Test Directions as a Critical Component of Test Design: 
Best Practices and the Impact of Examinee Characteristics

Joni M. Lakin
Auburn University

To appear in Educational Assessment

Author Note

Joni M. Lakin, Department of Educational Foundations, Leadership, and Technology, Auburn University.

This work was supported by a fellowship from the American Psychological Foundation. The author gratefully acknowledges the helpful comments of David Lohman, Kurt Geisinger, Brent Bridgeman, Randy Bennett, and Dan Eignor on earlier drafts of this article.

Correspondence concerning this article should be addressed to Joni Lakin, Department of Educational Foundations, Leadership, and Technology, Auburn University, Auburn, AL 36849. Email: joni.lakin@auburn.edu
Abstract

The purpose of test directions is to familiarize examinees with a test so that they respond to items in the manner intended. However, changes in educational measurement as well as the U.S. student population present new challenges to test directions and increase the impact that differential familiarity could have on the validity of test score interpretations. This article reviews the literature on best practices for the development of test directions as well as documenting differences in test familiarity for culturally and linguistically diverse students that could be addressed with test directions and practice. The literature indicates that choice of practice items and feedback are critical in the design of test directions and that more extensive practice opportunities may be required to reduce group differences in test familiarity. As increasingly complex and rich item formats are introduced in next-generation assessments, test directions become a critical part of test design and validity.
Test Directions as a Critical Component of Test Design: Best Practices and the Impact of Examinee Characteristics

Introducing new or unfamiliar item formats to examinees poses particular challenges to test developers because of the need for students to quickly and accurately understand what the test items require (Haladyna & Downing, 2004; Lazer, 2009). Item formats are defined not only by their response modes (e.g., multiple choice, constructed response), but also by the types of demands they are expected to place on examinee’s knowledge and cognition (e.g., basic arithmetic, scientific reasoning; Martinez, 1999). When introducing new item formats, the critical challenge is how best to introduce the task so that all students are able to respond to the format as intended by the test developers.

With new innovations in achievement testing, especially through computer-based testing and performance assessment, novel item formats are increasingly being introduced to achievement tests to increase construct representation and measure more complex and authentic thinking skills (e.g., National Center for Education Statistics [NCES], 2012; see also Bennett, Persky, Weiss, & Jenkins, 2010; Fuchs et al., 2000; Haladyna & Downing, 2004; Lazer, 2009; Parshall, Davey, & Pashley, 2000). In these new formats, simulations, computer-based tools, and items with complex scoring procedures are used for both teaching and assessment (Bennett et al., 2010; NCES, 2012). Furthermore, the role of innovative formats (and thus initially unfamiliar formats) will only increase in the near future because computer-based assessments with more complex and authentic test items form a central component of the general assessment consortia formed in response to Race to the Top (PARCC and SMARTER Balanced Assessment Consortia; Center for K–12 Assessment & Performance Management, 2011; Lissitz & Jiao, 2012).

When item formats are unfamiliar, test directions and practice become a crucial, but often overlooked, component of test development and a threat to valid score interpretation (Haladyna & Downing, 2004). Test directions serve to explain unfamiliar tasks to test-takers and are often the only means that test developers have of diminishing potential unfamiliarity or practice effects. Further, if the test developers suspect there will be preexisting group differences in test familiarity, test directions are
their last opportunity to address this validity threat as well. If groups differ in their understanding of the test task after the directions have been completed, bias can be introduced to the test scores, resulting in score differences between the two groups that do not reflect real group differences in ability (Clemans, 1971; Cronbach, 1984). As James (1953) advocated, test directions should “give all candidates the minimum amount of coaching which will enable them to do themselves full justice on the test. All candidates are then given a flying start and not just some of them” (p. 159). The goal of this article is to outline some key considerations in developing high quality test directions. This is particularly important in the context of a culturally and linguistically diverse society where test familiarity is more common among some cultural and linguistic groups and creates a barrier to educational opportunities for individuals from other groups (Kopriva, 2008). These disparities are discussed below.

