48

Writing Items and Tasks

Mary Schedl
Educational Testing Service, USA

Jeanne Malloy
Educational Testing Service, USA

Introduction

For high stakes testing a construct for testing is defined, and detailed test and item specifications exist before items (selected response questions) and tasks (constructed response questions) are written. Content specifications function as the blueprint for developing items and tasks that measure the abilities defined by the construct. But, once the blueprint exists, there are still many decisions left for item writers to make in producing the test items, such as creating items and creating or selecting stimuli for testing the abilities and determining which points to test. Item writers decide what questions to ask and whether questions are of appropriate difficulty and fair for the test population. While in textbooks and research there is a fair amount of agreement about appropriate guidelines for item writing, these sources provide very little explanation regarding the significance of individual guidelines, and they seldom include supporting examples and data. In this chapter we explicate important principles for both selected response items and constructed response tasks; and we provide representative examples of weak and strong items to illustrate why the principles matter. We use reading comprehension items to illustrate selected response principles, and writing and speaking tasks to illustrate constructed response principles; however, similar issues exist for other types of items and tasks.

Previous Views or Conceptualization

Before Haladyna and Downing (1989a) published a taxonomy of 43 multiple choice item-writing rules, very little attention was given to the design and construction of test items and tasks. Even today, very little empirical consideration has been given to the subject. Haladyna and Downing's taxonomy is a compilation

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.
© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.
DOI: 10.1002/9781118411360.wbcla025

of suggestions from textbook authors and other sources. Characteristics of the stimuli in item development are not considered. In another study, these two researchers analyzed the results of 96 theoretical and empirical studies to see what support they provided for each rule (1989b). Two rules received the greatest attention—a rule about the number of options an item should have and a rule about the need to balance (vary) the position of the correct answers in test questions. For nearly 50% of the rules no research was found. More recently Haladyna, Downing, and Rodriguez (2002) published a taxonomy of 31 multiple choice item-writing guidelines for classroom teachers. The authors considered the validity of each guideline both on the basis of collective opinions of textbook authors and on the basis of empirical research, but even here they state that "item writing is still largely a creative act" (p. 328). In 2005 Frey, Petersen, Edwards, Pedrotti, and Peyton compiled a separate list of 40 item-writing rules for classroom assessment, using an approach similar to that of Haladyna and Downing. There was substantial agreement in their results.

Also modeling their study on the taxonomy of Haladyna et al. (2002), Hogan and Murphy (2007) compiled advice on crafting and scoring constructed response tasks, identifying 12 recommendations for preparing constructed response tasks and 13 for scoring. Rationales underlying some of these recommendations are discussed, among others, in Schmeiser and Welch (2006) and McClellan (2010). In his book *Constructing Test Items* (1989), Steven Osterlind considers important issues in item writing, including the relationship of items to test validity and reliability. He also addresses the difficulty of establishing evidence for good items in absolute terms, and the fact that constructing test items demands complex technical skills and sophisticated levels of thinking. In addition, he attempts to synthesize the technical skills needed to construct test items. The Association of Language Testers in Europe (ALTE) published an extensive set of course materials for item writers (1995, updated 2005; we used the updated version) that presupposes the need for item writers to have a good background in models of language ability (Module 1), the test production process (Module 2), and item-writing issues and item types (Module 3). Nevertheless, while pointing out that "it is essential for an item writer to be trained in the techniques of item writing" (p.106), the authors list, but do not elaborate on, guidelines similar to those found in other taxonomies.

Given the paucity of empirical evidence related to item-writing rules, many testing programs produce their own version of tips and guidelines for item writers. In our experience taxonomies of item-writing rules alone are insufficient to guide novice item writers. They need detailed explanations and examples of both good and bad items. In this chapter we provide recommendations for writing good items and tasks, and we do so on the basis of many observations of items and their statistics over many years. To our knowledge, there are no extensive discussions in the literature that use pretest data to explain how writing items and tasks is related to principles of validity, fairness, reliability (the consistency of test scores across different forms of the same test), item or task difficulty, and discrimination (the power of an individual item or task to separate high ability test takers from low ability test takers). Explaining these relationships is our goal. We also consider characteristics of the stimuli as part of the item and task design.

