Technical Affairs

By Mike Aamodt, Associate Editor

A Test! A Test! My Kingdom for a Valid Test!

I thought it was going to be an easy task. Two clients wanted to do some employment testing and asked me to find vendors for them who had "good" tests. Previously, both clients had their supervisors create their own tests, and we advised them that this was not a good idea. While most supervisors understand the jobs they are responsible for supervising, they often do not have the technical expertise to develop legally defensible tests. Instead, I recommended that they should purchase "professionally developed" tests. The clients' needs weren't that complicated; they needed tests measuring mechanical ability, SQL, Java, Excel, and one of those strange computer languages that to this day I don't understand.

My plan was to contact vendors – hoping to get a "one-stop shopping" vendor – and ask for copies of their validation studies so that I could make a recommendation regarding the vendor(s) that would best meet the clients' needs. Ten vendors later, I began to realize that finding vendors who actually had information on their tests was going to be difficult. They all claimed that their tests were "valid," a few had reliability information, but none had any real evidence of validity other than vague statements about how their tests were content valid. Three of the interesting responses I received from major test vendors were:

- "We don't have any validation reports but our tests are in accordance with the Uniform Guidelines on Employee Selection Procedures (UGESP) section 14(C) of the Technical Standard Section; falling under the Classic Content Validation."
- "We do not have that information at this time. These assessments are taken from our practice test and we feel that they are very good."
- "I'm contacting you on behalf of your request for technical information for the SQL test we have online. Unfortunately, the publisher is getting ready to pull that test from the catalog. It is an older version and they are looking to add additional tests in the future. We currently do not have the Technical Manual online and there are no plans to get it since it will be removed."

A fourth vendor sent me the validity template that it uses for all its tests, describing the process for determining the reliability and content validity of a test, but not actually providing any reliability or validity information for the test itself! One vendor actually told me not to worry because, "Our test is EEOC approved."

As my frustration level increased, I took my wife's advice (always a smart thing) to take a step back and think about what information I actually needed rather than wanted. That is, what information *must* a vendor provide for a client to feel comfortable in purchasing a test? After all, because no test is valid across all jobs and all situations, what type of validity evidence or other psychometric information would one expect from a vendor? Here are the results of my musing.

Prior to using a test to select employees, there are two steps that must be taken: (1) determine if the construct measured by the test is appropriate for the job in question, and (2) determine if the test is a reasonable measure of that construct. It is the organization's responsibility to gather the information in Step 1 and the vendor's responsibility to provide the information needed for Step 2.

STEP 1: DETERMINING JOB RELATEDNESS

For a test to be useful for an organization, it must measure one or more competencies that are required to properly perform a given job. The term "properly perform" is a rather general term that includes such behaviors as properly performing tasks, engaging in safe work behaviors, arriving to work on-time and not leaving early, engaging in organizational citizenship behaviors, not engaging in counterproductive work behaviors (e.g., theft, harassment, sabotage), and not prematurely leaving the organization.

The process of determining job relatedness usually begins by conducting a job analysis to determine the tasks performed, the conditions under which the tasks are performed, and the competencies (knowledge, skills, abilities, and other personal characteristics) needed to perform the tasks under the identified conditions. Such a process helps establish the content validity of a competency to be measured, but does not establish the content validity of a given test. That is, a job analysis might indicate that basic math skills are needed for a job (i.e., a content-valid competency), but *Bob's Test of Math* might be so bad that it is not a valid *measure* of the competency. Likewise, the *Do You Excel in Excel?* test might be a valid measure of a competency (Excel knowledge), but the competency might not be needed to perform the job of a tow-truck driver.

Thus, it is important to distinguish the potential validity of a competency from the potential validity of a measure of that construct. As mentioned previously, it is the organization's responsibility to establish the validity of the competency for a given job and the test vendor's responsibility to establish the validity of the test as a measure of that competency.

(continued on next page)

Once a list of relevant competencies has been established, the next step is to find sound measures (tests) of these competencies. Potential measures might include interview questions, training and experience ratings, work samples, or paper-and-pencil tests.

STEP 2: DETERMINING IF THE TEST IS A REASONABLE MEASURE OF A COMPETENCY

Once an organization determines the job-relatedness of a competency, it has two choices: it can create its own measure of the competency, or it can purchase an existing measure. Creating a reliable, valid, and fair measure of a competency is difficult, time consuming, frustrating, costly, and just about any other negative adjective you can conjure up. Think of the frustration that accompanies building or remodeling a home and you will have the appropriate picture. Thus, creating a test internally should only be done if you have the professional resources to do so. In this case, professional resources include, but are not limited to, the technical expertise necessary to develop, implement, evaluate, and update a measure of a construct.

If you are going to purchase the test from a vendor, you should ask the following four questions and expect the vendor to have written documents (i.e., validity studies, technical reports, test manuals) to answer these questions. If you can't find a vendor that has the necessary information, you are better served not testing than to use a poorly documented test. Instead, you may be best served by having an external consultant develop, implement, evaluate, and update a test tailored to the job(s) in question that will be valid for years to come.

