CentACS White Paper on:

Validity and Reliability of the WorkPlace Big Five ProFile (long form)

Prepared by:
Pierce J. Howard, Ph.D.
Director of Research
May 18, 2007

What does it mean to call a test valid? Reliable? In some ways, it is like calling a person nice: nice could mean any of many things, depending on who is using the word, and about whom s/he is talking! Nice is neatly groomed, but also friendly, or usually smiling. Nice could also mean someone who makes good grades, or who is well-rounded. One who is studious, or helpful, or a hard worker, or just conversational, or even just quiet. Volunteers are nice. Well, you get the idea.

In fact, to call a test valid is something like calling a person nice. Just as nice does not mean any one specific attribute, but rather an accumulation of attributes, so validity is not a single attribute, but rather an accumulation. To quote the *Standards for Educational and Psychological Testing* (1999, p. 11):

> Validity…is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose.

So, let us take a look at the various kind of evidence in support of the validity of the WorkPlace.

1. **Item content**. The language of test items should reflect the context to which results will be applied. This applies in two ways:
    a. **Work Context.** The WorkPlace results are interpreted in terms of people's behavior at work. Accordingly, the language for test items is workplace language: "…work in solitude", "enjoys making calls on others", "…with work associates", "…wears many different 'hats' at work", "…imagines new business concepts", and so forth.
    b. **Global/Cultural Context.** The WorkPlace has been translated into several other languages, including "Asian English." In each case, we forward translated into the target language, then had a different linguist back-translate into English. We "juried" the resulting back translation to insure that the content matched the construct being measured by each item. Discrepancies were resolved by creating new items in the target language that were faithful to the construct in question. Hence, each translation uses language that is both natural for the target culture(s) and reflective of the construct to be measured.
2. **Item Format**. Studies have shown that items worded in the third person (as in "Is a talker" or "Interrupts others") elicits a wider range of responses than items beginning with the personal pronoun (as in "She/he is a talker" or "I am a talker"). For this reason, we use the third person format through the SchoolPlace and the WorkPlace.
3. **Response Options**. Studies have shown that the use of all positive anchors (as in 1 through 5) with a Likert-type scale (e.g., Strongly Disagree through Strongly Disagree) fails to elicit as wide a range of responses as do a mix of negative and positive anchors (as in -2 through +2). For this reason, we use the -2 through +2 format in both the WorkPlace and the SchoolPlace.

4. **Bandwidth**. In testing, "bandwidth" refers to the scope of a particular measure. "Overall IQ" has broad bandwidth, while "3-dimensional spatial rotation ability" has narrow bandwidth. In the WorkPlace, "Extraversion" has broader bandwidth, while "Sociability" has narrower bandwidth. Particular when using an assessment for selection and coaching, it is important to have both kinds of traits, as some work issues are better explained by broad bandwidth traits (such as achievement being explained by C), while other, more specific work issues are better explained by narrow bandwidth traits (such as sales achievement being explained by N4-, E2+, A3-, and C3+, in addition to C). For this reason, both the WorkPlace and the SchoolPlace have broader bandwidth measures (the five supertraits) and narrower bandwidth measures (the 24 subtraits).
5. **Internal Consistency.** Each trait is measured by a set of items. For example, C2: Organization is measured by:
    a. Gets organized before beginning a task
    b. Is neat and tidy
    c. Keeps everything in its place
    d. Organizes for work effectively

    To the degree that respondents tend to answer these four questions in the same direction (e.g., in this case, agreeing with all four or disagreeing with all four), then C2 is said to be internally consistent. This is measured by **coefficient alpha**, aka Cronbach's alpha. Alphas should never fall below the value of .5 for subtraits, .7 for supertraits. Low alphas indicate that the items are probably not measuring the same thing. On the other hand, if alphas are too high, e.g., over .9, it suggests that the items are too similar, and lack robustness, failing to capture the subtlety and complexity of human behavior. On the SchoolPlace and WorkPlace, all subtraits are between .5 and .8, and all supertraits are around .8. This is comparable to the best of assessments available, such as the NEO tests. Interestingly, traits that are more external and easier to observe tend to have higher alphas than traits that are more internal and trickier to observe. An example would be C2: Organization (= .81) versus N3: Interpretation (= .59)
6. **Cross-Validation**. A standard analysis to establish validity of a test is to take the norm group, divide it in half (randomly), and apply your scoring alrorthyms to each group separately. The results should be the same. For both the WorkPlace and the SchoolPlace, we performed a cross-validation study, each with excellent results.
7. **Norms.** Tests should be normed on the same kind of population that will use its results. For example, one test we know was normed on a small group (n=150) of managers, mostly male. The use of that test for sales (or other job categories) and females is thus not allowed. The WorkPlace is normed on adults from age 18 up who are working full time. The SchoolPlace is normed on full time students from age 12 to 22. Not only must the norm group be similar to the people who will be taking the test, but the norm group must reflect the diversity of that group. So, for example, in our tests, we use the most recent US census as an aid in creating a well-balanced norm group. In our most recent study (for version 4.0), we had over 10,000 subjects in our initial analysis, but the final norm group was reduced to under 2,000, so that the number of representatives for each sex, race, age, industry, and job category (or in the case of the SchoolPlace, favorite subject/s or college major, and private/public school) would be reflective of their normal distribution in the workforce. The details are available in the Professional Manual.