Current Views or Conceptualization: Writing Test Items and Tasks

Once a valid construct has been defined and a test framework has been created, test specifications are developed. The specifications indicate the item types, the number of items of each type, and the knowledge, skills, and abilities to be tested. Item writers must be knowledgeable about the test construct and the framework in order to write appropriate test items on the basis of the given specifications and to ensure that the materials, too, are appropriate for the specified test population and purpose—which are also defined in the construct and framework. Test-taker performance on items offers evidence from which we infer the degree to which test takers have or lack the knowledge, skills, and abilities of interest, so the validity of the measurement depends on the degree to which the test items assess these appropriately. Poor item writing causes construct-irrelevant variables to influence measurement. Something that is irrelevant to the construct is not part of the knowledge, skills, and abilities that a test is supposed to be measuring. Item and task validity, discrimination, and difficulty can be negatively impacted by construct-irrelevant variables.

Fairness is a fundamental assessment principle that is directly related to validity. Xi (2010) argues that fairness is an aspect of validity and conceptualizes it as comparable validity for all relevant groups. An item that is unfair allows some test takers or groups of test takers to perform better or worse than other test takers of the same ability. It is the item writer's responsibility to ensure that test materials are equally accessible to test takers from different backgrounds, because failure to do so may lead to construct-irrelevant variance. The reliability of measurement across different forms of a test is directly related to the item writers' ability to create comparable and valid test questions of appropriate difficulty and discrimination.

We divide this section into three parts. In the first part we consider the craft of item and task writing as it relates to validity and fairness in the design of test questions and tasks. In the second part (selected response items) and in the third part (constructed response tasks) we discuss more subtle and craft-oriented features of item construction, which can affect the difficulty and the discriminating power of items and tasks. In particular, we examine how various components of items and tasks can influence difficulty and discrimination.

The Craft of Item and Task Writing: Fairness and Validity

The language in which selected response items and constructed response tasks are presented must be clear, precise, and unencumbered by superfluous or difficult language. If the item text is more difficult than it needs to be, then it will measure a test taker's ability to understand the item text in addition to the test taker's ability to comprehend the stimulus. This is neither as fair nor as valid a measure as it could be. The following excerpt is from a passage about nestling birds, and the question that follows it is an example of poor item text. Here and in all the examples, the asterisk marks the correct answer.

Passage excerpt
Many signals that animals make seem to impose on the signalers costs that are overly damaging. A classic example is noisy begging by nestling songbirds when a parent returns to the nest with food. These loud cheeps and peeps might give the location of the nest away to a listening hawk or raccoon, resulting in the death of the defenseless nestlings. In fact, when tapes of begging tree swallows were played at an artificial swallow nest containing an egg, the egg in that "noisy" nest was taken or destroyed by predators before the egg in a nearby quiet nest in 29 of 37 trials.

Item
According to the paragraph, the experiment with tapes of begging tree swallows established which of the following?

*(1) By making excessive noise in order to obtain the attention of a parent returning to the nest with provisions, nestling birds may put themselves at the mercy of predators.
(2) Predators are drawn to nests inhabited by nestlings more frequently than they attack nests in which only eggs are available.
(3) Tapes containing the sounds of nestlings begging for food may entice more predators than the noise made by real nestlings.
(4) Predators have no means at their disposal other than the begging calls of nestlings to help them locate nests.

In this example each of the options contains difficult words and phrases that are unnecessary for testing the examinee's understanding of the information about the experiment. Option 1 (the correct answer) is complex grammatically and includes lower frequency vocabulary than is used in the passage excerpt. Examinees may understand the text but not the words "excessive" or "provisions," which are not used in the passage itself. The phrase "at the mercy of predators" is essential to expressing the result of the experiment but may not be known by many non-native readers.

Option 2 contains the idiomatic phrase "drawn to" and the participle "inhabited," which is more difficult than necessary; option 3 contains the difficult word "entice," which is not part of the passage excerpt; and option 4 contains the fairly uncommon idiomatic phrase "at their disposal." A version with simpler options would better allow test takers to demonstrate whether they understand what the experiment shows.

Another example of testing the item text rather than the author's intended point is provided below. The use of the negative in both stem and options is confusing.

Item
According to the paragraph above, which of the following is NOT true about the noisy begging by nestling songbirds?

(1) It may not go unnoticed by predators.
(2) It may occur when a parent returns with food.

(3) It may result in the death of nestlings.
*(4) It may not attract predators to the nest.

With negatives both in the stem and in the options, it is difficult to keep in mind what is true and what is false. In this case Options 1, 2, and 3 are true, but Option 1 includes a negative. Option 4 is not true but does include a negative. Revising options 1 and 4 to eliminate the negatives would make this a more reasonable item, as the following example illustrates.