**Importance of Developing Good Directions and Practice Activities**

Test directions are an instructional tool whose purpose is to teach examinees the sometimes complex task of how to solve a series of test items. Depending on the students and test purpose, these directions may include general testing taking tips (e.g., “Don’t spend too long on any one item!”; Millman, Bishop, & Ebel, 1965). Some directions rely heavily on examples accompanied by oral or written explanations of the task, while others consist only of brief oral or written introductory statements. The effectiveness of a set of directions depends not only on the characteristics of those directions, but also on the characteristics of the examinees, including their ability to understand the written or spoken directions and familiarity with the test tasks. Good test directions provide enough information to familiarize examinees with the basic rules and process of answering test items so that test items are answered in the manner intended by the test creator (AERA, NCME, & APA, 1999; Clemans, 1971). With increasingly complex computer-based or interactive tasks, efficiency in answering the questions and understanding the task (with minimal guidance from a proctor) is also critical. The *Standards* (AERA et al., 1999) further emphasize that examinees must have equal understanding of the goals of the test, basic solution strategies, and scoring processes for test fairness to be assured. The *Standards* suggest the use of practice materials and example items as part of the test directions to accomplish this goal.
On some tests, such as classroom tests, traditional achievement tests, and even aptitude tests for college applicants, it is assumed that understanding the task is trivial and so the directions are perfunctory (Scarr, 1981). In these cases, the directions can be brief because the examinees have been exposed to many similar tests (e.g., vocabulary or mathematical computation tests) throughout their education. Some researchers even complain that examinees entirely ignore the test directions and skip straight to the test items as soon as they are able (James, 1953; LeFevre & Dixon, 1986). In these cases, low-quality directions are expected to have a negligible impact on test scores.

On the other hand, it has long been acknowledged that practice can significantly affect performance on novel tasks and that differences in familiarity can lead to construct-irrelevant differences in scores (Haladyna & Downing, 2004; Jensen, 1980; Thorndike, 1922). Several researchers exploring the effects of training on item performance confirmed that more complex and unfamiliar formats showed the greatest gains in response to training (Bergman, 1980; Evans & Pike, 1973; Glutting & McDermott, 1989; Kulik et al., 1984; Jensen, 1980).

The size of practice effects likely has practical importance, but diminishes quickly after a few exposures to a particular test (Kulik et al., 1984; te Nijenhuis et al., 2001). Practice effects on traditional formats in parallel forms of a test have been estimated to be .23SD for a single practice test (meta-analysis; Kulik, Kulik, & Bangert, 1984). Jensen (1980) estimated this effect to be a .33 IQ SDs gain based on the literature up to 1980. A noteworthy point is that practice gains differ significantly by examinee type and test design.

Thus, on tests where the formats are less likely to be familiar, the demands of understanding the task may not be trivial and thus the directions must not be perfunctory. In the absence of a full practice test, directions with practice examples are test developers’ best chance to decrease practice gains on novel tests. With the introduction of complex item formats into achievement tests, such as laboratory simulations (NCES, 2012), the novelty of the items may be irrelevant and detrimental to measuring the desired construct in early applications of the format (Ebel, 1965; Fuchs et al., 2000).
These effects may have critical impacts on test use. For example, Fuchs et al. (2000) found that a brief orientation session on a new mathematics performance assessment (an unfamiliar assessment format for their sample of students) greatly improved the scores students received (by .54 to .75SD, depending on subtest). They concluded that unfamiliarity with the testing situation, which was particularly acute for performance assessments where both the format and grading scheme were unfamiliar, was an obstacle to the students’ ability to demonstrate their mathematical knowledge on the test. In their case, the unfamiliarity not only changed the construct measured, but also created spurious growth effects that could confound interpretations of the test scores longitudinally.

**Importance of Directions in Culturally and Linguistically Diverse Examinee Populations**

Over time and with additional test experience, students acquire important problem-solving skills and knowledge relevant to the format (not only general test-wiseness [Millman et al., 1965] but format-specific knowledge and strategies). Students with greater experience with the item format may use more sophisticated or practiced strategies in completing the items compared to students with less experience (Norris, Leighton, & Phillips, 2004; Oakland, 1972).

