Taras, V., & Kline, T. J. B. Scale validation via quantifying item validity using the Dm index. Psychological Reports, 2010, 107, 2, 535-546.

QUANTIFYING CONSTRUCT VALIDITY USING THE D_m INDEX

Abstract

While most validity indices are based on total test scores, this paper describes a method for quantifying the construct validity. The approach is based on the item selection technique originally described by Piazza (1980). However, Piazza's P^2 index suffers from some substantial limitations. The D_h coefficient provides an alternative that can be used for item selection and also to provide a validity index for a set of items. An example of how to use the technique is provided. This method may be especially useful when the sample of items and/or persons is small, rendering more traditional approaches such as factor analysis or item response theory inappropriate.

QUANTIFYING CONTENT VALIDITY USING THE D_m INDEX

The term “validity” lies at the heart of all test item construction and use. One of the most seminal discussions of validity is the paper by Cronbach and Meehl (1955) written over 50 years ago. In this paper they argued that the validity of a construct was based on a nomological network of relationships around a given construct. While reliability represents the expected consistency of the test results if the test was administered multiple times or multiple versions of the tests were used (Hogan et al., 2000), validity is a measure of correspondence between the test and the construct that the test is intended to measure.

There are various methods for assessing aspects of validity. A single validation method is not likely to be effective and an instrument validation should be based on multiple validation criteria (Kline, 2005). Three major types of validity are usually differentiated, although this differentiation is more superficial and based on validation methods rather than types of validity (Landy, 1986). *Construct validity,* a.k.a. content validity, represents the extent to which a measurement reflects the specific intended domain of content. *Construct validity* refers to the degree to which items reflect a single concept and have consistent relationships with theoretically important exogenous variables. *Criterion-related validity* is used to demonstrate the accuracy of hypothesized relationships between the measure and other measures (Carmines & Zeller, 1979).

To gather evidence that the inferences made based on tests scores are valid, each approach to validity assessment has been used. One characteristic of the criterion-related validity approach is that it provides quantitative values (and as a result, the capability to assess for statistical significance and report effect sizes). This has been of substantial importance when tests are used in high-stakes environments such as selection for employment, decisions regarding treatments, and educational streaming. In such situations, the criterion-related coefficient is expressed as a correlation or other measure of the relationship between, say test scores, and a criterion variable, such as job performance (e. g., Hollenbeck & Whitener, 1988; Timmreck, 1981).

Most of the criterion-related validity indexes refer to total test scores rather than specific items. While item-to-criterion assessment has been discussed as one potential way to decide to retain or discard items (e.g., Ghiselli et al., 1981), a summary index for the test, based on these item indices has not been fully described. This paper offers a method for doing so and was developed based on the item selection technique described by Piazza (1980). As will be shown Piazza's method, and the Piazza's P^2 index resulting from it, suffers from some serious limitations. An alternative method for assessing criterion validity is offered and a new D_h coefficient is introduced that summarizes a set of item criterion-related validity coefficients. This coefficient is similar in its approach to that suggested by Westen and Rosenthal in 2003 as an overall construct validity coefficient at the *test* level. Finally, an example using the D_h coefficient is illustrated.

PIAZZA'S INDEX OF PROPORTIONALITY: P^2

Although originally intended as an item selection technique that should be used in addition to factor and reliability analyses, Piazza's index of proportionality P^2 (1980) is closely related to the notion of an instrument's construct validity and has been used for in this manner (e.g., Booth et al., 1983; Roberts & Clifton, 1992). The method is based on the evaluation of the relationships between item scores and a set of theoretically relevant exogenous variables.

The set of exogenous variables used to evaluate an item's "construct validity" uses a similar procedure to that used to assess test criterion-related validity. However, the variables used in the latter often focus on outcomes different from that of the test (e.g., predicting success in college from SAT scores). The variables used for construct validation often focus on constructs similar to that of the test (e.g., the relationship between SAT scores and high school grades). In both cases, there is a "criterion" and so this term will be used throughout to refer to any exogenous variable, and is consistent with Cronbach and Meehl's (1955) reference to any variable being located within a nomological net.