Revised

(1) It may be noticed by predators.
*(4) It may keep predators away from the nest.

Similarly, the language of constructed response tasks must be as clear and unencumbered by superfluous or difficult language as possible. A test taker who does not completely understand a task is less likely to produce a work sample that accurately represents his or her ability. Precise action verbs and specific descriptive phrases delineating performance expectations are preferable because they better convey the nature or purpose of a task. For example, it is clearer and more precise to ask test takers to "summarize" another piece of writing than it is to ask them to "discuss" it.

Tasks or their directions should include information indicating the type and amount of detail or elaboration expected, hence the inclusion of comments such as "be sure to support your ideas with specific reasons and examples" and the mention of a typical word range for high quality responses. The absence of specific guidance concerning performance expectations can lead to construct-irrelevant variance. For example, in a constructed response assessment of writing proficiency, able but stylistically economical test takers may leave out examples or other supporting information, unless they are informed that this level of detail is expected.

Constructed-response tasks such as integrated skills tasks have multiple components, and directions are necessarily more complex when such tasks are administered. In TOEFL integrated writing tasks, for example, test takers read a brief passage and then listen to a lecture on the same subject before writing a response on the basis of what they just read and heard. In a staged item such as the one just described, directions for each component are supplied as appropriate: "Now listen to a lecture on the subject you just read about." If preparation time is an item component, the amount of preparation time allowed before responding should be indicated. For example, in a constructed response task measuring speaking proficiency, test takers may be given 30 seconds to make an outline or to mentally prepare a response after learning what the specific speaking task is, and then 60 seconds to deliver their oral response.

When multiple constructed response tasks are included on an assessment and test takers must make decisions about allocating their time, they should be told the point value for each task.

Items and tasks should avoid taking for granted content knowledge that might not be present to the same degree in all test takers and hence might unfairly

disadvantage or advantage certain groups in the test population. Cultural differences in outside knowledge could lead to a significant difference in performance of test takers on an English language test. Consider the following excerpt from a passage on European art:

> *Passage excerpt*
> Academic practice and theory were based on the study of officially approved models . . . and the belief that art was governed by rules akin to the laws of nature or grammatical structures. These precepts were challenged by the Romantic notion of individual genius, which cast the true artist as a rebel who necessarily rejected rules and conventions. In reality, the divide was sometimes less clear-cut than this: for example, J. M. W. Turner, the British artist who revolutionized landscape painting and was acknowledged as an important influence by many later avant-garde artists, remained a passionately loyal member of London's Royal Academy.

The assumption here is that the reader is familiar with Romanticism as a movement, has a good idea of what London's Royal Academy was, understands the implications of being a member of this society, and knows what avant-garde artists stood for. It is unlikely that non-Europeans would be as familiar as Europeans with these matters.

Care must also be taken that the stimuli for items and tasks are not too time-consuming. In a timed reading comprehension test, the amount of time needed to process passage information must be considered in the item construction process. If more time is required than is available to a test taker, the test taker may try to guess the correct answer or may omit an item, both of which are likely to affect item discrimination and validity.

Similarly, some constructed response tasks may burden the memory. If a stimulus is long, or if there is a delay between the presentation of information in the stimulus and the response (as sometimes occurs in staged tasks), individual differences in the ability to recall rather than in language ability may influence performance. Shortening the stimuli, shortening the time between presentation of the stimuli and response, allowing test takers to take notes during the presentation of stimuli, and giving test takers access to parts of the stimuli while they are responding are ways to reduce the need to recall. For example, in some writing tasks based on reading stimuli, test takers can view the reading stimuli as they write.

Difficulty and Reliability

In this part we consider the individual components of selected response items and how the construction of these components influences item difficulty and discrimination. As noted earlier, discrimination is the power to differentiate high ability test takers from low ability test takers. The higher the level of discrimination, the better. The range of the classical item analysis discrimination index is –1.0 to 1.0. Statistics are reviewed for pretest items, and items with low discrimination may

be revised and then re-pretested before being delivered operationally. TOEFL items that discriminate below 0.30 are routinely reviewed for item flaws.

Items should test knowledge and skills that are appropriate for the test purpose and population. For a typical test, items range in difficulty from easy to hard for the intended group, and the greatest concentration of items is in the range in which 30% to 70% of the test takers get the item correct. Items significantly easier or more difficult discriminate among members of a relatively small proportion of the test population.