When all students are affected by item format or test unfamiliarity, the influence of unfamiliarity can be detected through traditional validity methods such as retesting and experimental designs (such as those used by Fuchs et al., 2000). Unfamiliarity in this case can also be responded to with a standard set of directions. However, differential familiarity may, in fact, be an even greater threat than general unfamiliarity to valid score interpretation (Fairbairn, 2007; Haladyna & Downing, 2004; Maspons & Llabre, 1985; Ortar, 1960; Scarr, 1994). When familiarity only affects a subset of students or even individual students, detecting its effects (termed individual score validity [ISV] by Kingsbury & Wise, 2012) becomes more difficult because many validation methodologies rely on accurately identifying affected examinee subgroups, such as race/ethnicity. For example, differential item functioning methods can only detect intermittent group differences in item functioning, not differences that affect the entire test or an unidentified subset of students. Because of this difficulty, test developers should actively anticipate instances where unfamiliarity could be a threat to valid and fair use of test scores for all students.
When new item formats are introduced to a testing application, if some students are able to gain additional practice or test exposure, these differences could potentially contribute bias to test scores (Haladyna & Downing, 2004). These differences in ability to profit from test practice are pervasive. Prior experience with tests and practice activities provided by parents or extracurricular programs can provide some students with greater knowledge of typical tests and solution strategies (Hessels & Hamers, 1993). When tests are high-stakes, these influences are especially likely and worrisome (Haladyna & Downing, 2004). But extracurricular practice is not the only concern. For students who are new to U.S. schools or do not come from families who are engaged with their school system, access to knowledge about tests and test preparation activities may be restricted and also lead to differential effects of item format novelty (Budoff, Gimon, & Corman, 1974).

The critical assumption that an item format is equally familiar to all students (a foundation of valid score interpretation; AERA et al., 1999; Haladyna & Downing, 2004) seems especially problematic among culturally and linguistically diverse students whose experiences with testing may vary considerably (Anastasi, 1981; Fairbairn & Fox, 2009; Peña & Quinn, 1997; Sattler, 2001). Both students and parents can vary in their understanding of the purposes of testing, motivation, knowledge of the tests used, and access to test preparation opportunities (Kopriva, 2008). These differences are exacerbated when tests have high stakes attached to them (Reese, Balzano, Gallimore, & Goldenberg, 1995; Solano-Flores, 2008; Walpole et al., 2005).

The U.S. school population has always been diverse, but the challenges for educational assessment are increasing with rising numbers of Latino students and English learners (EL; Federal Interagency Forum on Child and Family Statistics; 2011; Passel, Cohn, & Lopez, 2011). Latino and EL students have a number of social and cultural differences—in addition to the linguistic differences—that may contribute to the widely documented differences in their test performance compared to White and non-EL students. These differences are described in the sections that follow.

**Test-wiseness.** There are apparent cultural differences between mainstream middle-class White families and families of other racial/ethnic background and SES with respect to their testing-related
parenting behaviors and beliefs about testing purposes. Although test-taking skills classified as test-wiseness are usually not addressed by test directions, they could be and might contribute to students’ ability to profit from test directions that focus on the particular of a given format. Cultural differences seem to affect students’ test-wiseness (general knowledge and strategies for testing). Greenfield, Quiroz, and Raeff (2000) and Peña and Quinn (1997) found differences in assessment-related behaviors for Latino mother-child pairs compared to White mother-child pairs. Peña and Quinn argued that differences in the interactions of mothers with their children, specifically the types of questions that mothers asked during playtime, related to differences in the way that the children worked on picture-labeling (vocabulary) vs. information-seeking (knowledge) tests. They observed that the Latino children had more difficulty understanding vocabulary tasks than they did understanding information-seeking tasks that elicited descriptions, functions, and explanations and argued that the latter task more closely resembled the way that the Latino mothers interacted with their children.

At school entry, test-wiseness continues to differentially affect students from culturally and linguistically diverse backgrounds. Dreisbach and Keogh (1982) found that test-wiseness training, devoid of test-relevant content, improved the performance of Spanish-speaking children from low SES backgrounds more than would be expected from the general population (d=.46 between trained and untrained students; effect size calculated from their reported data). Even among college students, differences may persist. Maspons and Llabre (1985) found that college-aged Latino students in the Miami-Dade area students benefitted significantly when trained in test-taking skills necessary for U.S. testing systems. In their experimental design, they found that the correlation between their knowledge test and later class performance in the class jumped from r=.39 for the control group to r=.67 for the experimental group that received training in test-taking skills.

Differences in cultural and educational backgrounds appear to affect EL students’ engagement with tests. Budoff et al. (1974) argued that EL students “differ in familiarity and experience with particular tasks, have a negative expectancy of success in test-taking, and are less effective in spontaneously developing strategies appropriate to solving the often strange problems on a test” (p. 235).
The latter point makes high quality test directions especially likely to help EL students if they support strategy development. Budoff et al. confirmed that their learning potential post-test (which allowed more opportunities for feedback on test strategies) created substantial practice gains for EL students: $d=0.68$ for primary students and $d=0.40$ for intermediate grade students. They also found that the post-tests correlated more strongly with achievement for EL students compared to their pre-test scores and that indicators of SES and English proficiency were correlated with pre-test but not post-test performance.

