EVALUATION IN EDUCATION AND HUMAN SERVICES

Constructing Test Items:
Multiple-Choice, Constructed-Response, Performance, and Other Formats

Second Edition

Steven J. Osterlind

Kluwer Academic Publishers

Constructing Test Items:

Multiple-Choice, Constructed-Response, Performance, and Other Formats

Second Edition

Evaluation in Education and Human Services

Other books in the series:

Smith, M.:
Evaluability Assessment
Ayers, J. and Berney, M.:
A Practical Guide to Teacher Education Evaluation
Hambleton, R. and Zaal, J.:
Advances in Educational and Psychological Testing
Gifford, B. and O'Connor, M.:
Changing Assessments
Gifford, B.:
Policy Perspectives on Educational Testing
Basarab, D. and Root, D.:
The Training Evaluation Process
Haney, W.M., Madaus, G.F. and Lyons, R.:
The Fractured Marketplace for Standardized Testing
Wing, L.C. and Gifford, B.:
Policy Issues in Employment Testing
Gable, R.E.:
Instrument Development in the Affective Domain (2nd Edition)
Kremer-Hayon, L.:
Teacher Self-Evaluation
Payne, David A.:
Designing Educational Project and Program Evaluations
Oakland T. and Hambleton, R.:
International Perspectives on Academic Assessment
Nettles, M.T. and Nettles, A.L.:
Equity and Excellence in Educational Testing and Assessment
Shinkfield, A.J. and Stufflebeam, D.L.:
Teacher Evaluation: Guide to Effective Practice
Birenbaum, M. and Dochy, Filip J.R.C.:
Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge
Mulder, M., Nijhof, W.J., Brinkerhoff, R.O.:
Corporate Training for Effective Performance
Britton, E.D. and Raizen, S.A.:
Examining the Examinations
Candoli, C., Cullen, K. and Stufflebeam, D.:
Superintendent Performance Evaluation
Brown, S.M. and Seidner, C.J.:
Evaluating Corporate Training: Models and Issues

Constructing Test Items:

Multiple-Choice, Constructed-Response, Performance, and Other Formats

Second Edition

Steven J. Osterlind
University of Missouri-Columbia

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47535-9
Print ISBN: 0-7923-8077-0

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

Contents

Chapter 1

What Is Constructing Test Items?

INTRODUCTION

Constructing test items for standardized tests of achievement, ability, and aptitude is a task of enormous importance—and one fraught with difficulty. The task is important because test items are the foundation of written tests of mental attributes, and the ideas they express must be articulated precisely and succinctly. Being able to draw valid and reliable inferences from a test's scores rests in great measure upon attention to the construction of the items or exercises that comprise it. If a test's scores are to yield valid inferences about an examinee's mental attributes, its items must reflect a specific psychological construct or domain of content. Without a strong association between a test item and a psychological construct or domain of content, the test item lacks meaning and purpose, like a mere free-floating thought on a page with no rhyme or reason for being there at all. Interpretability of a test's scores flow directly from the quality of its items and exercises.

Concomitant with score interpretability is the notion that including only carefully crafted items on a test is the primary method by which the skilled test developer reduces unwanted error variance, or errors of measurement, and thereby increases a test score's reliability. In one very complete sense, the aim of this entire book is to increase test constructor's awareness of this source for measurement error, and then to described methods for identifying and minimizing it during item construction and later review.

The task of constructing good test items is difficult because writing precisely and succinctly is challenging. The intended meaning must be clear. Additionally, the grammar, spelling, punctuation, and syntax must be correct and exact. Since many test items are no more than a single sentence, there is often little opportunity to garner meaning from context. Because good writing is difficult, it is distressingly easy for a test-item writer to inadvertently convey hints, biases, prejudices, opinions, or confusing information.