Different parts of a text may vary in lexical, syntactic, and conceptual difficulty. It is important that items that test different parts of such a text correspond in difficulty to the parts they are testing. Ideally, the specific part needed to answer a question should determine the difficulty of that question. Difficult items should not be written about easy parts of a text, because the inferences we draw about test takers' abilities are based on their responses to items. Low ability test takers should answer questions about the easy parts of a text correctly, but they should answer incorrectly questions about the difficult parts of a text. In the following discussion we consider each item component separately and provide examples of items we consider flawed. Some examples represent item development problems that new test developers commonly create and some are from actual TOEFL pretests. For the latter, we look at the item analysis after initial pretesting and compare it with the new item analysis after re-pretesting.

Item Stem

The stem can be written as a question or as an incomplete statement that is to be completed by selected response options, but the stem should not be undirected. For example, a stem that simply states "the author believes that . . ." is undirected because it forces the test taker to read the options in order to understand what is being asked. Test takers who understand a point being tested should be able to formulate an answer to the question without first reading the options.

Well-crafted stems are free of ambiguity and direct the test taker's attention to the part of the stimulus that contains the information needed for answering the item. The following example is from a TOEFL reading comprehension pretest.

Passage excerpt

The undisputed pre-Columbian presence on the Pacific islands of Oceania of the sweet potato, which is a New World domesticate, has sometimes been used to support Heyerdahl's "American Indians in the Pacific" theories. However, this is one plant out of a long list of Southeast Asian domesticates. As Patrick Kirch, an American anthropologist, points out, rather than being brought by rafting South Americans to Oceania, sweet potatoes might have just as easily been brought back by returning Polynesian navigators who could have reached the west coast of South America.

Question

Why does the author discuss the presence of the sweet potato on the Pacific islands?

(1) To present evidence in favor of Heyerdahl's idea about American Indians reaching Oceania
(2) To emphasize the familiarity of Pacific islanders with crops from many different regions of the world
*(3) To indicate that a supposed proof of Heyerdahl's theory has an alternative explanation
(4) To demonstrate that some of the same crops were cultivated in both South America and Oceania

This item was flagged after pretesting because, although 46% of the TOEFL test population chose Option 3, the intended answer, 14.2% of the most able test takers chose a different option. The stem was found to misdirect readers from the intended key and thus was revised to be more directed: "Why does the author mention the views of Patrick Kirch?" When the item was pretested again, 57% chose the intended option and discrimination improved significantly (from 0.44 to 0.55), only 5% of the most able test takers choosing an incorrect answer. The stems of items should also pose questions that are independent of each other.

Since every question on a test contributes to the inferences drawn about a test taker's ability, it is important that the items be independent in the sense that each test question tests a separate point. Lack of independence reduces overall test reliability. The following examples are based on a passage about nesting birds.

Passage excerpt
Further evidence for the costs of begging comes from a study of differences in the begging calls of warbler species that nest on the ground versus those that nest in the relative safety of trees. The young of ground-nesting warblers produce begging cheeps of higher frequencies than do their tree-nesting relatives. These higher-frequency sounds do not travel as far, and so may better conceal the individuals producing them, who are especially vulnerable to predators in their ground nests.

Item
This paragraph indicates that the begging calls of tree-nesting warblers

(1) put them at greater risk than ground-nesting warblers experience
*(2) can be heard from a greater distance than those of ground-nesting warblers
(3) are more likely to conceal the signaler than those of ground-nesting warblers
(4) have higher frequencies than those of ground-nesting warblers

If another item were to ask the following question, then test-wise examinees would know that it must be true that the begging calls of tree-nesting warblers can be heard from a greater distance:

Which of the following can be inferred from the fact that the begging calls of ground-nesting warblers do not travel as far as those of tree-nesting warblers?

Because the second question provides the information requested in the previous question, examinees may be able to answer the first question without understanding this point in the passage itself.

Item Key

The key, like the item stem, should be as precise and unambiguous as possible. The following is an example of an imprecise key taken from a TOEFL pretest.

Passage excerpt
More recent evidence suggests, however, that autonomic activity may not be as broad and diffuse as Cannon contended. Some studies of autonomic activity show clear differences in the autonomic patterns that accompany such emotions as anger and fear. And people across cultures report bodily sensations that differ depending on the emotion: they generally report a quickened heartbeat and tense muscles both when angry and when fearful, but they feel hot or flushed strictly when angry and cold and clammy strictly when afraid. However, even with these refinements, the fact remains that . . .

