Personnel Assessment Monographs

Test Wiseness: Cultural Orientation and Job Related Factors In the Design of Multiple Choice Test Questions

Volume 2, Number 1
April 1989

ipmaac

Personnel Assessment Monographs is published by the Assessment Council of the International Personnel Management Association (IPMAAC), 1617 Duke Street, Alexandria, Virginia 22314. Copies are provided without charge as a service to members of the Assessment Council. Members of IPMA and others may obtain copies by writing the Director of Assessment Services at the above address. The monographs are intended to provide information on current topics in the field of applied personnel assessment. Manuscripts or proposals for manuscripts are solicited by the editor. Reviews of research, assessment methods, applications, innovative solutions to personnel assessment problems, and related subjects which are not generally available elsewhere are appropriate. Monographs are intended to be of value to practitioners of personnel assessment in such fields as employee selection, performance appraisal, program evaluation, recruitment, organizational assessment, and related fields. Manuscripts are reviewed by the editor and consulting editors, and accepted for publication based on the technical and professional soundness of the manuscript, and the extent to which conclusions and other information is relevant to applications in the field of personnel assessment. The editor seeks to balance topics covered to insure that subjects of interest to all readers are addressed.

Test Wiseness: Cultural Orientation and Job Related Factors In the Design of Multiple Choice Test Questions

Edited by
William E. Tomes

CONTRIBUTORS

Brenda Morefield
Chuck Schultz
Christina L. Valadez

Editor's Preface

How many times have you heard this comment about a job applicant who performed poorly on a test: "He can do the job; he just doesn't do well on tests." On the other hand, there is the employee about whom we hear this comment: "He scored high on your test, but he doesn't have a lick of common sense and can't do the job."

In the three papers published in this monograph, Schultz, Morefield and Valadez explore the reasons why test performance is sometimes affected by factors that are not job related. The first two papers discuss sources of irrelevant test variance, with the final paper examining the effect of irrelevant test variance on test reliability and validity.

The test items appearing in the monograph are used to illustrate a particular point. We have not attempted to evaluate all psychometric properties of the items. In this monograph, our authors are sharing their ideas about a very interesting assessment topic. Others may have different ideas; we would like to hear them.

Finally, we hope you have noticed the change in appearance of the monograph. I want to thank the publications staff of the Institute of Public Affairs at the University of South Carolina, and especially Pinkie Whitfield, the Publications Manager, for their efforts in putting together this monograph. We hope you find this publication more attractive and easier to read.

William E. Tomes
Editor

Types of Multiple-Choice Questions That Malfunction

by
Chuck Schultz
and Brenda Morefield

We often assume that when a candidate fails a test he or she lacks the knowledges, skills and abilities the test is supposed to measure. However, factors besides what the test is intended to measure affect test scores. Because of these other factors, people who know how to handle a situation may not give the "correct" answer to a question about it on the test. Candidates wonder in what frame of reference to respond. They must decide whether to state a solution, obtain more information, or refer the case to someone else.

Over the years we have identified many types of test questions that have not worked as intended. Certain *question formats* result in candidate response patterns that cannot be explained in terms of question content. The formats seem to elicit responses that are more related to "response sets" than to an understanding of the subject matter. Different *candidates' expectations* about the test lead to different response patterns.

Let's look at some question formats that lead to malfunctioning questions and discuss how to improve them.

Negative wording. We might ask, "which of the following is not a factor in..." or "which of the following is least important to...." These produce peculiar results. The candidate may understand the question initially, but, in the process of analysis, the candidate concentrates on the issues and forgets the negative orientation.

Look at question 1 in Table 1. Read through the alternatives and see if you can decide which answer is keyed correct....

As I reviewed this question 14 years ago, I tried hard to figure out which was most important consideration. That's what many candidates did, and perhaps that's what you did.

Some of us got fooled. We forgot to look for the least important as we deliberated on which was most important. It

doesn't matter that least is underlined. With a question this involved, the tendency is to look for the best solution or most important reason. Enough candidates were fooled that the question did not work.

We discovered this phenomenon through strange item-analysis results from a 117-item test with 23 "not" questions. The subject-matter specialists for this test found it easier to write correct answers for some of their topics than to write distractors. The candidates were distracted though. The negative form of the statement seems to have confused the candidates on 16 of the "not" questions. The problem seems greatest on questions with the most complex alternatives.

Notice also that alternative *c* contains the word "cannot". Putting this together with the stem produces a double negative. Double negatives make the question even more confusing.

One smaller and one larger. When writing questions with quantitative answers, novice test writers will include one distractor smaller and one larger than the correct answer. Test-wise candidates who do not know the answer increase their probability of guessing correctly by choosing between the second and third alternatives.

Look at questions 2, 3, and 4 in Table 1. Assuming you do not have a calculator handy and haven't memorized the square root table, question 2 is rather difficult.

But if you are test-wise, you may know about "one smaller and one larger". You narrow the choice automatically to *b* and *c*. Further, if you know that 60 times 60 has to end in zero and 55 times 55 has to end in five, then the answer has to be *b*.

As author Charles Schulz put into the mouth of one of his Peanuts' characters, "If you're smart, you can get the right answer without knowing anything."

You can use the same approach on question 3 about the French Revolution. If you know whether it began before or after 1776 (the date of the American Revolution) you have the clue you need to choose between *b* and *c*.

TABLE 1

"Not" Question

1. In the direct seeding of logged areas, which of the following is the *least* important consideration?

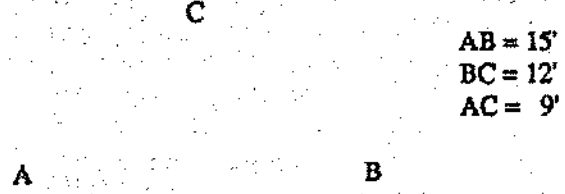
   (a) Indigenous species should be selected for planting.
   (b) Species should be selected based on relative cost of planting stock.
   (c) For purposes of seeding, a black duff layer cannot be considered mineral soil.
   (d) Seed should be applied as near as possible to the period of late November through December.