In his illustration, Piazza (1980) used a five-item test to study opinions about the nature of racial economic inequalities. The items were different formulations of the basic question: "Who is to blame for the fact that blacks are not as well off as whites?" To evaluate the validity of the test items Piazza assessed the relationships between responses to each of the five test items and five exogenous criterion variables that he theorized to be alternative predictors of the answers to the questions. The variables he included were: age, education, sex, income, and number of children. The relationships between item responses and the criterion variables were assessed using Pearson's product-moment correlation coefficients. According to Piazza, in a good test, the profile of the items' relationships to the exogenous should be very similar. That is, it is expected that correlation coefficients between all of the test items with each of the exogenous variables will not differ greatly from one item to next within the same test.

To assist in identifying "good" items from "poor" items obtained correlations are plotted. The relationship patterns are then examined to weed out items that "contaminate" the test. Figure 1 is a reproduction from Piazza's (1980) original article (p. 590). Based on the analysis of the graph, Piazza concluded that items 4 and 5 (bold lines on the graph) displayed criterion relationship patterns inconsistent with the rest of the test items and, therefore, should be dropped from the test.

Insert Figure 1 about here

Piazza further provided a quantitative index of the degree of item profile consistency with the *P*^{2}. The focus in the equation he used was the extent to which each pair of items had proportional correlations across the set of theoretically relevant variables. He called this the index of proportionality *P*^{2} and is shown in Equation 1.

(1) $P^2 = (\sum_k rx_1z_k rx_2z_k)^2 / (\sum_k r^2x_1z_k \sum_k r^2x_2z_k)$, where:

k=1, …, N and z_k are the *N* criterion variables.

The *P*^{2} statistics equals +1.0 if the item scores have exactly proportional correlations with each of the exogenous variables. It equals -1.0 if the correlations are proportional but always are opposite in sign. It equals 0.0 if there is no consistent proportionality (Piazza, 1980:592).

Limitations of Piazza's Technique

The technique described by Piazza appears at first glance to be useful for item selection. However, his proportionality index has some serious flaws that render it ineffective as a validity index. First, the proportionality index is completely driven by the degree of similarity the test items display in their correlation with criterion variables. Therefore, the P^2 is a measure of item internal consistency than validity.

Second, the P^2 ignores the magnitude of the correlations. Consider the following cases in which the proportionality index yields values close to +1.0, yet it is obvious the items display no construct validity. For example, suppose all the test items do not correlate with the criterion variables, (i.e., all correlation coefficients between test item scores and criterion variables are close to zero (Figure 2). In this case, the P^2 will be close to one suggesting perfect properties of the test. Yet, from the construct validity perspective the test items are useless, as they do not relate to any exogenous variables of interest. In fact, Piazza's original example may be suspect for this very reason. Most of the observed correlations between the item scores and criterion variables were weak. If they are plotted on a graph that includes the entire range of possible correlation coefficient values ($\pm$ 1.0) the plots are much less impressive than that used in the original figure where the range was -0.2 to +0.4 (see Figure 3).

Insert Figures 2 and 3 about here

Consider another case in which each item score strongly correlates with each criterion variable and the item correlation profiles are close to identical, but the directions of the correlations are opposite to the expected effects (see Figure 4). In this case, too, the P^2 will be close to one suggesting perfect construct validity. Yet, this conclusion would most certainly not be appropriate. Because P^2 is only based on item-criterion correlation profiles, ignoring the direction of the hypothesized relations, the coefficient may be useless or even misleading.