Item
The word "refinements" in the passage is closest in meaning to

*(1) adjustments
(2) variations
(3) findings
(4) applications

In the original version above, 46% of TOEFL test takers answered correctly, with a discrimination value of only 0.28, which is below the 0.30 threshold for item review for TOEFL items. The key was replaced with "small improvements"; this was designed to make it more precise, after which 45% of test takers answered correctly, with an improved discrimination of 0.41.

Distracters

The purpose of distracters, or incorrect answer choices, is to make it possible to discriminate test takers in terms of the knowledge, skills, and abilities being tested. Able test takers select the correct answer (the key) and less able test takers select distracters.

Because distracters must be wrong but plausible, it is usually more difficult to create distracters than it is to create the stem or the key. Distracters can be based on a statement or idea that is taken from the passage and then modified so as to become incorrect, or they can be plausible answers to the question that are not supported by information in the stimulus. In general, the finer the distinctions that must be made between the key and the distracters, the more difficult the item. The abilities of the test population and the purpose of discriminating among the test takers must be kept in mind in determining how fine the distinctions need to

be. There are two major considerations in designing distracters and many ways for item authors to go wrong, as illustrated in the following examples.

First, distracters must be attractive to test takers who do not sufficiently understand the stimulus material or the point being tested. Therefore they should be at least superficially related to the stimulus or topic. If the key uses vocabulary from the stimulus, so should the distracters. For questions covering only a small part of a large text, distracters are generally drawn from the same area of the text as the key, because this is the area of the text where test takers expect to find the answer. The item testing the following text includes poor distracters that do not utilize vocabulary or ideas from the stimulus.

Passage excerpt

Off and on throughout the Cretaceous period, large shallow seas covered extensive areas of the continents. Data from diverse sources, including geochemical evidence preserved in seafloor sediments, indicate that the Late Cretaceous climate was milder than today's. The days were not too hot, nor the nights too cold. The summers were not too warm, nor the winters too frigid.

Weak version

According to the paragraph above, which of the following is true of the Late Cretaceous climate?

(1) The climate was very similar to today's.
(2) The climate supported a large number of species.
(3) The climate was extremely dry.
*(4) The climate did not change dramatically from season to season.

In this weak version, Options 2 and 3 are unlikely to attract test takers who are guessing because they do not include vocabulary from the stimulus, which does not mention "species" or "dryness."

A reasonable item can be created by revising these two options:

Revised version

(2) Summers were very warm and winters were very cold.
(3) Shallow seas on the continents caused frequent temperature changes.

Similarly, distracters need to be written so that they cannot be eliminated on the basis of common sense or common knowledge. If a question were to ask for a reason why dinosaurs became extinct, a distracter stating that humans hunted them to extinction would be easy to eliminate because virtually everyone knows that humans and dinosaurs did not coexist.

Test takers are sensitive to positive and negative connotations in stimuli, even when they do not understand specific details, so care should be taken that distracters do not violate test-taker expectations in this regard. In the following example, the immediate context for the word tested is more negative than positive, so the distracters should be either negative or neutral.

Passage excerpt
To the extent that the coverage of the global climate from these records can provide a measure of its true variability, it should at least indicate how all the natural causes of climate change have combined. These include the chaotic fluctuations of the atmosphere, the slower but equally erratic behavior of the oceans, changes in the land surfaces, and the extent of ice and snow.

Item
The word "erratic" in the passage is closest in meaning to

(1) dramatic
(2) important
*(3) unpredictable
(4) beneficial

Option 4 (in context, "beneficial behavior of the oceans") does not fit the comparison to the "chaotic fluctuations of the atmosphere," making this a distracter likely to be eliminated by test takers who are guessing.

Distracters are also unattractive when they include absolute terms, such as "never" and "always." It is easy to eliminate a distracter that is absolute, because very few things are either always or never true.

The second major principle concerning the development of distracters is that they need to be clearly false. In the following example, one distracter proved to be too close to the key, resulting in an item that discriminated poorly.

Passage excerpt
Over long periods of time, substances whose physical and chemical properties change with the ambient climate at the time can be deposited in a systematic way to provide a continuous record of changes in those properties over time, sometimes for hundreds of thousands of years. Generally, the layering occurs on an annual basis, hence the observed changes in the records can be dated. Information on temperature, rainfall, and other aspects of the climate that can be inferred from the systematic changes in properties is usually referred to as proxy data.