Quantitative questions: one smaller, one larger

2. What is the square root of 3025?
   (a) 45 (b) 55 (c) 60 (d) 65

3. When did the French Revolution begin?
   (a) 1758 (b) 1773 (c) 1789 (d) 1798

4. What is the area of triangle ABC?



   (a) 48 square feet
   (b) 54 square feet
   (c) 60 square feet
   (d) 67.5 square feet
   (e) some other area

Question 4 is another kind of numerical question that shares the one-smaller-one-larger bias. If I don't know how to solve for the area of a triangle, I can figure out that the area cannot be more than half the product of the two shorter sides. Therefore, I'll pick 54 as the better choice of *b* and *c*.

We can make numerical questions more fair by giving *a* and *d* equal time. "When in doubt picking *b* or *c*" will no longer give the test-wise an advantage. Then all candidates have only one chance in four of stumbling into the correct answer.

In question 4 we have added a fifth alternative, *e*. In numerical questions we include something like "some other amount" to alleviate another factor that interferes with measuring a candidate's ability. Without the *e* alternative, candidates who make the biggest mistakes get a second chance.

Since distractors in quantitative questions are designed to represent the most likely wrong answers, test takers who make reasonable mistakes mark one of the alternatives offered. People who make unreasonable mistakes won't find their answers among the distractors, so they have to try again. Therefore, the person who makes the worst mistake gets a second chance, while the person who makes a common mistake happily chooses one of the distractors we provided, and misses the question. "Some other amount" fits any outlandish solution. We use it as the keyed answer one time in five to neutralize the test-wise.

True-false. True-false questions are sometimes placed in a quasi-multiple-choice form by asking something like, "How many of the following statements are true?" We do not like the connotation of absolutes implied in true-false questions. A statement has to be blatant to be false in every conceivable situation. Candidates differ in judging how true a statement must be to be called true.

Take for example the true-false questions 5 through 8 in Table 2. You can make a case that any one of these is true. You can also make a case for any one of these being false. Question five: There are considerations other than utilization of staff for assigning tasks, so 5 can be false. Question six: While employee preferences should be considered, the organization's mission is more important, so 6 can be false. Statements seven and eight are contradictory, so if one is considered true the other could be considered false.

On questions such as these, whether a person answers true or false depends on issues other than the person's understanding of the issues. For example, it depends on how one interprets the situations. Questions like this sometimes appear on an objective test, but how objective are they?

TABLE 2

True-false

T F 5. A supervisor should assign tasks to maximize utilization of staff.

T F 6. A supervisor should consider an employee's preferences when assigning work.

T F 7. Tasks should be assigned to employees with their career development in mind.

T F 8. A task should only be assigned to an employee who has been trained to do it.

9. Which of the following would you use as the basis for assigning tasks to workers?

(a) Effective utilization of staff
(b) The workers' preferences
(c) The workers' career development
(d) Which workers can do the task without further training

All of the above

10. Which of the following would you use as the basis for assigning tasks to workers?

(a) Effective utilization of staff
(b) The workers' preferences
(c) The workers' career development
(d) Which workers can do the task without further training
(e) All of the above

*** Effective format

Can we make the true-false questions work objectively by incorporating them into a multiple-choice form? In Question 9, you must balance the four alternatives against one another. Here, we would key *a* since, compared to the other alternatives, it is the important consideration. From a management point of view, the other considerations are only viable if they help achieve *a*.

The asterisks you see to the left of question 9 imply that this is a question that functions well. Question 9 puts truth and falsity into perspective. Asterisks used in the tables signify solutions to malfunctioning items. There is an X drawn through the asterisks at question 9. Our reviewers did not all agree that we had demonstrated a successful format with it. There is some correctness in each of our distractors. Perhaps too much. The point we want to make is that there is always some truth in distractors. By pitting them against one another, you define a point of reference for determining how true is true.

"All of the above". Test writers often use this alternative when they can't think of three wrong answers. You find it keyed correct much more often than the alternatives teamed with it. On a test for which "all of the above" is keyed correct only one time in four or five, candidates still choose it much more often.

We put ten five-choice all-of-the-above questions among other questions on a classroom test. On eight of these questions, an alternative other than "all of the above" was keyed. On each of these, candidates chose "all of the above" more often than the other distractors. The item analysis showed that these distractors were not measuring the common factor, achievement in the class. "All of the above" attracted both poor students and students who were creative enough to imagine the distractors as true. Or is it that they attracted students with the response set to say "all of the above"? Or those who were test-wise and who know that test writers usually include "all of the above" only when they can't think of three wrong answers?

"All of the above" shares with true-false questions the necessity of determining which statements are absolutely true. Now let's look at question 10, where we take question 9 and add "all of the above". If we can define a correct answer in 9, wouldn't the same answer be correct for 10? But statements 6, 7, and 8 can be considered true as well. Therefore, alternatives *b*, *c*, and *d* can also be considered correct. With "all of the

above" as a choice, the alternatives are no longer balanced against one another. What answer a candidate chooses may have less to do with understanding the issues than with the decision rule used.

Social Desirability. The social desirability response set has been studied extensively in personality tests. Social desirability is active in multiple-choice tests as well. All too often the correct response is clearly the most socially acceptable thing to do. In questions 11 through 13 in Table 3, we leave off the item stems and present only the alternatives. As you read through those alternatives you will probably see that some of the actions are quite socially desirable. The numbers in the left hand column show how many candidates picked each alternative. We had data for 130 candidates for Questions 11, 12, and 13.

Almost all of the candidates chose the socially desirable response in 11, 12, and 13. On question 11, there were some candidates who believed in confrontation as well.

For those questions, the socially desirable response was the keyed answer. The problem we see here is that you don't have to respond to the question itself--you don't even need the stem to get the question correct.

In multiple-choice test questions, we have seen that often the keyed response differs from one of the distractors only by the social acceptability of the phrasing. Question 14 is such a question. Alternatives *a* and *b* are two different ways of saying "do nothing", one of which is more attractive than the other. Alternatives *c* and *d* are two ways of saying "ask her to be nice." This is a two-choice question. Candidates will select either *b* or *d*.

On questions 15 and 16 in Table 4, the problem takes on a different hue. The subject-matter specialists told us that Case-workers need to know when to close the case. They created situations in which you have done everything you are supposed to do for the client and there is nothing more you legally can do. The keyed answer for 15 is *a* and for 16 it is *b*. However, at least on the test, candidates find a variety of services that are preferable, more socially desirable. If they have the option of saying they would close the case or saying they would do something more friendly, candidates pick the more friendly answer.

Are the few people who said they would close the case the

best candidates? That is not clearly so. The social desirability response set seems to be working against us.