**Perceptions.** There are also important differences in perceptions of tests, too, as suggested by Budoff et al. (1974). Students and parents new to the U.S. school system may differ in their understanding of the purposes of testing and the basic strategies of completing tests (Kopriva, 2008; Reese et al., 1995). For example, Walpole et al. (2005) found that students from low-SES backgrounds enrolled in low-performing schools lacked information about the availability of test preparation materials and vastly overestimated how much improvement could be expected from re-testing. Such differences are believed to impact students’ motivation to perform well on these tasks and to exert effort (Kopriva, 2008; Schwarz, 1971). When the tests have high stakes (such as for gifted and talented identification or college admissions), there are often large differences between White and minority students in the perceived importance of preparing at home for school-administered tests (Briggs, 2001; Chrispeels & Rivero, 2001; Kopriva, 2008; Solano-Flores, 2008; Walpole at al., 2005). Better advanced information about tests and their uses should therefore be targeted to these parents.

**Summary.** Although test unfamiliarity can affect all students when test formats are novel, the literature strongly suggests that there is a possibility of differential familiarity for any test format when there is diversity in the student population. As a result, test developers should take care with the development of test directions and practice activities to address any potential unfamiliarity effects that may influence test performance.

**Best Practice in Designing Test Directions and Practice Opportunities**

Several older measurement texts tout the importance of good directions to valid test use and offer some advice on the features of good directions (e.g., Clemans, 1971; Cronbach, 1984; Traxler, 1951). The
attention to the construction of test directions appears to have diminished over time, as exhibited by the content coverage of successive editions of *Educational Measurement*: Lindquist (1951), Thorndike (1971), Linn (1989), and Brennan (2006a). The first two editions each contained a full chapter on test administration that contained extensive advice on creating the test directions (Clemans, 1971; Traxler, 1951), whereas the latter two editions restricted their discussion to issues of standardization with little specific advice on developing test directions (e.g., Bond, 1989; Cohen & Wollack, 2006; Millman & Greene, 1989).

Even in these early guides, specific guidance for how to create directions was uncommon. To some extent, this is because the best directions differ by test and item types. However, the research also lacks systematic efforts to discover exactly what features make for good directions for a given format especially in recent years when test item formats have been evolving considerably.

Given the infrequent inclusion of guidance on the development of directions in recent measurement texts, test directions may appear easy to construct. However, when actually trying to write directions, it can be quite difficult to explain an unfamiliar task clearly and succinctly so that all examinees fully understand. This is especially true when writing directions for children because even basic concepts must be minimized (Kaufman, 1978). Children have less experience with tests and need simple language and adequate practice. Furthermore, when tests are less familiar and more complex, students of any age may depend more heavily on the test directions to understand the task. From a review of the literature, the following guidelines for creating directions and practice items are suggested. These guidelines are then expanded upon in the following sections.

1. Treat directions as an instructional activity and build in flexibility for additional practice or explanations for students who do not immediately understand.
2. Use the simplest language needed, but no simpler.
3. Use more practice items with greater variety, dependent on examinee characteristics.
4. Provide elaborative feedback to make examples efficient and effective for all.
5. Gather evidence about the quality of the directions to support the intended validity inferences for students who vary in test familiarity.

**Treat Directions as an Instructional Activity**

We know from extensive the work on practice effects (Jensen, 1980; Kulik et al., 1984; Thorndike, 1922) that examinees improve their test scores when repeatedly administered similar tests. What is the examinee learning that results in practice effects on later examinations even when the items are not the same or no feedback is given? For examinees first encountering novel item formats, part of the challenge of novel item formats is that they require examinees to assemble a strategy and rely more heavily on metacognitive monitoring skills rather than well-practiced and automatic strategies (Marshalek, Lohman, & Snow, 1983; Prins, Veenman, & Elshout, 2006). The items may also require specific knowledge about the response format or rubric (Fuchs et al., 2000). Test directions with a number of practice items (or a separate practice opportunity) may be sufficient to help students overcome this initial learning phase.

Depending on the age of the examinees and how much they vary in their experience with the test format (or tests in general), the detail of the directions and practice may need to be adapted at the student or class level so that students are neither bored nor overwhelmed by the length and detail of the directions or practice. This would, in fact, be a sharp departure from typical standardized test directions.