A further reason why constructing good test items is difficult is that the task challenges the writer to be creative. Crafting test items requires more than employing good compositional skills. Imaginative and novel ways of expressing ideas can frequently be useful in test item construction. And, creativity includes an intuitive appreciation of how a particular test item may be perceived by examinees. Such rich understanding of test items will assist one in gaining a “sixth sense” about constructing them.

Recent research has demonstrated that the wording and format of test items can greatly influence the psychological perspective that the examinee brings when considering a response (e.g., Wolf, et al, 1995). Anxiety, motivation, and ultimately, performance are affected by the item’s wording and format. As the craft of preparing good items grows ever-more sophisticated, attending to the examinee’s psychological perspective becomes correspondingly more important. This fact was made manifestly clear when in late 1995 a task force of eminent psychologists was convened by the Board of Scientific Affairs of the American Psychological Association. Their report, entitled “Intelligence: Knowns and Unknowns” (Task Force, 1995), identifies many important, known facts about intelligence and provides test-item writers with an invaluable resource for understanding the psychological perspective of examinees.

An additional reason why constructing test items is important is that there are manifold technical considerations in item preparation that can influence its quality. Features such as employing an appropriate item format, the level of vocabulary, determining the optimal number of response alternatives, and whether to permit negatively worded items (e.g., “Which is not . . .”), are only a few such considerations. The writer must attend to them with care and skill.

Still further, persons involved in assessment are keenly aware of the increased attention given to alternative formats for test items in recent years—item formats other than multiple-choice or true/false, or matching. These alternative formats are, with increasingly regularity, the sole format for items on a test. Yet, in many writers’ zeal to be “curriculum-relevant” or "authentic" or "realistic", the items are often developed seemingly without conscious thought to the interpretations that may be garnered from them. This book argues that the format for such alternative items and exercises, too, requires rigor in their construction. In fact, it is the author's hope that this book may draw attention to this problem, and even offer some solutions, as one chapter is devoted to just these alternative formats.

The perils of writing test items without adequate forethought are great. Decisions about persons, programs, projects, and materials are often made on the basis of test scores. If a test is made up of items haphazardly written by untutored persons, the scores could be effected, and validity impacted by the resulting erroneous decisions. Such decision errors can sometimes have serious consequences for individuals. Incorrect levels of achievement or performance may be inferred. Programs, projects, and materials could be misjudged. Obviously, such a disservice to examinees as well as to the assessment or evaluation process should be avoided if at all possible.

Although there is abundant literature explaining measurement theory, test construction, and analysis of test results (see Anastasi, 1988; Cronbach, 1984; Ebel, 1979; Gulliksen, 1950; Hambleton & Swaminathan, 1985; Lord & Novick, 1968; Nunnally, 1978; Thorndike, 1982; Weiss & Davidson, 1981; Wright & Stone, 1979; and many others), there is woefully little information about planning, designing, and writing test items themselves. Cronbach observed in 1970 that "the design and construction of achievement test items have been given almost no scholarly attention" (p. 509). And Bormuth (1970), remarking on the lack of concern for information about constructing test items, noted that most writers of test items have only their intuitive skills to rely upon. Nitko (1984a) lamented the dearth of item-writing research with gentle humor:

Elder item writers pass down to novices lists of rules and suggestions which they and their item-writing forefathers have learned through the process of applied art, empirical study, and practical experience, (p. 204)

Even more disturbing is the conclusion of Haladyna and Downing (1989) after their scrutiny of 46 authoritative textbooks and other sources in the educational measurement literature: ". . . the body of knowledge about MC item writing seems not to be particularly well established, yet the practice of item writing is extensive and certainly warrants more scholarly attention than it appears to have received."

Researchers—notably, Millman and Greene (1989), Roid and Haladyna (1982), and Wesman (1971)—have similarly commented about the lack of significant research or practical guidance on this subject. Wood (1977), in a report titled "Multiple Choice: A State of the Art Report," offers a comprehensive review of topics related to multiple-choice testing but offers little guidance about how to actually construct test items. And, a 1984 survey of topics selected to be of interest to the membership of the prestigious National Council on Measurement in Education did not even include anything related to the problems associated with constructing test items, although it did cite issues related to test construction and even writing-skills assessment specifically (Berk and Boodoo, n.d.).