TABLE 3

Social desirability. Without the stems, can you tell which of these alternatives are keyed? The number choosing it is shown to the left of each alternative.

11.
- 0 (a) Make sure you are just as forceful as he is.
- 84 (b) Try to understand what is causing him to behave this way.
- 4 (c) Ask him for equal time to answer his complaints.
- 42 (d) Confront him about his behavior and the effect it is having.

12.
- 2 (a) Tell the person you can't help.
- 2 (b) Refer the person to someone else.
- 123 (c) Seek the answer among other resources.
- 3 (d) Give the person the best answer you have with the information currently available.

13.
- 118 (a) You understand her point of view.
- 1 (b) If she calls often enough, she will get help.
- 10 (c) We will do whatever she asks.
- 1 (d) If she treats us well, we will treat her well.

14. You have several reports that Jane, who reports to you, has been rude and insulting to her co-workers. How would you handle the situation?

(a) Ignore it for now.
(b) Wait to see whether this becomes a problem.
(c) Ask Jane to watch her mouth.
(d) Talk to Jane about being more tactful with her co-workers.

Question 17 avoids the social desirability response set while testing for whether a candidate will close cases when they need to be closed. Here we have not pitted "control the caseload" with "be nice". Instead we matched four ways of controlling it.

In each of the types of questions we have referred to, there are characteristics other than content that influence candidates' responses. Solution: Avoid formats that, regardless of question content, elicit systematic responses. A second solution: use a mixture of formats so as not to put all of a candidate's eggs in one basket.

Test characteristics produce what Campbell and Fiske called method variance in their landmark 1959 article on the multi-trait multi-method matrix. When they measured the same factors through different methods they found the resulting correlations to be the result of both content and method.

In some cases the correlation between trait A and trait B measured by the same method is higher than the correlation between trait A measured by one method and trait A measured by a second method. As an example, when a supervisor rates employees on leadership and technical knowledge you may get a high correlation between the two variables. When you measure leadership by supervisor ratings and leadership by a multiple-choice test you may get a low correlation. The high correlation when two unrelated variables are measured by the same method is attributed to method variance.

Candidates' expectations. We have been talking about response sets in terms of question formats. Now let's look at them from the point of view of the candidates. We must acknowledge that a test is an artificial situation and that the implicit ground rules may not be universally understood.

Questions are often presented such that the candidate has to make assumptions about the level on which the question should be answered. Must the candidate give a solution or decide whether to solve the problem or let someone else solve it?

Usually the candidate will enter the test with the set to solve the problem. Often the test writer may have something else in mind. The test writers may expect the candidate to do something such as:

gather additional information
leave the situation to the proper authority
refer it to an expert

let the client decide
help the client decide
ask for guidance

TABLE 4

Low desirability of correct choice

Suppose you should close the case?

15. 3 Tell Louisa you intend to close her case.
17 Ask Louisa what she sees as solutions.
6 Offer Louisa suggestions about what she can do.
4 Schedule follow up visits to see how Louisa is doing.

16. 2 Arrange housecleaning for Mrs. Morris.
4 Since Mrs. Morris is not in a dangerous or negligent situation, close the case.
17 Schedule a follow up to see how Mrs. Morris is doing.
7 Contact Mrs. Morris' family and enlist their help.

***

17. Suppose your caseload has become heavy and as you now manage it you cannot adequately service all of your cases. If all of the following options were available to you, which one would you do to make it possible to better perform your casework duties?

(a) Close cases that are not technically eligible for further services.
(b) Limit the hours each day you will be available to take phone calls.
(c) Set aside cases in which the client is hostile or uncooperative.
(d) Work on only cases with the most severe problems.

*** Effective format

Using the same words in every test booklet does not ensure that all candidates have the same question. The words mean something different to each of us.

We may want to see whether the candidate knows that acting without more information is premature. We expect the candidate to know that, this time, more information is needed. Other times we fail to provide all the information one would have on the job and expect the candidate to extrapolate. How can the candidate tell which is the case on a particular question?

Look at question 18 In Table 5. We have given you some information about Carl and his family. Do you have enough information or should you gather more before charting a course of action? We find that some good caseworkers choose to solve the problem on what we have given them, while others feel they need to have more information. To put all candidates on the same wave length, make the alternatives parallel.

Frame the question one way or the other. If you want to find out about the candidate's ability to determine what information is needed, ask about information, as we do in question 19. If you want the candidate to make a decision with the information at hand, give as alternatives various courses of action, as we do in question 20.

What if the worker should do nothing? In some situations one should wait until the time is ripe, but the candidate expects, "They must want me to do something now, or they wouldn't have included this question in the test."

A subject-matter specialist told us, "In this job it is important for the incumbent to be patient." So we wrote a "be-patient" question something like question 21 in Table 6. Since you know our rationale, perhaps you will accept *d* as the correct answer. But, empirically, the question did not work. Candidates, even the better candidates, came up with creative solutions. They may act like stodgy bureaucrats once they are in the job, but on the test they are proactive and innovative.

Question 22 may be a better way to see whether a manager will expend resources on a program enhancement that has not been funded. While we are discussing "do nothing", let's again look at the problem of closing cases when the objectives have all been met. In questions 15, 16, and 17, we were discussing "close the case" in terms of the social desirability of alternatives. In the spirit of considering the multiple influences on test items, it is

relevant to consider that issue here also. The candidate may perceive that we want to solve a problem for the client rather than adhere to our bureaucratic principles. But when we wrote those questions, we had bureaucratic principles in mind. How is the candidate to know?

Should you always follow rules and procedures? What if following the letter of the law in a particular situation is unreasonable? Candidates may believe that we would expect them to say that they would not break the law. In a Liquor Enforcement Senior Officer test, we had a situation in which the subject-matter specialists and most candidates thought the candidate should break the law.

The situation was: "The law says you may not sell liquor to a certain class of business unless the customer shows you authorizing identification. A regular customer whom the liquor store manager has done business with for ten years comes in five minutes before closing time. The customer needs a case of scotch. The customer left the authorizing identification at home. What should the manager do?"

The candidate is put into a quandary. Regardless whether the candidate thinks it is proper to be lenient or hard-nosed, the candidate's chore is to decide whether the test writers want the reasonable response or the lawful response. What does this question measure? As a result of an appeal, a hearings examiner canceled the register built from this test.