Item
According to this paragraph, scientists are able to reconstruct proxy temperature records by

(1) studying regional differences in temperature variations
*(2) studying and dating changes in the properties of substances
(3) observing annual changes in the properties of substances as they are deposited
(4) inferring past climate shifts from observations of current climatic changes

When the item was first pretested, 30% of the top-ability group selected Option 3. Only 25% of the TOEFL population selected Option 2, the intended key. The

item discrimination was only 0.23. When Option 3 was revised to "observing changes in present day climate conditions," 47% of the TOEFL population selected Option 2, so the item became easier and its discrimination value increased to 0.48.

The Craft of Writing Constructed Response Tasks: Difficulty and Discrimination

In constructed response tasks, discrimination in test-taker performance levels is achieved by assigning scores along a performance continuum with well-defined score points. A primary reason for trying out constructed response tasks before administering them operationally is to detect tasks that are easier or more difficult than desired, so they may be revised or eliminated. Task difficulty is typically determined by analyzing score distributions and by computing the mean or average score. Normally scores should be distributed across the full range of score points, and score averages for supposedly comparable tasks should be similar.

For a high stakes decision scores should be highly reliable, meaning that a test taker would receive the same score on a different but comparable task of the same type, but one scored by different raters. For lower stakes uses such as providing diagnostic feedback, a lower level of reliability may be adequate. Typically, reliability in constructed response tasks is measured in terms of consistency across applications of the measurement procedure. One method for achieving consistency is to have clear, detailed specifications for prompts and stimuli. For example, in prompts based on stimuli, stimuli characteristics such as length and complexity should be defined.

The method or methods for determining reliability depends on the testing situation. If multiple raters are used to score responses independently, for example, the consistency of test-taker scores across raters (inter-rater agreement) can be used to measure reliability. Reliable scores require reliable scoring procedures.

Developing a Scoring Rubric

A scoring rubric is essential for reliable scoring. A scoring rubric delineates the criteria by which responses to constructed response tasks are discriminated.

Typically, the scoring rubric for a given task type is developed as the item specifications are being determined, and it is refined as the task types are being prototyped and piloted. The criteria for scoring must reflect the purpose for which the item has been designed and must focus on the response characteristics necessary for evaluation. As mentioned above, these characteristics are defined along a continuum.

Scoring rubrics typically are either analytic or holistic. In analytic rubrics, each desired feature of a response is identified and awarded a specific point value. In holistic rubrics, score points are defined on the basis of the overall impression of a response. In both cases, a range of possible score points is specified and verbal descriptors are created for each score point. Generally, as many score points are used as can be consistently and meaningfully delineated and evaluated.

Identifying, Training, and Monitoring Raters

Raters must have the necessary educational qualifications and experience to rate responses. They must also be able to demonstrate mastery of the scoring training materials. In TOEFL and GRE, for example, this is achieved by requiring raters to pass a certification test upon completion of training. Raters retrain briefly before each scoring session and perform satisfactorily on a calibration exercise before being permitted to score operationally.

Rater training materials include benchmarks and range finders. Benchmarks are responses that have been selected as exemplars of responses at each score point on the rubric. Range finders are responses selected to guide raters in scoring responses that may be harder to match to the rubric. They may be examples of responses that, for example, are almost, but not quite, good enough to be awarded the higher of two adjacent score points.

Benchmarks and range finders are selected as soon as it is possible to obtain an adequate sampling of responses, and raters should be able to consult these materials as needed throughout operational scoring.

To ensure that raters are making appropriate distinctions at each score point on the rubric, rater performance is monitored during operational scoring using both statistical methods (inter-rater agreement rates, rater agreement rates with monitor responses, and the distribution of scores assigned) and qualitative measures (having scoring leaders selectively read rated responses to check for accuracy during scoring sessions).

Refining Constructed Response Tasks on the Basis of Review and Tryouts

It is difficult, perhaps impossible, to judge whether constructed response tasks are clear and appropriate for a given population without subjecting them to meaningful review and tryouts, preferably both, for tasks on high stakes assessments. Like many other aspects of test development, crafting high quality constructed response tasks is a recursive process of successive refinements rather than a linear process.