The neglect of research into the field is reinforced by the fact that none of the articles in the most recent edition of Handbook of Educational Psychology (Berliner & Calfee, 1996), arguably the single major reference source chronicling recent advances in the field, did not include any description of crafting test items. And, the publication of an international handbook on educational research, methodology, and measurement devotes only a scant, four-page article (out of over 800 pages) to item-writing techniques (Herman, 1988). Haladyna (1994) suggests two reasons for the paucity of credible research into item construction, including, 1) that the terms of cognitive psychology are not adequately defined so as to allow meaningful interpretation, and 2) the absence of a validated taxonomy for identifying and classifying complex cognitive behavior. He concludes that, "Item writing in the current environment cannot thrive due to the existence of these two barriers" (p. 185).

A sad testimony to the widespread neglect of this important part of testing is the fact that the single most popular introductory textbook to the field of psychological testing devotes only three paragraphs to effective item writing. And, the introductory textbook is not alone in its inadequate coverage of the topic. A simple review of nine primary texts in educational psychology—all published since 1990 and all citing tests and measurement in their title—reveals that none give more than rudimentary coverage to constructing test items.

Considering that test items are the backbone of most assessment instruments, the dearth of advice about how to construct them is remarkable. Regretfully, it seems that Ebel's 1951 comment on the insufficiency of relevant research and guidance is still applicable today: "The problems of item writing have not received the attention they deserve in the literature on testing" (p. 188). Haladyna (1994), too, suggests that the problem has not improved very much over the years when he echos Ebel's lament more than forty years later: "Item writing lacks the rich theoretical tradition that we observe with statistical theories of test scores" (p. 193). Test-item writers are routinely left to their own devices because there is no developed theory to undergird item writing, nor is there a comprehensive resource identifying the distinctive features and limitations of test items, the function of test items in measurement, or even basic editorial principles and stylistic guidelines. Further, as item development becomes necessary for modern psychologically-based instruments, such as is attempted in many performance-based and other constructed-response instruments, this deficiency for item writing becomes even more pronounced.

The little guidance that is available to assist in constructing test items is frequently perfunctory or trivial, often consisting of a list of "dos" and "don'ts." A review of many of these lists reveals that they are predominantly comprised of idiosyncratically selected rules for achieving good writing (e.g., "avoid wordiness," "focus each item on a single idea," etc.). Further, typical lists of item-writing rules intermix with basic suggestions for good writing certain technical and editorial guidelines, such as "Avoid 'All of the Above,'" or "Keep options in a logical order." Although particular rules offered in such lists may be acceptable, a simple list neither captures the complexity of the task nor conveys why certain features are requisite to producing test items of merit. Even when psychometric analysis is employed as apart of such criterion (e.g., Frary, 1995), such lists are of dubious utility.

This book does not suggest one list of things to do when preparing test items; rather, the emphasis is on understanding criteria for meritorious test items, as well as recognizing the importance of good writing generally for this type of technical writing test-item construction and learning how it may be achieved. Working from this viewpoint, the skilled item writer will be both cognizant of good writing and sufficiently informed to employ whatever rule of writing or editorial style is appropriate to the particular item he or she is preparing. Two chapters are devoted specifically to communicating editorial rules for test items. Another chapter is devoted exclusively to performance-type items.

Lest there be confusion on the point of not citing lists, one final comment. There is nothing wrong with lists of item-writing rules; however, I submit that it is more important to understand criteria for meritorious items as well as to stress principles of good writing generally and to learn the specific editorial rules for this kind of technical writing than to attempt to identify just some particular rules that may apply to some test items in a few circumstances. Certainly to list all applicable rules of writing style or of editorial mechanics for test items would be an enormously long compilation and of questionable utility.