1. Is there evidence that the test is tapping what it purports to measure?

This is a question that addresses both the construct and content validity of the test. Let's use a personality inventory as an example to discuss construct validity. If a vendor has a personality inventory measuring the trait of conscientiousness, it is important that the vendor demonstrate that scores on it's inventory correlate highly with scores on other measures of conscientiousness and not as highly with tests of similar, yet different constructs (e.g., integrity, motivation). When a test correlates highly with other tests of the same construct, it is said to have *convergent validity;* and when it correlates less highly with similar constructs, it is said to have *divergent validity.*

Whereas the *construct* validity of a test can be rather straightforward, the content validity of a commercial test can be complicated. If a test is designed for a particular occupation - for example, a math test for law enforcement personnel – the content validity of the test is established by demonstrating that the types of math (e.g., addition, subtraction), types of mathematical measurements (e.g., decimals, fractions, whole numbers), and the object of the math (e.g., time, speed, height) are in fact necessary to properly perform the job in question. This is simply a matter of conducting job analyses and identifying the math needs of the job.

Most tests, however, are not designed for a particular occupation and thus establishing content validity becomes more complicated. Let's consider a test of Excel knowledge as an example. Excel is used in many occupations ranging from clerical to human resources to accounting positions. The knowledge level needed to use Excel for a compensation analyst is much greater than the typical clerical job. Thus, a test might actually tap Excel knowledge but might be too difficult for one job and not difficult enough for another job. As a result, vendors should identify the difficulty level of a test (e.g., beginning, intermediate, advanced) and the employer would then determine the test level most appropriate for the job in question.

Another issue that goes to the validity of the test is the passing score. That is, how high does an applicant need to score in order to "pass" the test? Although some vendors provide recommended passing scores with their tests, the passing score should probably be individually set by each organization unless the vendor has a tremendous amount of research to support its recommendation. Perhaps setting passing scores could be a topic for a future ACN column.

2. Is there evidence that the test scores are reliable?

If an applicant took the test a second time (test-retest reliability) or took a different version of the test (alternate forms), is there evidence that the applicant would receive similar scores? If two people scored the test, is there evidence that the applicant would receive similar scores (interrater reliability)? To answer these questions, it is essential that the vendor provide reliability coefficients for the test as a whole as well as any subscales.

If there are alternate forms of the test, the vendor must demonstrate that scores on the two forms are not only highly correlated (parallel forms), but also have similar means and standard deviations (equivalent forms). If reliability information is not available, or if the reliability coefficients are not high enough (in general, reliability coefficients of .70 or higher are considered acceptable), the test should not be purchased.

3. Are there sex or race differences in test scores?

It is essential that vendors provide normative information on the test scores and that this information be broken down by sex and by race (and any other protected classes of interest). It is a fact of life that many types of tests (e.g., cognitive ability, physical ability) are going to have adverse impact, so the proper way to evaluate normative information is not so much whether the test will have sex or race differences, but rather, how these differences compare to similar measures. For example, a meta-analysis by Roth, BeVier, Bobko, Switzer, and Tyler (2001) indicated that on the typical cognitive ability test, whites will score 1.10 standard deviations higher than African Americans and 0.72 standard deviations higher than Hispanics/Latinos. So, if the cognitive ability test you are considering has a white-black difference of 0.80 standard deviations, you might be more likely to adopt it than a test

(continued on next page)

with a difference of 1.20 standard deviations, all other things being equal.

As a contrast to cognitive ability tests, integrity tests show white-black differences of only 0.07 standard deviations and white-Hispanic/Latino differences of -0.05 standard deviations (Ones & Viswesvaran, 1998). As with reliability information, if a vendor does not provide norms broken down by sex and race, you probably don't want to purchase the test. As an aside, it is important that the vendor provide the source of the people used to compute the norms because race and sex differences are greatest in the general population, smaller in job applicants, and smallest in incumbents who have already been screened on some measure.

4. Is there evidence that the test will predict performance in the job in question?

If a test is designed to predict performance in a particular occupation (e.g., law enforcement), the vendor should be able to provide validity studies demonstrating that the test predicts performance in the law enforcement training academy or on some on-the-job measure, such as supervisor ratings, commendations, or discipline problems. Such studies are called criterion validity studies. This may also be a topic for a future ACN column.

For tests that can be used for a variety of occupations, it is difficult for a vendor to demonstrate criterion validity. That is, a vendor can demonstrate that scores on the Excel knowledge test were significantly correlated with performance ratings for accounting clerks, but that does not mean that the same test will significantly predict performance in other jobs such as a compensation analyst or a secretary.

Obviously, the more research a vendor has, the more comfortable the test user will feel. With that said, however, it is essential to remember that no test is valid across all jobs and that criterion validity is established by occupation, and depending on who you talk (argue) with, perhaps by individual location.

In summary, the proper use of a test is the responsibility of both the user and the vendor. It is the vendor's job to ensure that the test itself is psychometrically sound and it is the user's job to ensure that the test measures a relevant competency for the job in question. If a vendor cannot adequately answer the four questions previously discussed, that vendor should not be used.

References

Ones, D. S., & Viswesvaren, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant data sets. *Journal of Applied Psychology,* 83(1), 35-42.

Roth, P. L., BeVier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology,* 54(2), 297-330.

HR Humor

An efficiency expert concluded his conference presentation with a note of caution. “These methods will work, but you should probably not try them at home.”

“Why not?” asked a human resource director from the audience.

“I watched my wife's routine at breakfast for years,” the expert explained. “She made lots of trips between the refrigerator, stove, table, and cabinets, often carrying a single item at a time. One day I asked her, ‘Honey, why don't you try carrying several things at once?’”

“Did it save time?” the person in the audience asked.

“It sure did,” replied the expert. “It used to take her 20 minutes to make breakfast. Now I do it in seven!”