IJT Test Validity and Adaptations Section

The International Journal of Testing (IJT) publishes research on testing and assessment in psychology, education, counseling and organizational behaviour. This is the stated objective of IJT, but an examination of the papers published in the journal indicates that few of the papers address the validity of tests and the appropriate adaptation of tests developed in one culture, country, or language to another. There are a number of reasons that these papers are not typically published in the journal including the fact that they are often narrow applications of tests in one context or situation that don't make a substantial contribution to the literature. The editorial staff of the journal believes these articles have value in themselves and as part of an archive on test use, validity, and adaptability across cultures and contexts.

Hence, we are experimenting with a new format and section of the journal in which these papers might be published. We envision relatively short papers (no more than 4 manuscript pages or 2000 words). These papers would describe very briefly the test in question, the context in which it was used, the number and nature of the participants in the research, the type of validation design used, the results and implications. To provide some guidance on the type of information that such papers should provide, we provide the following checklists of information that should be included in various types of papers that would fit in this section of the journal. We also envision an abbreviated review process designed to assess whether the critical information described below is included in each paper.

We invite submissions that fit these descriptions and will be happy to answer questions about the appropriateness of potential papers.

Information to be Included in Papers for this Section of IJT

Criterion-related validation studies should describe:

- The context or setting within which the study was conducted and how and when the data were collected.

- The name and description of the test being evaluated and the purpose of the study. This should include (whenever possible) all test items or, at a minimum, example items that are descriptive of the test as a whole. The full set of items is preferable but not necessary for this section.

- The criteria or outcomes examined in the validation study.

- Any control variables or additional measures that contribute to an understanding of the primary predictor or test of interest.

- Reliability information for both the predictor and criteria used in the study.

- The number of participants and a description of the sample being studied.

- A description of how the data were summarized and analyzed including correlational analyses, item analyses, regression, or factor analyses.

- Tables of descriptive data (means, standard deviations, correlations) and tables or brief descriptions of the results of all analyses conducted.

- Brief statements of the implications of the study for the use of the test.

Construct validation studies should include:

- All of the descriptive information about the study that is mentioned above should be provided (sample, setting, etc.).

- The context or setting within which the study was conducted and how the data were collected.

- The nature and description of the construct that is measured and why it is important to measure this construct.

- All test items or sample test items that provide the reader with a sense of the content of the measure. Whenever possible, the full set of items is preferable but not necessary for this section.

- Information collected to support conclusions about the nature of the construct. This might include a series of hypotheses about the degree to which alternate measures of the same construct ought to be related to the measure being examined. It would also include a list of the measures used to evaluate these hypotheses and why those measures are appropriate for establishing construct validity.

- Information about the reliability of all measures examined in the study.

- A description of the analyses conducted to support the construct validity of the measure. These would most likely include item analyses, correlational analyses, and exploratory or confirmatory factor analyses.

- Descriptive data (means, standard deviations and correlations) should be included in all papers. Multivariate analyses, including exploratory and confirmatory factor analyses, targeted to the evaluation of the hypotheses identified above should also be summarized. Item response theory (IRT) studies may also be appropriate and should be accompanied by an assessment of dimensionality and model-data fit.

- A short description of the implications for test use, perhaps as an alternative to existing measures.

Test Adaptations: Test adaptation occurs when a test created in a "source" culture is imported to a "target" culture in such a way as to change the psychological meaning of the test as little as possible. It may or may not include translation of the test linguistically.

Studies focused on test adaptation should include:

- A description of the construct being measured, its purpose, and the population on which it is used in the source culture. It should also describe the nature of the population for whom the test is being adapted.

- A description of the context in which the test is being used and the purpose of testing as well as how the data were collected.

- If translated, a description of the methods used to derive the translation should also be included. Test or sample test items should be reported to provide the reader with a sense of the translated content of the measure. Whenever possible, the full set of items is preferable but not necessary for this section.

- Whenever possible, the source test and target test should be compared. This includes an examination of the internal structure of these tests and may include reliability analyses, item analyses, and factor analytic or IRT tests of equivalence.

- Correlations with external variables should be reported for both the source and target cultures. This includes correlations with demographic variables (e.g., gender, ethnic status, educational status) that may indicate differential consequences of test use for some groups.

- A brief description of the implications for test use in the target culture. This should describe how the adapted test can be used effectively in the target culture and the equivalence of test scores when compared with the source culture?

Example of a Criterion-Related Validity Report

Author: Validity IJT

Context or setting within which the test is evaluated: Data on the test were collected at a large undergraduate institution in the United States for which students applied for admission.

Name and description of the test. The test (containing seven subscales) is a biographical data measure intended to predict success as an undergraduate student. These scales contain items that assess hobbies, interests or past experiences that might indicate an applicant will be successful as an undergraduate student. The names of these scales and descriptions of the item content follows. Responses to all items were made on continuous five point response scales.

Knowledge and mastery of general principles
Gaining knowledge and mastering facts, ideas and theories and how they interrelate, and the relevant contexts in which knowledge is developed and applied. Grades or GPA can indicate, but do not guarantee, success on this dimension.
Sample Item: I usually got good grades on all homework or assignments.