WHAT THIS BOOK IS ABOUT

Four Major Issues in Item Construction

Constructing test items is a comprehensive field of endeavor which may be categorized by particular issues. This book addresses major issues included in constructing test items by focusing on four ideas. First, it describes characteristics and functions of test items. Characteristics of test items involve classifying and describing test items by various item formats, that is, the depiction of test items as multiple-choice, true-false, matching, or some other type regardless of whether for they are intended for traditional or performance-types tests. It also includes problems of definition, terminology, and identification of relevant assumptions. Conjoined with characteristics of test items is an understanding of the various functions they serve in measurement, along with an awareness of their limitations and some familiarity with alternatives to test items in measurement. While this information is necessary background to constructing good test items, it is precursory to actually writing them.

A second feature of this book is the presentation of editorial guidelines for writing test items in all of the commonly used item formats, including for test of constructed-response formats and performance tests. Editorial guidelines are prescriptive rules for style and form. They dictate the placement of punctuation, writing style, and many test-item protocols, such as where and how to place directions to test items, or when to use boldface type or italics.

The practice of measuring human attributes and capabilities by means of test items is so common that style rules for writing test items in the various formats should be articulated, standardized, and accepted throughout the industry. There is a greater likelihood for a test item to do whatever it is intended to do if its conception and writing follow prescribed rules. Currently, there does not exist such a comprehensive or prescriptive set of editorial guidelines for writing test items. Clearly, this kind of guidance is needed.

A third aspect of this book is the presentation of methods for determining the quality of test items. Determining the quality of test items may be categorized into two interdependent issues: 1) procedures for gauging the proper content for test items, which revolve around concerns of validity and 2) procedures for examining test items for either random errors or systematic bias, which reflect considerations of reliability. Both of these issues are addressed by judgmental procedures as well as statistical models. The methods described can be applied to making judgments about test items written by others as well as to test items written by the reader.

A fourth component of this book is the presentation of a compendium of important issues about test items. Some examples of issues discussed in this compendium are procedures for ordering items in a test, ethical and legal concerns for using copyrighted test items, item scoring schemes, computer-generated items, and more. A compendium of subjects important to constructing test items could be very large, but this one is arbitrarily confined to cover only a few topics of paramount importance.

Type of Items Addressed

The issues discussed in this book are intended for test items that will be used in both standardized tests and many teacher-made tests. Standardized tests are tests whose initial construction, as well as conditions for administration and scoring, have a uniform procedure so that the scores yielded by the measure may be interpreted in a consistent manner from one administration to the next (Ebel & Frisbie, 1986; Mehrens & Lehmann, 1987; Wiersma & Jurs, 1985). This can include both tests made up of traditional multiple-choice items and tests comprised of constructed-response formats or performance exercises. Teacher-made tests are typically not made according to specific and uniform procedures and the results from various administrations of a teacher-made test are difficult to compare. Regardless of the differences between standardized and teacher-made tests, the quality of the items is important to all kinds of tests.

Also, the material in this book applies to a variety of tests, regardless of whether they are administered to groups of examinees or to an individual. Such standardized or teacher-made tests of achievement, ability, or aptitude can be administered in a wide array of situations, including large- and small-scale assessment programs, clinical testing, educational and psychological testing in schools, tests used in counseling, employment testing, as well as professional and occupational licensing and certification testing. Also included are tests used in the evaluation of educational programs, projects, and materials. Additionally, the information in this book may be suited to tests used in special ways, as for example, testing people who have handicapping conditions, or testing linguistic minorities.

Furthermore, the information presented in this book describes test items as they may be used in tests of achievement, ability, or aptitude. Heavy emphasis, however, is given to test items found in achievement tests because achievement tests, as the most common type of test, are used for assessing many times the number of examinees tested with ability or aptitude measures. Such tests are the rule of the land in both school settings as well as in employment or certification and licensing testing.