The traditional understanding of standardization is that it refers to the behaviors of the test administrator and the testing environment rather than to the readiness of the examinees to engage in the test. In this behaviorist framework, fairness is achieved when every student hears the same directions and sees the same practice items (c.f., Cronbach, 1970; Geisinger, 1998; Norris et al., 2004). However, when students vary greatly in their familiarity with testing, the typical administration procedures and test directions provided may fail to homogenize the test experience of examinees. For complex and unfamiliar formats, test developers would benefit from broadening their definition of standardization to include homogeneity of the examinee experience and change the focus of directions from standardization to effective instruction.
To accomplish this change, mainstream testing might benefit from incorporating elements of the dynamic assessment and learning potential assessment tradition to adapt directions and practice opportunities. Following these traditions, test direction activities would ensure that every student is on equal standing with respect to understanding the task, motivation to do their best, and familiarity and test-wiseness related to the format (Budoff, 1987; Caffrey, Fuchs, & Fuchs, 2008; Kaufman, 2010; Ortar, 1972; Rand & Kaniel, 1987; Tzurkiel & Haywood, 1992). In some circles, this is called the “Betty Hagen rule”\(^1\): “Do not start the test until you are sure every child understands what he or she is supposed to do.”

Like Hagen, these interactive testing traditions recognize that fairness comes from adaptation to students rather than standardization of administration and that, in fact, equitable treatment can produce inequality (Brennan, 2006b). Hattie, Jaeger, and Bond (1999) capture this tension between standardization and adaptation as the key to fairness in item formats and administration:

“The manner under which tests are administered has become a persistent methodological question, primarily because of the tensions between the desire to provide standardized conditions under which the test or measure is administered (to ensure more comparability of scores) and the desire for representativeness or authenticity in the testing condition (to ensure more meaning in inferring from the score).” (p. 395)

When students vary widely in their familiarity with tests and their readiness to show their best work, fairness is likely best achieved through increased flexibility in test administration (e.g., additional practice for some, changes in directions, comprehension checks) rather than rigid adherence to standardized procedures.

In treating test directions as an instructional activity, the directions must be adapted to examinee characteristics and pay attention to motivation. Young students may benefit especially from basic test-wiseness training, including self-pacing and tactics to reduce impulsive responding. In a study of test

---

\(^1\) Hagen was the co-author of several forms of the Cognitive Abilities Test (CogAT) among many other tests and well-versed in designing tests for children. This is hear-say wisdom from her CogAT Form 6 co-author, David Lohman.
directions for young students, Lakin (2010) found that multiple tactics and teacher interventions were needed within the directions to slow students down and to encourage them to consider all of the response options before selecting an answer. Even young students seem quick to assume they understand the task, consistent with the minimal effect of written directions on older students in an experimental design (LeFevre & Dixon, 1986).

For abstract item formats, such as those used on ability tests and critical thinking tasks, the directions should also try to introduce a basic strategy for solving the early items so that students feel successful and can adapt the strategy as items increase in difficulty (Budoff et al., 1974). Familiarity with successful strategies is certainly an advantage for more experienced test-takers (Byrnes & Takahira, 1993; Marshalek et al., 1983; Prins, Veenman, & Elshout, 2006), but may be minimized by teaching a basic strategy. In this way, the first few test items (also called “teaching items”) should start easy and increase in difficulty so as to lead the child to develop appropriate strategies by gradually introducing more complex rules (Cronbach, 1984). For introducing complex simulations or tools, test developers should similarly plan to introduce tools as needed and preferably one at a time so that the examinee is not overwhelmed with new cognitive demands (Gee, 2003).

**Use the Simplest Language Needed, But No Simpler**

In writing test directions, the primary goal must be clarity in conveying information to examinees. Directions that are too long lead to inattention and directions that are too vague or brief will not serve the intended purpose. There is thus a fine line to walk when every student will receive the same directions. The trend in abilities assessment seems to be towards using the shortest instructions possible (even to the somewhat extreme “nonverbal directions” procedure; Lakin, 2010; McCallum et al., 2001) with the idea that less language is more culture-fair. However, this may not be the best method for closing gaps in understanding between culturally and linguistically diverse students. In fact, shorter directions—where repetitions, explanations, and logical conjunctions are dropped—may actually increase the burden of listening comprehension for examinees (Baker, Atwood, & Duffy, 1988; Davison & Kantor, 1982).
Furthermore, even directions that seem short and clear to the adult test developer may contain an undesirable number of basic concepts when testing young students. In traditional test administration, basic concepts are essential to explaining the task, especially when a practice item is presented. Telling the examinees to “look at the top row of pictures” or asking, “What happens next?” both involve basic concepts that young students may not know. As a result, the directions for most widely used intelligence tests contain many more basic concepts than one would expect or desire on a test for young students (Flanagan, Kaminer, Alfonso, & Rader, 1995; Kaufman, 1978).