Responsibility to solve the problem. I talked to a subject-matter specialist about a Contracts Specialist 2 essay test she had worked on. She insisted that it was a great test. However, she said, they had a couple of excellent candidates who failed it on the first try. But after the subject-matter specialist coached them on what approach to take, they had no trouble with the test the second time. The test wasn't written well enough that candidates could pass without being coached about what approach to take.

TABLE 5

Get more information or solve the problem?

18. Carl is on your caseload. Carl's foster parents call and ask you to remove him from their home. Carl got into a fight with their son at school and knocked out the son's tooth. What would you do?

(Non-parallel alternatives)

(a) Find another placement for Carl.
(b) Refer the family for mental health counseling.
(c) Gather more information before you take action.
(d) Have a meeting with Carl and his Foster family to work out their differences.

***

19. (What additional information would you gather?)

(a) Find out who has financial responsibility for the dental work needed by their son.
(b) Determine the age of the foster parents and how long they have been providing foster care.
(c) Find out more about Carl's adjustment in the foster home and the circumstances of the fight.
(d) Determine whether your co-workers have had any cases in which violence led to injury of a member of the foster parents' family.

***

20. (How would you focus your treatment?)

(a) Mediate a reconciliation.
(b) Find another placement for Carl.
(c) Refer Carl for mental health counseling.
(d) Hold the foster parents to their contract.

*** Effective format

TABLE 6

Take no action

21. You have pilot tested a procedure for an experimental program and it works. Your supervisor is seeking funding for the program. It will not be funded for at least six months, if it is funded at all. You have a full workload. What should you do about the experimental program?

    (a) Write a report discussing anticipated results.
    (b) Adjust workload to fit the program into current resources.
    (c) Arrange periodic meetings of interested people to keep up enthusiasm.
    (d) Be patient but keep interested people informed of any developments.

When should you take action?

***

22. To which one of these non-funded activities would you devote staff and other resources?

    (a) Responding to requests from citizens for information about your rules and procedures.
    (b) Collecting signatures for a petition that would lead to improvement of procedures in your agency.
    (c) Double checking the work of another unit that has been damaging the reputation of your agency through careless errors.
    (d) A successfully pilot tested program your supervisor has requested funding for, but which will not be funded for at least six months.

*** Effective format

Another question asked how the candidate should handle a situation. For a Contracts Specialist 2, the appropriate response

was to notify a higher authority. Instead these excellent candidates told how the situation should be handled. Shame on them! Or shame on the test writers?

Question 23 is an example of a multiple-choice item that forces the candidate to choose between solving the problem and referring the case to the proper authority. I believe the answer is *b*, but many candidates may think we want them to do something positive rather than pass the buck. Again, we should make the alternatives parallel, and either give four ways to solve the problem, or four ways to get someone else to handle it. Question 24 presents alternatives at a Clerk Typist's level of involvement. Question 25 deals with how to handle the problem, but it is not directed to a Clerk Typist.

Multiple-choice questions need to be stated in such a way that each candidate will be able to see the level on which the question should be answered. To this end, the alternatives should be parallel. Should the candidate collect more data, refer the case to someone else, or close the case rather than solving the problem? Make it clear whether the candidate should select a solution to the problem or a way of dealing with the case preparatory to formulating a solution.

In conclusion: we need to ensure that the test score results from intended content not from incidental factors. Don't use faulty formats. Use parallel alternatives. Phrase the questions so that candidates know what tasks to address.

TABLE 7

Responsibility to solve the problem

23. You are a Clerk Typist in a social welfare office. You answer a call from a distraught client who has been receiving public assistance, but whose grant has been terminated. The client will be evicted tomorrow if assistance is not continued. What would you do?

    (a) Advise the client to appeal the decision.
    (b) Transfer the call to the client's caseworker.
    (c) Say you will do what you can to get the assistance continued.
    (d) Discuss the client's options for handling any financial difficulties.

***

24. The clerk gets the call . . .

(a) Transfer the call to the client's caseworker.
(b) Have someone look into the situation and call the client back.
(c) Say that you will pass the message along to the proper authority.
(d) Set up an appointment for the client to come in and talk to the caseworker.

***

25. The caseworker gets the call and has determined that the client is not eligible for any assistance . . .

(a) Advise the client to appeal the decision.
(b) Say you will do what you can to get the assistance continued.
(c) Discuss the client's options for handling any financial difficulties.
(d) Explain to the client that you can do nothing more to help.

*** Effective format

Problems of Bias and Test-Wiseness in Measuring Oral Communication and Problem-Solving Skills Through Multiple-Choice Items

by
Christina L. Valadez

The first paper addresses structural features of multiple-choice items and answers that can affect accurate measurement of candidates' abilities. The second paper discusses how the content of multiple-choice items and answers may, in some instances, result in measuring candidates' style rather than the ability it was intended to measure.

Although varying in degree of importance from job to job, good communication skills are identified as an important element in almost every job analysis we conduct. There are several aspects of communication skills relating to style and verbal socialization that are not typically considered when developing multiple choice items, and yet have the potential of affecting test results.

The test writer faces two challenges in testing for these skills. The first is to get the subject matter specialists to define what constitutes good communication for their jobs, and next to determine how best to measure their definition of "good communication." This is particularly challenging when part of the communications skills needed are oral communication skills, yet the testing format is to be multiple-choice. We face this dilemma when we need to conduct continuous or frequent testing for large numbers of candidates in different geographic areas.

The solution we typically rely on is to present a situational problem involving verbal interaction, and ask the candidate how to best solve this problem. A number of verbal strategies are offered as alternatives, and candidates are asked to choose the one they believe is the best response to the situation described. This approach typically assumes some measure of problem-solving ability, another element prevalent in most job analyses, as well as

"oral communication skills." Depending on the level and nature of the jobs, other elements, such as "interpersonal skills," "dealing with the public" or "supervision" may be part of such a situational item. The essence of such items, however, remains "how to effectively respond to a communication problem in a given context."

In an effort to anchor the responses the subject matter specialists identify as important to on-the-job performance, we ask them for behavioral examples. How does your best performer respond? How does a poor performer respond? We use their answers to those questions to build our keys and distractors.

But are subject matter specialists really providing us observations of successful oral communication strategies? Or are they instead providing us examples of their own style, or perhaps their assumptions of a strategy they believe produces the desired outcome?