Finally, this book does not differentiate between items that may be used in tests of psychological assessment and those incorporated into instruments designed for educational measurement. A word of explanation may be needed to clarify this point. Most test developers, psychometricians, and psychologists make a distinction between the assessment of psychological constructs and measuring educational achievement. For example, Messick (1980) wrote an influential paper emphasizing the importance of gathering evidence for valid interpretations in differing kinds of assessments, and urged persons to consider to consider both "evidential" and "consequential" basis for test interpretation and test use. These are useful distinctions for guiding one in judging the ethical grounds for a test's application as well as appraising potential social consequences of testing; however, the items to be included in these differing assessments are not different in kind. Items for both types of assessment instruments require the same degree of care and technical skill in their construction. Hence, the information presented in this book on constructing items will apply to either type of assessment.

Types of Items Not Covered

Since the description so far has been of what is included in this book, it seems logical to also cite some kinds of test items and tests and measurement issues that are not covered. To begin, this book does not cover assessment done by essay, or writing sample. There are many and varied considerations when developing the essay prompt as well as an array of measurement problems to be dealt with in scoring essays; however, such issues are not addressed in this book. Also, some measures of personality and interest require questioning strategies beyond the scope of the material in this book. For example, this book does not directly address interview strategies, selfreport measures, semantic differential, or Likert and Likert-type scales (i.e., scales that present a range of responses from "strongly agree" to "strongly disagree").

Further, many types of exercises used in intelligence testing are not included in this book, particularly questions and situations that involve inductive reasoning. Inductive reasoning is the ability to apply specific experiences to general rules. An example of an inductive-reasoning prompt may be the statement that a furry, four-legged animal that says "meow" is a cat, therefore, any furry, four-legged creature that makes the "meow" sound is a cat. These kinds of problems are commonly expressed as analogies. This book does not specifically address analogies.

Finally, there are manifold issues in measurement generally that are indirectly related to constructing test items but are not specifically addressed in this book. Some of these issues may contribute in some way to constructing good test items, but one simply cannot include everything about such a large topic as constructing test items in one book.

Criteria for Material Included

Two criteria were used in deciding whether to include a particular issue in this book. The first criterion was whether the issue related directly to test items as individual entities. Issues dealing with items assimilated into an entire test (e.g., generalizability of items, scaling items for various interpretations, etc.) are not included here. A second criterion for deciding what to include was the utility of the information. All of the information included in this book was selected because it has some potential for application in actual construction of test items. Sometimes this potential for application is very direct, such as when and how to incorporate a graphic in a test item; at other times, the information is more supportive and will enable one to understand more about test items generally, such as becoming aware of their proper role in measurement.

MAJOR PURPOSES OF THIS BOOK

Goal for This Book

The primary goal for this book is to contribute to the improvement of tests and measurement by aiding good test-item construction. It is hoped that this goal is accomplished in a number of ways. First, this book can make a significant contribution to the field by presenting complete and up-to-date information about test items and how to construct them. The information provided is of two types: information derived from the work of others and information that is original.

The review of the relevant literature will help in providing information by critiquing and synthesizing the best of what is currently known about test items and their construction. It will also serve as documentation of recent advances in knowledge about test items, item-writing technologies, and item-writing methods for commonly used types of tests. Nowhere else in the literature does there exist such a comprehensive review of information important to constructing test items.

Additionally, the author contributes some new ideas about constructing test items. These new ideas are not mere guesses and speculation about strategies; rather, they represent useful ideas and successful item-writing techniques derived from the author's experience in planning and constructing items for a wide variety of tests.

Standardizing Editorial Guidelines for Writing Items

A principal feature of this book is to prescribe a uniform editorial style for particular test-item formats and a rationale for doing so. As far as the author is aware, nowhere else is such a suggested set of prescriptions put forth. Hence, this book represents a suggestion to test-item writers, test developers, and other interested persons to consider the development of a set of editorial guidelines for constructing test items as an industry standard.