From their work in cross-cultural testing, van de Vijver and Poortinga (1992) argue that directions should “rely minimally on explanations in which the correct understanding of the verbalization of a key idea is essential” (p. 20). In their view, visual examples and verbal descriptions must work together with appropriate verbal repetition and feedback to support comprehension so that a single moment of inattention or misunderstanding does not derail an examinee’s performance. Indeed, rather than arguing for short directions, van de Vijver and Poortinga (1992) argue that increasing the amount of instructions and practice can actually reduce the cultural loading of a test. More language decreases how critical it is that students quickly grasp the rules of the task from a few examples. A liberal number of examples and exercises, they say, can overcome relatively small group differences in familiarity.

Biesheuvel (1972) likewise argued that familiarization with test demands is essential before beginning a test in a cross-cultural context: practitioners often fail “to recognize the desirability of prolonged pre-test practice in dealing with the material which constitutes the content of the test and of introducing a learning element in the test itself” (Biesheuvel, 1972, p. 48).

The best compromise therefore seems to be appropriately detailed explanations that are firmly concretized by examples and visuals. Visuals are a way of reducing language load without reducing information (Mayer, 2001). Another solution is to rely on computer animations: one important benefit of computerized tests is the potential for directions to show rather than tell test-takers where to direct their attention. This approach reduces the amount of explanation required and the corresponding language load.
Use More Practice Items

In treating the directions as an instructional activity, the test developer should pay special attention to choosing the best practice items and tactics to teach the students how to respond to the task. For example, the first few items should not provide opportunities to succeed using inappropriate strategies. Misconceptions about the task introduced by the test directions and early test items are a serious threat to validity. For example, Lohman and Al-Mahrzi (2003) found that the use of certain item distractors (that is, the incorrect options to an item) unintentionally reinforced a misunderstanding of the task that was only discovered through an analysis based on personal standard errors of measurement. Likewise, Kaufman (2010) reported that the content of early items on the WISC Similarities tests unintentionally steered students towards giving lower-point answers. Thus, it is clear that test directions, practice items, and early test items must work together to prevent misunderstandings.

The choice and number of practice items is no small consideration in the design of good test directions and practice. In fact, research shows that test developers should design directions and examples keeping in mind that students will rely more heavily on the examples than verbal descriptions to get a sense of the task. LeFevre and Dixon (1986) compared the performance of students on two item formats when examples and instructions for the tests conflicted (e.g., the words described a figure series test but the example showed a figure classification problem). No matter how the experimenters varied the emphasis on the directions, the college-aged students used the example as their primary source of information on how to do item and virtually ignored the written directions.

However, LeFevre and Dixon (1986) warn that “the present conclusions probably would not apply to completely novel tasks; in such cases a single example might be insufficient to induce any procedure at all” (p. 28; see also Bergman, 1980). LeFevre and Dixon argue that examinees need both examples to help concretize procedures for tests and verbal instructions to help abstract the important features from the examples. Feist and Gentner (2007) concur, finding that verbal directions were essential to drawing attention to important features of test items, teaching basic strategies, and giving feedback on practice items. “In speaking, we are induced by the language we use to attend to certain aspects of the
world while disregarding or de-emphasizing others” (p. 283). Although examples by themselves may be less useful to naïve test-takers, clearly, they play an important role in helping examinees develop specific strategies and procedures for solving test items (Anderson, Farrell, & Sauers, 1984).

An additional benefit of using more practice items is that they provide a more representative sample of test items (Haslerud & Meyers, 1958; Kittell, 1957; van de Vijver & Poortinga, 1992). Presenting a range of realistic examples is important. Examples that are too simplified do not expose students to the full range of item types and may lead them to develop ineffective strategies for the task, while a range of examples can introduce strategies and allow examinees to anticipate the range of challenges in the test (Jacobs & Vandeventer, 1971; Morrisett & Hovland, 1959). Furthermore, several studies showed that schemes are better built from multiple practice items than from long explanations (Anderson, Fincham, & Douglass, 1997; Klausmeier & Feldman, 1975; Morrisett & Hovland, 1959; Ross & Kennedy, 1990).