It is interesting and instructive to compare how we attempt to measure oral communication skills in a multiple-choice format with how we measure more quantifiable skills, math for example. When we present a math problem to solve, we focus on the end result. Any given problem may allow numerous ways to work out the answer. My personal observations have shown differences in the process used across generations due to changes in teaching methods, and also differences due to the various teaching methods in different countries. The validity of different approaches is recognized through testing for the ability to correctly reach the final result, rather than testing for knowledge of a particular process.

However, when using multiple-choice testing for oral communication skills, by anchoring responses to behavior, subject matter specialists proclaim that they are measuring the knowledge of the process rather than the ability to attain a successful outcome. It is this assumption of the superiority of one approach in producing the desired outcome that may present problems due to differences in socio-cultural orientation, or due to test-wiseness.

After reviewing a wide variety of multiple-choice exams, some response patterns emerge for many of the items dealing with verbal interaction. I discovered one example by exploring with Agency X the reasons behind the low multiple-choice scores for a group of promotional candidates that were considered to have great management potential, and the high test scores of a

group of promotional candidates whose performance ratings were low. Analysis of answers revealed a tendency on the part of the good candidates with low scores to choose an authoritarian approach to resolving interpersonal supervisory situations rather than the participatory approach, which was keyed. The high scoring poor candidates chose the keyed participatory approach.

Interestingly enough, the overall management style of this agency tended to be more authoritarian than participatory. Although subject-matter specialists agreed that the participatory approach was best, the decision was not based on their observations of current and past job performance so much as on what was currently thought to be ideal. Test-wise candidates who knew the ideal could respond correctly to these questions; those who relied on the observed behavior of those their agency considered good performers did not.

A similar problem can occur when relying on an organization's verbal behavioral norms for keying a particular aspect of the communication process as "correct." Besides individual and organizational differences in communication style, socialization in what constitutes appropriate communicative behavior varies across ethnic, gender, geographic, and socioeconomic lines. Some examples of distinct understandings of appropriate oral communication from the sociolinguistic and organizational culture literature point out the depth of some of these differences.

First, an example of gender difference. Key (*Male/Female Language;* 1970) and others, in analyses of women's and men's speech patterns, conclude that many women have been socialized to use a more "polite" or "correct" form of speech than what one expects to hear from men. This includes a higher frequency of indirect expressions particularly in imperative construction where questions are often used in place of a statement (e.g., Should we leave now? rather than Let's leave now). A higher frequency in the use of confirmatory tag questions (e.g., It's time to leave now, *isn't it?*) has also been noted.

Indeed, part of the current literature directed towards professional women speaks to some of the communication differences that have resulted from the differing verbal socialization (e.g., Harragan's *Games Your Mother Never Taught You,* Gilligan's *In a Different Voice*). Although the focus of this literature seems to be teaching women a different communicative behavior, social changes such as those that have occurred in the work place over the past two decades typically produce linguistic

shifts on the part of all participants--not just one segment. Therefore, observant, or test-wise women will likely know which of four alternate choices offered to a multiple-choice question is "correct" in the context of the traditional work environment. They may or may not find this "correct" approach to work most effectively for them on the job.

Another example comes from Patricia Clancy's article, "The Acquisition of Communication Style in Japanese" (1986). She documents the efforts of Japanese mothers to teach their children how to express themselves, particularly their desires, in an indirect manner, and how to interpret the indirect requests of others. This focus on indirect expression contrasts sharply with the expressive values of directness found in many of our test items. Again, test-wiseness or other awareness of norms calling for directness will lead a candidate to the key, regardless of whether the candidate believes in or uses directness as the "better" strategy.

Assuming the superiority of verbal strategies involving a high degree of explanation may also exclude candidates for the wrong reasons. Shirley Brice Heath, in a recent article titled, "What no bedtime story means; narrative skills at home and school" (1986) shows the results of socioeconomic differences in early linguistic socialization of children. She compared the language socialization patterns of adult-child interactions, in three neighboring communities--a mainstream middle-class community and two neighboring communities--one working-class community whose livelihood depended on a textile mill, and the other an ethnic working class/farming community whose livelihood was derived from a combination of farm and mill work. Major differences were found in how adults taught children to verbalize information.

In "Maintown", the middle-class community, she found a high level of usage of written narratives, labeling, and what-explanations, including running commentaries linking new knowledge to old knowledge in the communicative socialization patterns of preschool children. Children are explicitly taught to label two-dimensional book representations and relate them to their "real" three-dimensional counterparts through repeated questions such as "what is this called?; where have you seen this?" The child's reply is then confirmed and expanded upon, with further, often detailed explanations of the topic at hand. In learning and testing situations, they encounter later at school, the

what-explanation they were taught is replayed in such activities as picking out topic sentences, writing outlines, and answering standardized test questions. There is a tight linear order to the instruction, questioning, and expected answers to the questions or description of events.

In "Roadville," the working class community, there is a reduced reliance on written narrative in daily life. Children are taught to label items and provide what-explanations, but adults do not continually link old and new knowledge for them. Children are not coached on the similarities or differences between literary and real events. Activities are modeled to children rather than being verbalized in great detail as in "Maintown."

"Trackton", the farming community, provided an even greater departure from "Maintown" verbal socialization. There were few bedtime stories, and a limited use of written materials in other aspects of life as well. There was instead, more emphasis on nonverbal expression, and the ability to narrate events. Labeling objects was not a factor in children's verbal socialization, and explanations requested of the children by adults were for reason-explanations rather than what-explanations. Adults believe children "come to know" how to speak, in contrast to the focus of *teaching* children to verbalize found in "Maintown." Nonverbal expression in toddlers is reinforced more than their attempts at verbal expression, which is viewed as simply noise. Verbal behavior as well as other activities are modeled rather than explained or taught. As children learn to speak, they are asked more analogical questions requiring nonspecific comparisons of one event, item, or person with another rather than the "what-questions" used in the other two communities. (Ex: "What's that like?" rather than "What is that?")

The data gathered were compared to grade school student performances and the school expectations of literacy learning and narrative style. The school expectations fit only one of the three communication patterns found in the area (the mainstream). Those children whose linguistic socialization best matched the linguistic expectations at school, not surprisingly, had the best academic and test performance. The article concluded that the academic environment could benefit from using a variety of approaches to verbal strategies, rather than maintaining the expectation that all students conform to one standard of expression and learning.