Various approaches can be used for reviewing tasks. For example, test developers can perform a task themselves and then use the rubric to score their responses. However, because test developers' abilities tend to be significantly different from those of the test takers, trying out tasks on a subgroup of the test population generally provides more meaningful results. Tryouts should be administered under the conditions to be used for operational administrations, and responses should be scored by experienced raters.

Tryouts can be helpful in determining whether (a) the test takers understand what they are supposed to do; (b) the tasks are appropriate for the test population; (c) a particular subgroup of test takers seems to have a nonconstruct-related advantage over other subgroups; (d) the tasks elicit responses of the length and complexity desired; (e) responses are distributed across the full range of score points, or they cluster at selected score points; and (f) the responses can be easily and reliably scored using the existing rubric (for example, responses scored independently by more than one person are awarded the same or adjacent scores).

It should be also be noted that, for practical reasons, it is not always possible to obtain enough responses through the tryout process to reliably determine score distributions and mean scores.

Decisions concerning which items of a given type to use operationally are based on item analysis and rater input. For test forms to be comparable, tasks of a given type should have similar mean scores and similar score distributions across the rubric score points from form to form.

Analyzing rater data is helpful in selecting pretested items for operational use. In cases where multiple raters score the same response, high inter-rater agreement is a possible indicator of quality. However, inter-rater agreement must be examined in light of score-point distributions, as it is necessarily high when only a limited number of the available score points are being awarded to responses.

The TOEFL integrated writing item discussed below was crafted with care and received multiple reviews by test developers before it was tried out, yet some problems were not apparent until test-taker responses were examined.

Example: Stegosaurus Plates

In this item test takers read a short passage explaining three theories about why stegosaurus dinosaurs had bony plates on their backs. The reading is illustrated with a drawing of a stegosaurus, so that test takers are sure to understand the type of animal being discussed, and their ability to visualize is thus minimized as a possible source of construct-irrelevant variance. After completing the reading, test takers listen to a part of a lecture in a biology class in which the professor rebuts each of the three theories presented in the reading. Test takers hear the lecture only once but are permitted to take notes while listening to it. A few seconds after the lecture concludes, test takers are presented with the prompt, which asks them to explain in writing how the lecture they just heard challenges information in the reading. They are given 20 minutes to write and told that good responses are typically between 150 and 225 words long. They can view the reading passage (and the illustration) as they write.

One of the theories presented in the reading is that the plates protected the dinosaurs against attacks by predators. In the lecture, a professor rejects this explanation by arguing that the plates were ineffective at providing protection. In a tryout version of this part of the lecture, the professor says that the plates were thin and "could have been bitten through easily." Tryouts revealed that some test takers who write well misinterpreted the word "bitten" as "beaten." It was hypothesized that the comprehension problem was due to the short vowel sound in the word "bitten." Accordingly, the wording of the lecture was changed to "would be able to bite through them [the plates] easily," in which the vowel sound is more distinctive.

Another of the theories presented in the reading is that the plates helped lower body temperature when the animal became overheated. The reading points out that the plates contained blood vessels and that blood vessels can carry heat to the body's surface, where it then radiates into the atmosphere. In the lecture, the professor rebuts this argument by pointing out that the blood vessels were not

located where they would have been useful for this purpose, namely near the surface. The rebuttal of this point was presented in the lecture as follows, when the item was first tried out:

> Second, the temperature regulation theory. A closer look at the actual pattern of the vessel channels in the plates undermines this theory. If the cooling theory were correct, the vessels would be leading the blood along the surface of the plates where the blood would cool, and then carrying the cooled blood back into the body. But the actual pattern of the vessels seems different, suggesting that their real function was to direct blood toward the living tissues in the plates, supplying them with nutrients and helping them grow. So, the blood flow pattern inside the plates was suitable for supplying nutrients to living tissues rather than for temperature regulation.

In the tryout it was discovered that some high ability test takers had difficulty understanding why the blood vessels were unsuitable for radiating excess heat from the body, possibly because the lecture was not explicit about the location of the "living tissues of the plate." The contrast between the plate surface and the inner tissues of the plate was made explicit in the revised version. The revised version was also simpler: the information that blood vessels carry cooled blood back into the body was removed as nonessential. Here is the revised text of this part of the lecture:

> Second, the temperature regulation theory. This theory is inconsistent with how the blood vessels were arranged in the plates. If the cooling theory were correct, the vessels would lead the blood along the surface of the plates where the blood would cool. But the vessels were not arranged in this way. Instead, their arrangement suggests a very different function: the blood was mostly directed toward the living tissues inside the plates, supplying them with nutrients and helping them to grow. So the main function of the blood flow in the plates was to supply nutrients to living tissues rather than temperature regulation.