Morrisett and Hovland (1959) also found that presenting a range of different practice items was essential because it taught examinees to discriminate between task-relevant and task-irrelevant problem features. However, they also found limits to how much variety is beneficial. Although they found that extensive practice on just one type of item led to negative transfer on diverse items, they also found that sufficient practice on a set of three item types was more effective than brief practice on 24 item types. Thus, the best approach in selecting practice items is to identify a few broad categories of items and help examinees develop workable but general strategies that can be adapted to different items within these categories.

The relative timing of directions and example items is also a consideration. Lakin (2010) found that the preamble of test purpose and describing the task that many tests use may not be ideal and that good test directions should engage students (especially young students) in responding to practice items almost immediately rather than expecting them to listen to extensive explanations first. In fact, some research has shown that it may be better to let the students see examples before they are given any
description of the task at all in order to make the description more concrete (Sullivan & Skanes, 1971). This puts the onus of explanation on the feedback to responses rather than directions.

**Provide Elaborative Feedback**

Practice or example items can play an important role in demonstrating the format rules to students. However, practice alone does not lead to optimal performance when it is not paired with appropriate feedback. The research indicates that (1) feedback makes practice more efficient for learning and (2) elaborative feedback that goes beyond yes/no verification improves learning further.

**Practice with feedback is more efficient.** The importance of feedback was established early in the history of psychology (Thorndike, 1927). Morrisett and Hovland (1959) showed that feedback increased the amount of learning from each practice item, whereas providing no feedback resulted in the need for many more practice items to produce the same amount of learning. Furthermore, practice without feedback has the tendency to make most students faster but not better at task (Thorndike, 1914, Ch. 14), because such practice leads to automatization of whatever strategy the student is using, whether efficient or not, and not necessarily to improvement in performance.

Several studies support the need to provide feedback for practice items on tests (e.g., Hessels & Hamers, 1993; Parshall et al., 2000; Sullivan & Skanes, 1971; Whitely & Dawis, 1974). In one example, Tunteler, Pronk, and Resing (2008) conducted a microgenetic study of first-grade students solving figure analogies and showed that although all students improved their performance from practice alone, targeted feedback in a dynamic training session greatly increased performance. Importantly, Sullivan and Skanes (1971) found that feedback had a greater effect for low-ability examinees, who gained much less from practice without feedback than high-ability examinees. This latter finding is consistent with research showing that high-ability examinees learn more from practice without feedback because they can more often correct their own errors (Shute, 2007).

The positive effect of feedback for culturally and linguistically diverse students has been confirmed in cross-cultural research as well. Hessels and Hamers (1993) and Resing, Tunteler, de Jong,
and Bosma (2009) found that dynamic testing approaches with tailored feedback and practice significantly increased the performance of language- and ethnic-minority students on an ability test.

**Elaborative feedback increases learning from practice problems.** Several researchers have investigated what degree of feedback and explanations are optimal (e.g., Judd, 1908; Kittell, 1957; Sullivan, 1964). In her review of the literature, Shute (2007) concluded that elaborative feedback provided in tutoring systems resulted in larger gains in comprehension compared to simple verification of the answer. Elaborative feedback could include both a brief explanation of why the chosen answer was correct as well as information on why other attractive or plausible distractors are wrong.

Ideally, the person administering the test should also be aware of what feedback the examinees are receiving as well so that the examiner can determine that the examinee has sufficiently understood the directions (van de Vijver and Poortinga, 1992). This is rarely a component of traditional test directions, but in many cases could be easily added to the administration process (e.g., directing the administrator to walk around during the practice items). Note that examiners would also need to be provided with information on how to provide additional feedback to students who are not completing the practice items correctly. Such comprehension checks should allow the examiner to be confident that all students understand before the test begins. Computer-based administration of tests or practice tests would also be effective in tracking practice item performance and routing students to additional practice.