These same differences and conclusions also can be relevant

in considering expectations regarding verbal behavior in the work place and in testing. Test performance can be affected by different candidate expectations regarding the level and amount of verbal interaction, types of explanations needed, and assumptions regarding narrative style, listening behavior, and what information can remain implicit *vs* what needs to be explicit.

We deal with this to a certain extent by reducing reading difficulty, and attempting to provide complete and clear information in all test items. However, different distractors may continue to appeal to candidates of differing backgrounds as long as, in their experience, those approaches have proven successful.

Testing for the verbal behavioral process, then, rather than for the final outcome or for considerations for reaching the final outcome, may therefore have the effect of testing for verbal socialization patterns. How closely the applicant's patterns of verbal interaction conform to the ideals of a certain organization is likely to be reflected in the test score. This is not the same as testing a candidate's *ability* to communicate orally in the way necessary to do the job well.

How can we test for this ability without unnecessarily excluding good candidates? Oral communication is a complex web of vocabulary, grammar, structure of narrative, nonverbal cues, social cues, and paralinguistic speech features such as accent, use of "fillers" (hm, uh), rhythm and speed, etc. All of these features combine and interact. Two speakers of the same background share these features and therefore are likely to derive the same meaning from an interchange. There is no doubt that differences in the interpretation of these features can increase the potential of miscommunication.

Consider the following example of an interchange between a Greek employee and an American supervisor as a demonstration of how differing assumptions in verbal interactions can lead to misunderstanding. First it is pointed out that varying degrees of authoritarianism are perceived as appropriate for the supervisor. As we read the example, it is clear that although the supervisor and the employee understand each others' words, they misunderstand the meanings. This misunderstanding leads each of them to a negative evaluation of the other.

This example is quoted from Harry C. Triandis and is taken from the files of a Greek psychiatrist. As background information it is important to remember that Greeks perceive supervisory roles as more authoritarian than Americans who prefer participatory decision making. Read the verbal conversation first then the attributions being made by the American and the Greek.

Verbal Conversation	Attribution
American: How long will it take you to finish this report?	**American:** I asked him to participate. **Greek:** His behavior makes no sense. He is the boss. Why doesn't he tell me?
Greek: I do not know. How long should it take?	**American:** He refuses to take responsibility. **Greek:** I asked him for an order.
American: You are in the best position to analyze time requirements.	**American:** I press him to take responsibility for his own actions. **Greek:** What nonsense! I better give him an answer.
Greek: 10 days.	**American:** He lacks the ability to estimate time; this time is totally inadequate.
American: Take 15. Is it agreed you will do it in 15 days?	**American:** I offer a contract. **Greek:** These are my orders: 15 days.

In fact the report needed 30 days of regular work. So the Greek worked day and night, but at the end of the 15th day, he still needed one more day's work.

Behavior	Attribution
American: Where is the report?	**American:** I am making sure he fufills his contract.

	Greek: He is asking for the report.
Greek: It will be ready tomorrow.	Both attribute that it is not ready.
American: But we had agreed it would be ready today.	**American:** I must teach him to fufill his contract. **Greek:** The stupid, incompetent boss! Not only did he give me wrong orders, but he does not even appreciate that I did a 30-day job in 16 days.
The Greek hands in his resignation.	The American is suprised. **Greek:** I can't work for such a man.*

*Excerpt from Harris, Philip R. and Robert T. Moran, *Managing Cultural Differences:* 1987: 78-9

This interchange contains a series of events that could easily be found in supervisory communication test items. For example, we might describe the problem as how to make sure an employee submits a report on time. We might then ask "Which of the following would be the best approach?"

(a) Determine a reasonable deadline and inform the employee.

(b) Tell the employee what needs to be done, and monitor progress.

(c) Explain the project and agree on a deadline with the employee.

Although any of these approaches could attain the desired outcome, alternate (c), the participatory approach, would most likely be seen as the *best* approach and, therefore, keyed as the

"correct" answer. Yet in the situation described, this strategy did not work. There will be employees, like the Greek, who have a different understanding of how a supervisor should communicate deadlines to employees. And what about employees of organizations such as the previously mentioned Agency X, where expressed values do not match actions? What is likely to happen in a test situation is that these employees, if they choose the response reflecting what they would do, will choose a distractor, while candidates who are test-wise and know what should be done will choose the key.

One could argue that the candidates choosing such distractors would not function as well on the job as those whose verbal styles better matched the organization's. Yet in the example provided, how well did the supervisor carry out the supervisory task? The "right" approach did not work.

One of the most interesting features of human communication is not the knowledge of a particular set of rules, but the ability to learn and adapt. Those who are skilled in the art of oral communication can use communicative differences and resulting miscommunication as a source of expanding their understanding, and can adapt to new interactions.

We adapt daily to different modes of communication between work and home environments; between co-workers and the public. Every time we move into new social env. onments, we begin to learn new ways of interacting with others. How well or how quickly this is accomplished varies from individual to individual. It is this variability that is a truer measure of oral communication skills than knowledge of a preferred communication model. Do current multiple-choice items presumed to measure good communication skills test for this variability? I strongly suspect that most do not.

We frequently receive comments from candidates that depending on circumstances which we have not addressed in the multiple-choice item stem, they could choose any of the distractors offered as the best response. We tell candidates to rely solely on the information provided to choose the best response. Yet there is so much paralinguistic information (e.g., tone, volume, word spacing, etc.) and nonverbal information (stance, gestures) not to mention social information (individual history, rank, relationships) that we take into account both consciously and unconsciously. Indeed, training in management and communication encourages us to consider numerous factors in communicating

with different individuals, rather than always using what falls into our own communicative comfort zone. After much reflection, I am inclined to agree with the candidates who say "it all depends."

To summarize, given some of the problems outlined, how valid is it to test for current knowledge of the norms of appropriate verbal behavior in a particular environment? Even if we argue that current superior workers conform to the organization's communication style or values, how job related is a reliance on one approach, given the diversity of the ever-changing modern workforce? What are other alternatives? If we need to rely on a multiple-choice format, how can we better test for true communicative abilities?

In oral exams, we can be much more flexible about crediting a variety of approaches that will achieve the desired outcome. In multiple-choice tests with only one allowable "correct" response, testing for these skills is much more problematic.

Perhaps we need to focus multiple-choice items more on the criteria for achieving desired outcome of communicative problems rather than a "correct" process. And, of utmost importance, we need to make sure the SMS' description of communication problem solving goes beyond their perceived reality based on norms, to the factual observation that we request of them.