Continuous learning, and intellectual interest and curiosity
Being intellectually curious and interested in continuous learning. Actively seeking new ideas
Sample Item: When presented with interesting ideas in class, I always sought out additional information on the idea/concept.

Leadership
Demonstrating skills in a group, such as motivating others, coordinating groups and tasks, serving as a representative for the group, or otherwise performing a managing role in a group.
Sample Item: When starting a new project, I was usually expected to provide direction.

Social responsibility
Being responsible to society and the community and demonstrating good citizenship. Being actively involved in the events in one's surrounding community, which can be at the neighborhood, town/city, state, national, or college/university level. Activities may include volunteer work for the community, attending city council meetings, and voting.
Sample Item: I often participated in special fund raising events in my school.

Adaptability
Adapting to a changing environment (at school or home), dealing well with gradual or sudden and expected or unexpected changes. Being effective in planning one's everyday activities and dealing with novel problems and challenges in life.
Sample Item: I was usually able to get my work done even when an outside event made it difficult.

Perseverance
Committing oneself to goals and priorities set, regardless of the difficulties that stand in the way. Goals range from long-term goals (e.g., graduating from college) to short-term goals (e.g., showing up for class every day even when the class isn't interesting).
Sample Item: Even when I faced a problem that seemed too difficult to solve, I worked until I was successful.

Ethics
Having a well-developed set of values, and behaving in ways consistent with those values. In

everyday life, this probably means being honest, not cheating (on exams or in committed relationships), and having respect for others.
Sample Item: Even when it would have been easy to copy someone else's work, I always refused to do so.

To compare the efficacy of these predictors with more traditional measures, we also collected data on high school grade point average (HSGPA) and standardized test scores (SAT).

Criteria and Outcomes Measured.

Outcomes against which the predictors were evaluated included self-ratings of performance as students (1-7 behaviorally anchored rating scales), grade point average (measured from 0.0 to 4.0) at the end of the first year of school, and retention (0 or 1, with 0 indicating the student had withdrawn).

Control or additional measures of interest.

Gender (Male = 1, female =0) and ethnic status (Majority=1, other = 0) were included to assess the potential for gender or ethnic bias.

Reliability.

Reliability was assessed using coefficient alpha when appropriate and is reported in the table below.

Sample.

Participants included 1,184 applicants to major US undergraduate universities. The data were collected by administering surveys at two separate time points to participants. Additional criterion data were also collected from teachers.

Analysis.

The analysis included computation of descriptive statistics (means, standard deviations) and correlations between predicators and outcome variables. In addition, regression analyses of the three outcome variables on the set of nine predictors (biographical data scales and SAT and HSGPA) were conducted. Correlations between gender and ethnic status with each outcome and predictor were also computed.

Results.
The table that follows contains means, standard deviations, reliabilities, and correlations between predictors and the three outcomes.

Variable	Mean	SD	R with GPA	R with self-rating	R with retention
Knowledge	2.73	2.01	.22	.47	.12
Cont. Lrng.	2.84	2.50	.06	.40	.01
Leadership	3.24	3.00	.14	.41	.02
Soc. Resp.	2.42	2.53	.08	.39	.13
Adaptability	3.21	2.86	.21	.24	.08
Perseverance	1.47	2.84	.16	.45	.15
Integrity	3.62	3.15	.14	.35	-.01
HSGPA	3.52	.62	.36	.20	.19
SAT	578	45	.33	.18	.16
Gender	.55	.50	-.12	.03	-.09

Ethnicity	.82	.38	-.18	.08	.17
HSGPA	3.02	.69			
Self-Rating	4.88	.71			
Retention	.92	.27			

Regression analysis of GPA on the nine predictors yielded an R of .62. As reported in the table, HSGPA and SAT were most predictive of this outcome. A similar regression analysis of self-ratings of performance on the predictors yielded an R of .55; in this case, the biodata scales were the strongest predictor of this outcome. Finally, a regression analysis of retention on the predictors yielded an R of 29. Only a few predictors (adaptability, knowledge, SAT and HSGPA) contributed significantly to the prediction of this outcome.

Gender was correlated significantly with SAT (Males did better) and with college GPA (women did better). Ethnic status was particularly highly correlated with SAT, HSGPA, and college GPA (in all cases the majority group did better).

Implications

HSGPA and SAT are superior predictors of college GPA but several biodata scales were also meaningfully related (above .15) with college GPA.

Biodata scales are most predictive of self-ratings of performance though both HSGPA and SAT also contributed to the Multiple R for this outcome.

Retention was significantly predicted by four predictors, though the multiple R is modest in magnitude.

Gender and Ethnic status were relatively uncorrelated with biodata predictors but SAT and HSGPA were highly correlated with these demographic characteristics. If used to predict college GPA and make college admissions decisions, it will be important to assess differential prediction for members of different ethnic groups. The same can be said for gender subgroups. By contrast, there would be little concern for differential prediction if one were to use biodata scales to make admissions decisions.