**Collect Evidence on the Quality of Directions**

Test directions are a critical component of test design that influence both the construct measured and the precision of that measurement (Clemans, 1971; Cronbach, 1970), yet they rarely receive systematic evaluation to confirm that they are effective in reducing pre-existing differences in test familiarity among examinees. Ideally, directions would be evaluated as a part of the validity argument for a test to assure that they adequately familiarize all students with the task. Examples of appropriate research methods for this purpose have been reviewed earlier in this article (LeFevre & Dixon, 1986; Fuchs et al., 2000; Lohman & Al-Mahrzi, 2003; te Nijenhuis et al., 2007; see also Leighton & Gokiert, 2005). Likewise, researchers should evaluate the magnitude of practice effects (overall and for specific
students) as one check to determine how much students gain from changes in test familiarity alone, because this can threaten score interpretation (particularly longitudinal growth estimates).

The age, test experience, and background characteristics of the examinees clearly matters in the design of directions and warrants additional research. First, young students clearly require more orientation and distributed practice opportunities2 to help them understand the test tasks (Dreisbach & Keogh, 1982; Lakin, 2010), although finding the optimal method requires more research. In contrast, the particular issue with older examinees seems to be getting them to attend to the directions (LeFevre & Dixon, 1986), a challenge that has also not been addressed with empirical research.

Second, the background characteristics of the students, whether in terms of cultural differences that influence test-wiseness (Biesheuvel, 1972) or linguistic differences (Dreisbach & Keogh, 1982; Ortiz & Ochoa, 2005), may influence the type and amount of directions and practice that are needed. In schools with large populations of culturally and linguistically diverse students, the adequacy of test directions to create scores that reflect valid comparisons of student performance is an even greater concern and requires additional fairness research evidence. Other background characteristics should also be considered, when appropriate. One example is the Matthew effect, where gains from practice opportunities differentially benefit the most able students, at least on cognitive ability tests (Jensen, 1980; Sullivan, 1964; te Nijenhuis et al., 2007). One estimate of the Matthew effect comes from Kulik et al.’s (1984) meta-analysis, which estimated the effects of practice to be .17SD for low ability students, .40 for middle ability, and .82 for high ability (note that this was only for identical tests—the differential gains were much smaller for parallel forms, d=.07, .26, .27, respectively). Test developers should consider such complications and avoid unintentionally exacerbating disparities.

---

2 It is important to note that teaching test-related skills does not necessarily have to happen in isolation and extend the time spent on test preparation. Integrated instruction/assessment systems, such as the CBAL assessment system (Bennett et al., 2010; see also DiCerbo & Behrens, 2012), embed instruction related to learning the assessment system and tools into meaningful content instruction.
Conclusions about the Design of Test Directions

Complex and novel item formats play an important role in both cognitive ability and achievement testing. Introducing new item formats can yield large familiarity or practice effects when students first encounter the format which can influence evaluation at the individual or group level (Jensen, 1980; Fuchs et al., 2000). When access to test preparation varies, particularly when tests have high individual stakes attached to their results, differential novelty between students becomes a serious threat to fairness and the validity of interpretations from scores as well. In the diverse populations many schools serve, equal familiarity with testing and access to test preparation cannot be assumed.

Based on the literature review, several guidelines for creating directions were substantiated. It was clear that directions should be treated as an instructional activity that engages students in the task, use clear examples, provide elaborative feedback, and use age/level-appropriate language. The literature also suggests that when students vary widely in their familiarity with tests, fairness is likely best achieved through increased flexibility in test administration (e.g., additional practice for some, changes in directions) rather than rigid adherence to standardized procedures.

A critical area for future research is the use of computer-based adaptive directions that could provide the needed flexibility in practice and feedback. Even for paper-based tests, computer-based practice activities could be beneficial in addressing the suggestions for directions and practice made in this article. Computers can respond adaptively to what students can already do and provide more or less practice as needed. Computer-based practice and directions could even provide basic test-wiseness strategies for students who lack any familiarity with tests; training that would be inappropriate and boring for other students. They can also permit the sort of semi-structured trial-and-solution strategy that students commonly use when learning new computer games (Gee, 2003). Computer-based directions could also target the feedback to provide only useful (and brief) corrections for errors made on practice items and support this feedback with appropriate visual demonstrations. A further benefit of computer-based practice would be that examinees could elect to hear the directions in one of several languages.
Regardless of the format of the test, the most important lesson from the literature is that test directions are a vital contributor to test validity, should be systematically studied and designed, and should be evaluated as part of the process of compiling a validity argument. Unless test developers and users are certain that all students have had equal opportunity to understand the task presented, bias may be introduced through systematic differences in test familiarity.
References


