paper presented at the annual meeting of the NCME, Washington, DC.

*Schmeiser, C. B., & Whitney, D. R. (1975b). Effect of two selected item-writing practices on test difficulty, discrimination, and reliability. *Journal of Experimental Education, 43*(3), 30-34.

*Schrock, T. J., & Mueller, D. J. (1982). Effects of violating three multiple-choice item construction principles. *Journal of Educational Research, 75*(5), 314-318.

Scott, P. D. (1980). *A study of mathematical problem solving in a multiple-choice format*. Unpublished doctoral dissertation, The Florida State University.

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. 231-244). New York: Russell Sage Foundation.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

*Statman, S. (1988). Ask a clear question and get a clear answer: An inquiry into the question/answer and sentence completion formats of multiple choice items. *System, 16*(3), 367-376.

*Strang, H. R. (1977). The effects of technical and unfamiliar options on guessing on multiple-choice test items. *Journal of Educational Measurement, 14*(3), 253-260.

*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*(5), 425-429.

Tamir, P. (1991). Multiple choice items: How to gain the most out of them. *Biochemical Education, 19*(4), 188-192.

*Tamir, P. (1993). Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation, 19*(3), 311-325.

*Terranova, C. (1969). *The effects of negative stems in multiple-choice test items*. Unpublished doctoral dissertation, State University of New York at Buffalo. (30, 2390A)

*Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using the none of the above and one correct response options. *Educational and Psychological Measurement, 47*(2), 377-383.

*Tollefson, N., & Chen (1986, October). *A comparison of item difficulty and item discrimination of multiple-choice items using none of the above options*. Paper presented at the annual meeting of the Midwest AERA, Chicago, IL.

*Tollefson, N., & Tripp, A. (1983, April). *The effect of item format on item difficulty and item discrimination*. Paper presented at the annual meeting of the AERA, Montreal, Quebec.

*Trevisan, Sax, Michael (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, Sax, Michael (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement, 54*(1), 86-91.

*Tripp, A., & Tollefson, N. (1985). Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options? *Journal of Nursing Education, 24*(3), 92-98.

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

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*(1), 1-21.

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

*Weiten, W. (1982). Relative effectiveness of single and double multiple-choice questions in educational measurement. *Journal of Experimental Education, 51*(1), 46-50.

*Weiten, W. (1984). Violation of selected item construction principles in educational measurement. *Journal of Experimental Education, 52*(3), 174-178.

*Wesman, A. G., & Bennett, G. K. (1946). The use of ‘none of these’ as an option in test construction. *Journal of Educational Psychology, 37*, 541-549.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed). Washington, DC: American Council on Education.

*Williams, B. J., & Ebel, R. L. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. *The 14th Yearbook of the NCME,* 63-65.

*Williamson, M. L., & Hopkins, K. D. (1967). The use of none-of-these versus homogeneous alternatives on multiple-choice tests: Experimental reliability and validity comparisons. *Journal of Educational Measurement, 4*(2), 53-58.

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