Standardization will not stifle creativity in constructing test items; rather, it will provide a coherent framework for proven strategies as well as allow for new and untried approaches. This set of principles and standards for constructing test items will contribute significantly to the principal goal for this book.

Other Purposes

Still another major purpose for this book is to provide instruction in techniques and strategies for judging the quality of test items. An item merely written is unfinished; it should then be subjected to scrutiny to determine its worth. Poor test items should be discarded, salvageable items repaired, and good ones offered forth on tests. Such a probe into particular test items should be guided by specific criteria for good test items, which this book offers. This close examination may be done by a variety of means, including subjective (but not arbitrary) reviews by knowledgeable persons making informed judgments about whether a particular test item meets the criteria for good test items. And, of course, there are manifold statistics, statistical procedures, and research methods available to lend assistance in the review of test items. Some of these will be described as well.

Yet another purpose for this book is to serve as both a text for instructing students and practitioners in constructing good test items as well as a reference source for a wide array of persons who have a need for information about test items and allied issues. Currently, neither a text nor a comprehensive reference source exists for the important enterprise of constructing test items. Students may gain from the book's organization into chapters of increasing sophistication, guiding them from an initial understanding of the necessary considerations when constructing test items and why they are important, to learning the mechanics of how to construct the items, to becoming familiar with a host of relevant issues. Practitioners who may require a reference source will appreciate the book's thoroughness. Serving the need for a text and reference source in this field is an important purpose for this book.

WHY THIS INFORMATION IS IMPORTANT

Unquestionably, the principal reason the information presented in this book is important is because it may contribute to better decision making, the *raison d'etre* for any testing. Knowing about constructing test items will assist persons in being better informed and, hence, they will be more likely to make reasoned decisions, either as test developers, as test takers, or as test users. Of course, evaluative judgments about persons or programs made on the basis of scores derived from standardized examinations are common. Such decisions occur regularly in schools, counseling situations, business and industry, in licensing and certification programs, and elsewhere. It is estimated that annually over 25 million Americans take a standardized examination (Merrow, 1997). In each instance, the examination was administered so that a decision of some kind could be made. The consequences of decisions made on the basis of test scores can range from innocuous to dramatic. But, regardless of how frequently such decisions occur or the significance of the consequences, it is ethically responsible to promote the notion that test items should be constructed by informed persons.

Another reason the information presented in this book is important is that standardization is called for in this burgeoning field. There are simply so many tests and so many people involved in developing them that it is not realistic to hope that quality test items will result each time, independently, without some generally accepted standards.

Haphazard approaches to constructing test items may or may not produce good items. More likely, lack of uniformity can lead to confusion about relevant information, with a resulting diminution in quality. Buros (1972), lamenting the generally poor quality of tests produced, has remarked that "at least half of the tests currently on the market should never have been published" (p. xxvii). One hopes that with accurate, readily available information, the problem expressed in Buros' sad lament will be gradually alleviated. To that end, this book standardizes a core of basic information about test items.

Further, the guides and editorial rules for constructing particular test-item formats set down in this book could eventually—after review, deliberation, discussion, and appropriate modification—become industry standards for constructing test items. In this book they can serve as a first-draft model insofar as they may generate public discussion and deliberation among professionals in the field. Just as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985) and the *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee, 1981) serve professionals as guides by articulating important, shared information that promotes quality work in the field, the standards offered in this book may serve a similar purpose.

Another reason the information described here is important is that this book documents a significant amount of information about test items, both recent advances and traditionally accepted knowledge. In doing so this book will, in some measure, alleviate the paucity of scholarly materials about test items mentioned above. Additionally, disseminating the information will diminish myths and other inaccuracies about constructing good test items.

Yet a further reason for the importance of this book is the timeliness of the information. With the recent surge of interest in assessment, an enormous number of tests are currently used. The *Mental Measurements*