I hope through these means we can develop multiple-choice items that will work better to select the best candidates from a diversity of backgrounds and avoid the test-wise who simply know the rules.

References

Brice Heath, S. (1986). What no bedtime story means: Narrative skills at home and school. In B. B. Schieffelin and E. Ochs (Eds.), *Language socialization across cultures* (pp. 97-126). New York, NY: Cambridge University Press.

Clancy, P. M. (1986). The acquisition of communicative style in Japanese. In B. B. Schieffelin and E. Ochs (Eds.), *Language socialization across cultures* (pp. 213-250). New York, NY: Cambridge University.

Gilligan, C. (1986). *In a different voice.* Cambridge, MA: Harvard University Press.

Harragan, B. L. (1977). *Games your mother never taught you: Corporate gamesmanship for women.* New York, NY: Warner Books.

Harris, P. R. and Moran, R. T. (1987). *Managing cultural differences.* Houston, TX: Gulf Publishing Co.

Key, M. R. (1975). *Male/female language.* Metuchen, NJ: Scarecrow Press.

Schieffelin, B. B. and Ochs, E. (Eds.). (1986). *Language socialization across cultures.* New York, NY: Cambridge University.

Irrelevant Reliable Variance

by
Chuck Schultz

The first two papers have discussed the test-item characteristics and linguistic practices that affect test variance. Method variance, test-wiseness and cultural bias are unwanted sources of variance, irrelevant to the purposes of the test. In this paper, I want to distinguish among irrelevant, relevant, and random variance and how the different components of variance affect test reliability and validity.

First, I contend that the more irrelevant test variance you have, the higher the reliability. Anything that increases total variance relative to random error increases reliability. Reliability tells how consistently the test measures whatever it measures.

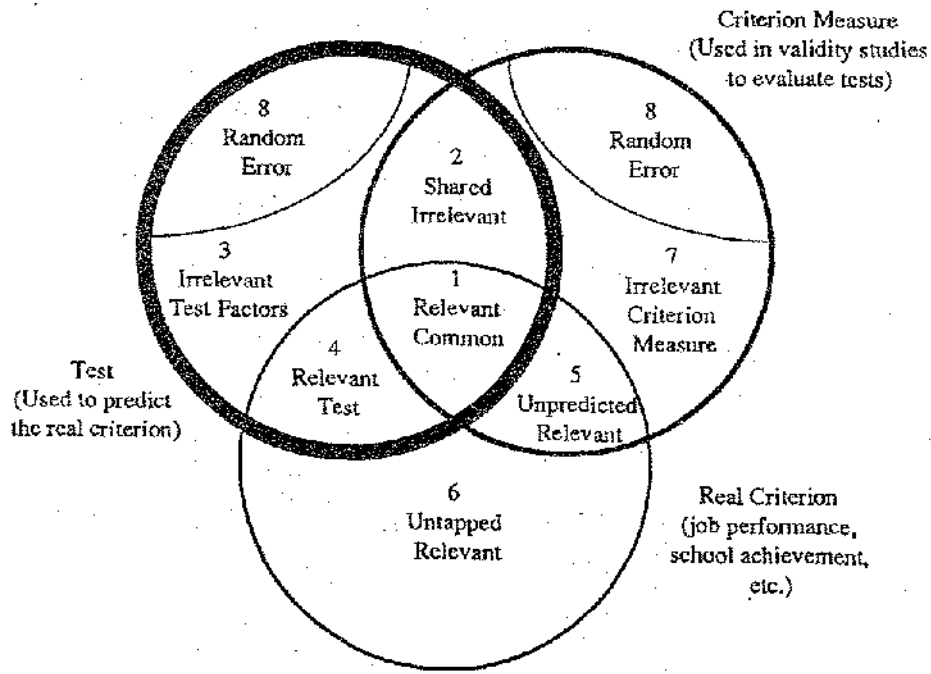
Second, I contend that having the *same* bias present in the test and the criterion measure inflates the validity coefficient. If *correlated* biases are present, the same thing will happen. For example, if for some erroneous reason a rater thinks group A members can't do the job well, and for another erroneous reason group A members do poorly on the test, you have correlated biases. Two wrongs make an enhanced validity coefficient. I emphasize validity coefficient as only one indication of test validity.

A validity coefficient is the correlation between a test and a criterion MEASURE. The criterion measure may or may not be an adequate reflection of the criterion (for example, job performance). You may use any one of a number of criterion measures in a validity study, each of which measures something different. You could use measures as diverse as number of units produced, supervisory ratings, or attendance. Each criterion measure gives you a different validity coefficient. The criterion measures probably overlap with one another and each probably overlaps with the hypothetical real criterion. Let me illustrate the relation between variance components and reliability and validity.

Table 8 pictures the components of variance in a test, a criterion measure, and a hypothetical pure criterion. Let's say you built a test to predict job performance and you designed the

Table 8
Reliable Variance

Variance components in a test, a criterion measure, and a hypothetical real criterion.



Reliability (How consistently a test measures whatever it measures)

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_e^2}$$

Validity (How well a test predicts a criterion measure)

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}} = \frac{\sigma_1^2+\sigma_2^2}{\sqrt{\sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2+\sigma_e^2}\ \sqrt{\sigma_1^2+\sigma_2^2+\sigma_5^2+\sigma_7^2+\sigma_o^2}}$$

criterion measure to check the validity of the test. The criterion itself is a hypothetical construct -- it is the quality of job performance that we are trying to measure with the test and the criterion measure.

In the table, the dotted circle stands for this "real" criterion, the heavy circle is the test, and the light solid circle is the criterion measure. Different variance components are represented by the numbered segments of the diagram. Segments 1, 4, 5, and 6 fall within the dotted circle representing the real criterion. These segments are what we were trying to measure, so I call them relevant variance. Segments 2, 3, and 7 contain the various factors that we were not trying to measure, but that, nevertheless, consistently affect test scores or criterion measures. These are the main topic of the paper: irrelevant reliable variance.

The two segments numbered 8 are random error. How we implicitly define random error depends on how we measure reliability.