8. **Compliance with ADA**. The original item list for the SchoolPlace was 800+ statements. We submitted them to a labor/employment attorney and asked to redline any item that we could not ask prior to hiring. All red-lined items were eliminated from further consideration, as well as a couple of subtraits. For example, we have no items which inquire about political, religious, or social beliefs.
9. **Assessing for Disorders.** The court cases that have lost, or rather, the tests that have been found to be discriminatory/unallowable for use in hiring, have been those whose results provide information about psychological disorders. Our tests neither measure for disorders, nor report them. The courts have said that tests measuring normal personality, so long as they have been proven through validity studies to be job relevant, are permissible.
10. **Court Record**. Of course, one of the best indicators of validity is a) the absence of court challenges, and/or b) the successful defense of court challenges. At this point, neither the WorkPlace nor the SchoolPlace have been challenged. In fact, to our knowledge, no Big Five test has "gone to court."
11. **Social Desirability**. Many tests have "lie scales" (also called validity scales or social desirability scales) to determine whether or not the respondent is being truthful. Research indicates that such scales do not work (references available upon request). We follow the suggestion of Costa and McCrae, who suggest that the use of raters can control for socially desirable responding in the case of high risk assessments, and that careful attention to instructions given to respondents can minimize socially desirable responding. In some groups (e.g., sales), the tendency to self-enhance (sales folks tend to do this on A and C) is so widespread that it makes no difference!
12. **Normative vs. Ipsative**. Ipsative scales should not be used for selection, as the results do not allow you to accurately determine how the individual compares to the population at large. From an ipsative scale (e.g., the DISC, AVA, or MBTI), all you know is how an individual is "measured against himself," while in normative scales you know how the individual is "measured against others." In an ipsative scale (e.g., all forced choice scales are ipsative—"Which is more like you, a or b (or c or d or…)?"), you count up how many times you answer in a particular direction. The problem with this approach is that while you may choose a over b, you don't know whether this is a huge or a trifling preference. In a normative scale ("To what degree do you like a?"), you are only evaluating one issue at a time, and not confounding it with other issues.
13. **Comprehensive vs. Partial**. Partial scales do leave out some aspect of normal personality. For example, the MBTI is partial, and leaves out Need for Stability. The Big Five is regarded as comprehensive, especially when subtraits are included. The more subtraits, the more comprehensive.
14. **Empirical vs. Theoretical**. Tests based on a theory of personality can only be used with that theory of personality, so you have to buy into that theory in order to use the test. This is true for the MBTI, AVA, DISC, and so forth. Empirical tests generally try to measure the basic structure of personality, the basic "building blocks." The results of an empirical test, in fact, can be used with almost any theory. The Big Five is empirical.
15. **Predictive Power**. Across all studies we have done with the WorkPlace and the SchoolPlace, we have found that individual traits typically correlate from .15 to .30 with performance criteria. When you combine traits into a multiple regression (e.g., using N4, E1, E2, E3, A3, and C4) to predict the criterion (e.g., sales volume), we typically get coefficients right around .40. As traits are only a

part of the total person, you must combine trait predictors with other predictors relevant to the performance criterion being measured. Other predictors include mental ability (e.g., numerical analysis), physical ability (e.g., hand-eye coordination), background checks (e.g., credit, police, academic), and experience factors (e.g., military service, previous work experience, hobbies). By combining these other factors (none of which alone associate more than .50 with performance), you can get a multiple regression around .90.

16. **Results from Selection Applications**. Across a variety of industries (e.g., banking, entertainment, government, manufacturing, utility, transportation, and so forth), these are the kinds of results we get from entering WorkPlace results into the selection equation:
    a. Reductions in employee turnover
    b. Increase in performance levels
    c. Improved information for the acquiring manager to use in coaching, team building, and so forth
    d. Excellent discriminant validity (in one recent study for a state government department, six jobs within the department were analyzed, with each job yielding its own unique profile associated with high performance in that job)

Finally, let's have a word about reliability. To begin with, this is a much simpler concept than validity. A test is said to be reliable if, when you repeat it, you get essentially the same thing that you got the first time. In other words, it is consistent. Reliability is typically measured by test/retest studies. Give the test today, and again tomorrow, and compare the results. These studies are typically of two types:

a. *Short term test/retest.* Here you administer the test on one day, then again to the same people anywhere from one month to two or three months later. A well-constructed test should yield short term test/retest reliability around .90. The WorkPlace and the SchoolPlace both achieve this level of short term test/retest reliability (details available in the Professional Manuals)

b. *Long term test/retest.* Here you administer the test on one day, then again in one to three years. A well-constructed test should yield long term test/retest reliability around .70. Both the WorkPlace and the SchoolPlace achieve this level. Reliability declines over time, because experience and context can have a small influence on how one responds to behavioral questions. The Big Five dimensions of N and A are more susceptible to environmental influence that the other three dimensions, as one's levels of stress (N) and position within a hierarchy (A) influence response patterns. It is not the genetic part of trait structure, but rather the environmentally-influenced part that changes.

Reliability is also assessed in two other ways: split/half and coefficient alpha. Split/half methods are used more commonly with ability tests, where there are right and wrong answers. We have not done split/half studies with our Big Five tests. Coefficient alpha (described above in paragraph #5 in the Validity section) is a versatile statistic, as it is accepted as an indicator both of validity and reliability. Good alphas support validity, in the sense that they suggest that all items are measuring the same thing, while they also support reliability, in the sense that they suggest that the respondents are being consistent in their answering.