In subsequent administrations there was no significant pattern of test takers with high writing ability having trouble understanding the information conveyed in the revised wording. The revised wording appears to have improved the validity, fairness, and discriminating power of the task.

Current Research and Future Directions

Proposals for research that compares test tasks and the abilities they require to real-world tasks and abilities are called for in the TOEFL Committee of Examiners 2013 Research program. A study evaluating the relationship between authentic stimuli and test stimuli was conducted for IELTS (International English Language Testing System) in 2010 (Green, Ünaldi, & Weir, 2010) and one is currently underway for iBT TOEFL (the Internet-Based Test of English as a Foreign Language) (Sheehan, in press).

A promising area of research is work on text analytics tools, which automate basic linguistic analyses of stimuli or other materials. These tools may, for example, analyze word frequency, syntactic complexity, and lexical and semantic cohesion in a given stretch of text in ways that are relevant to predicting difficulty. In addition, tools are being created to model item difficulty and to support item authoring by test developers. ETS (the Educational Testing Service) is also currently researching and developing automated engines for scoring both spoken and textual responses.

Technology plays an increasing role not only in testing but also in learning. New assessments will likely re-examine language constructs in light of computer learning and investigate whether new abilities are required and new items and tasks are needed to assess them. The TOEFL program is currently updating the language frameworks that guided the iBT TOEFL in light of possible changes to these constructs over time.

Challenges

As the examples in this chapter indicate, there are many possible challenges to item and task validity, and many design and language variables that can influence item and task difficulty, discrimination, and reliability. For this reason, pretesting of items and tryouts of tasks are highly desirable.

Perhaps the greatest challenges for programs using constructed response tasks are the time and expense involved in using human raters and ensuring that item difficulty is consistent across forms. As the previous discussion makes clear, high quality human scoring requires a considerable investment of time and resources. Fortunately progress is being made in creating and improving engines for automated scoring, and some engines for measuring writing performance produce results comparable to those produced by human raters. However, less progress has been made in developing effective strategies for detecting and minimizing variations in item difficulty in constructed response tasks across forms. Although some techniques exist (e.g., establishing mean item scores, determining comparable score distribution), these methods depend on adequate sampling and high quality scoring. For practical and test security reasons (e.g., it is easier to memorize constructed response tasks than it is to memorize multiple choice items), it may not be possible to obtain large enough samples through tryouts to detect and eliminate some item flaws and variations in difficulty across forms.

SEE ALSO: Chapter 13, Assessing Integrated Skills; Chapter 17, International Assessments; Chapter 33, Norm-Referenced Approach to Language Assessment; Chapter 34, Criterion-Referenced Approach to Language Assessment; Chapter 53, Field Testing of Test Items and Tasks; Chapter 57, Standard Setting in Language Testing; Chapter 80, Raters and Ratings; Chapter 94, Ongoing Challenges in Language Assessment

References

Association of Language Testers in Europe (ALTE). (1995). ALTE materials for the guidance of test item writers. Retrieved October 12, 2011 from www.alte.org/downloads/index.php?docid=89

Association of Language Testers in Europe. (2005). ALTE materials for the guidance of test item writers. Retrieved October 12, 2011 from www.alte.org/downloads/index.php?docid=89

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education: An International Journal of Research and Studies*, *21*(4), 375–64.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, *27*(2), 191–211.

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., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–34.

Hogan, T. P., and Murphy, G. (2007) Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, *20*(4), 427–41.

McClellan, C. A. (2010, February). *Constructed-response scoring: Doing it right. R&D connections, 13*. Princeton, NJ: Educational Testing Service.

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

Schmeiser, C. B., & Welch, C. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–53). Westport, CT: American Council on Education/Praeger.

Sheehan, K. (in press). Are TOEFL iBT reading passages characteristic of the types of reading materials typically encountered by students in university settings? A study funded by the TOEFL Committee of Examiners Research Program, 2011. Manuscript in preparation.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–70.

Suggested Readings

Alderson, C. J. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Brown, J. D., & Hudson, T. (1998). Alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653–75.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge, Taylor & Francis Group.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge, Taylor & Francis Group.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 387–432). Westport, CT: American Council on Education/ Praeger.

Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–27). Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.