Table 9 names the variance components and lists some of the variables that influence them. The numbers of the various components are the same in both Tables 8 and 9. The part of variance that accounts for the validity coefficient is the football-shaped portion of Table 8 made up of segments 1 and 2, which is formed by the overlap of test and criterion measure. Segment 1 is the relevant part and segment 2 the irrelevant part of the variance common to the test and the criterion measure. This common variance is responsible for the correlation between the test and the criterion measure; that is, the validity coefficient.

Those characteristics of the examinees that are reflected in both the test and the criterion measure cause these variance components. Everything the test and criterion have in common that isn't job-related appears in segment 2. For example, a characteristic reflected in method variance on the test may also be reflected in a rater's perception of job performance. Having a large vocabulary may result in a higher test score and may lead to a higher criterion rating, while it may be "really" irrelevant to the quality of job performance.

The other part of irrelevant variance that concerns us appears in segment 3. This is test material that applicants respond to consistently, but that has nothing to do with job performance. This is the material that favors the test-wise, the fortunate, or the person who is in tune with the test writers. It allows applicants to get on the top of the hiring list for reasons irrelevant to the job.

TABLE 9
Independent Components of Test Variance

1. Relevant variance common to test and criterion measure

2. Irrelevant variance common to test and criterion measure

    - Misconceptions shared by test and criterion
    - Shared biases
    - Criterion contamination
    - Test-wiseness
    - Abilities that do not pertain to the job

3. Irrelevant variance unique to the test

    - Test-wiseness
        - Specific determiners
        - Test practice
        - Problem solving set
    - Method variance
        - Test-item formats
        - Response sets (such as social desirability)
        - Cultural Bias

4. Relevant test variance not included in criterion measure

5. Relevant part of criterion measure not predicted by test

6. Relevant variance not measured by test or criterion measure

7. Irrelevant part of criterion measure independent of test

    - Poor definition of job performance
    - Rater Bias

8. Random error

Segment 4 contains any test factors related to job performance but not to the criterion measure. When we get a low validity coefficient, we claim segment 4 is large. We say, "The test is really a better measure of the criterion than our criterion measure is." And we frequently believe it, but nobody else does. Well, you can see it right here in the venn diagram. As an example, the test may inadvertently measure reading comprehension, which turns out to be important to job performance, but which we did not include in the criterion measure.

Perhaps the test measures the criterion better than the criterion measure does. But a good criterion measure likely measures the "real" criterion better than the test does. Segment 5 represents the part of job performance that is measured by the criterion measure and not by the test. You design your test to emphasize 1 and 4. You design your validity study to emphasize 1 and 5. If you do both well, 1 will be large and 4 and 5 will be small.

Segment 6 is the part of quality of job performance that is measured by neither the test nor the criterion. You can never measure how big this is. How well you get at the real criterion is determined only by judgement.

You could have a large overlap between test and criterion measure and still have a large segment 6: a large part of the criterion that is not measured. For example, you could identify 12 job elements in a job analysis and decide to measure only one of them. You could measure that one perfectly and still not measure much of job performance. Specifically, of all the things a Secretary does, you could test for typing speed and validate against typing performance. A validity coefficient of 1.0 would not assure good prediction of the job performance described in the job analysis.

Segment 7 represents the unique part of the criterion measure, the part that is associated with neither the test nor job performance. Segments 2 and 7 together constitute the most frequent flaw in validity studies, the failure of the criterion measure to represent the real criterion. This occurrence attenuates the validity coefficient. This attenuation can not be corrected for by the statistical correction for attenuation. That formula considers only the attenuation due to random error.

Segments 2 and 3 include the irrelevant test variance that we want to reduce. These are the variables that bias our test results. Be aware that when we reduce these components we lower

reliability, because we reduce total variance without reducing random error. At the same time we increase validity, because the relevant variance is now a larger proportion of total variance.

The formulas at the bottom of Table 8 show this phenomenon. The reliability coefficient, r(xx), will increase if you add to the variance of segments 1, 2, 3, or 4. The diagram illustrates the same concept. If segments 1, 2, 3, and 4 are increased while random error stays the same, the test will have proportionately less error and will appear more reliable. Only random error adversely affects reliability coefficients.

We have often been told that we should keep our reliability as high as possible. I'm telling you that is not necessarily so. When the reliability is the result of irrelevant variance it is of no use. It is worse than of no use. It makes our tests unfair. I would rather the non-relevant variance be error variance and lower the reliability coefficient, than to have variance that favors who knows whom. Whether the variance favors Shakespeare buffs, people who have taken introductory psychology, or truck drivers, if it is not related to job performance, it should not be in the test.

Selecting items using item analysis against total score can contribute to an unwanted reliability. If total score contains a good share of irrelevant variance, item analysis will identify the items consistent with the irrelevant variance.

The validity formula, r(xy), shows that validity is the common variance divided by the product of the square roots of the total variances. If the total variance of either the test or the criterion measure goes up, without an increase in the common elements of segments 1 or 2, the validity coefficient goes down. What's more, the validity coefficient will look better if you increase the shared irrelevant variance in segment 2. In the first two papers we were talking about increasing validity by decreasing component 3, irrelevant test variance.

You can also increase the validity *coefficient* by decreasing component 4: that is by removing *relevant* material from the test that is not included in the criterion measure.

What happens in a meta analysis of validity studies? It is likely that the variance common to a wide variety of validity studies in a wide variety of settings is irrelevant variance of the kinds we have been talking about. This implies a caution concerning validity generalization. The validity coefficient being generalized may contain a large dose of shared irrelevant variance.

We must be judicious when we use validity coefficients to demonstrate that are tests are valid. We may be fooling ourselves consistently. We may have some blind spots or misconceptions that apply equally well to the test and to the criterion measure. Our tests include method variance, test-wiseness, and cultural bias, which increase the reliability of our tests at the expense of job-relatedness.

CONTRIBUTORS

Brenda Morefield has 14 years in personnel in Washington State with the Departments of Personnel and Wildlife. She has worked in classification audits, recruitment, and test development in the central personnel agency. She is currently the personnel officer for wildlife.

Chuck Schultz worked five years in psychological research, seven in college teaching. He has been with the Washington State Department of Personnel for 17 years, first as director of validity studies and currently as test development manager. He has a Ph.D. in psychology from the University of Washington, including specialities in research design and testing.

Christina L. Valadez has worked in personnel and affirmative action for over 10 years with two Washington State agencies. With the department of personnel, she has worked in affirmative action, classification, test development, and rule interpretation. She is currently manager of the Recruitment Planning/Exam Administration Unit.