Insert Figure 4 about here

THE D_m VALIDITY COEFFICIENT: AN ALTERNATIVE TO PIAZZA'S P^2

A basic tenet for assessing item construct validity should be that we expect item scores to be significantly correlated with theoretically relevant exogenous variables. We also expect that the observed directions of the correlations will be consistent with the hypothesized directions of the effects. This is the same argument put forth by Westen and Rosenthal (2003) with regard to total test scores. They make the case that contrast analysis should be used more frequently by researchers trying to demonstrate construct validity. Through this approach, a single coefficient can be generated. It requires the researcher to specify in advance what the expected relationships are going to be between the construct of interest and other constructs within the nomological net.

On the graph similar to that produced by Piazza (1980) a good predictive validity would be indicated by a *substantial deviation* of each value from zero in the *hypothesized direction*. Below is an equation (2) for calculating a construct validity coefficient that captures both the amount deviation from zero and the direction of the relationship between each test item and exogenous variable(s). This index is denoted as the coefficient as D_0 as it is based on the analysis of the deviations of observed correlations from zero.

(2) $D_0 = 1 - \sqrt{\frac{\sum_{1-k} (r_o)^2}{k}}$, where:

r_o is the observed correlation between the item and exogenous variable, and

k is the number of exogenous variables.

As can be seen, the coefficient is derived by analyzing the deviations of observed item-criterion correlations from zero and its magnitude depends on the strength of association between the item scores and the criterion variables. Although the coefficient addresses one of the limitations of the P^2, it has a serious flaw: a high coefficient can be obtained only when the item-criterion correlations are close to one. However, perfect correlations are extremely unlikely in social sciences and one would not hypothesize such a case. We can refine the index by including comparisons between the observed correlations against theoretically meaningful values.

Equation 2 can be modified to resolve this problem and is shown in Equation 3. (3)

$D_h = 1 - \sqrt{\frac{\sum_{1-k} (r_h - r_o)^2}{k}}$, where:

r_h is the hypothesized correlation between a test item and an exogenous variable;

r_o is the observed correlation between a test item and an exogenous variable;

k is number of exogenous variables.

The resulting coefficient is a measure of absolute deviation of observed correlations from their hypothesized values. As an aside, this modification is similar to one suggested to Westen and Rosenthal (2003) by a reviewer to their index (Smith, 2005). The index is denoted coefficient D_h because it takes into account both the direction and the magnitude of the hypothesized relation between the items and the exogenous variables. The range of D_h values is typically 0.0 - 1.0 with a D_h equal to +1.0 representing perfect criterion validity and is obtained when each observed correlation between the item score and exogenous variable perfectly matches the hypothesized value. Values close to zero indicate no criterion validity. While negative values of D_h are possible, they are extremely unlikely as they can be obtained only when very strong correlations were hypothesized and very strong correlations but with opposite signs were obtained. The strength of this approach is its link to theory. That is, the researchers must be explicit regarding not only the variables they expect the items of a test to be related to, but also the effect sizes of those relationships.

One of the major limitations of the D_h index is its subjective nature. It is left up to the researcher's discretion to choose the values of hypothesized correlations. The value of coefficient can be easily manipulated by changing the size of hypothesized relationships. The D_m index value can be increased simply by including exogenous variables with very low hypothesized correlations with the item scores into the analysis. For example, a researcher may choose a number of exogenous variables that are very weakly related to the item scores. In this case, the hypothesized correlations will be close to zero, as will the observed correlations. Therefore, the value of the D_h coefficient cannot be reliably compared across studies.

This limitation can be overcome by using a combination of D_0 and D_h techniques. Rather than measuring deviation of observed item-criterion correlations from zero or from a hypothesized value, a measure of deviation from 0.5 (or -.05) would be more appropriate. This way, the equation for computing D_m is the following.

$$Dm = 1 - \sqrt{\frac{\sum_{1-k}(0.5_{hd} - r_o)^2}{k}}$$, where:

0.5_{hd} is the constant of 0.5 with the sign representing the direction of the relationship (negative/positive effect);

r_o is the observed correlation between a test item and an exogenous variable;

k is number of exogenous variables.

A correlation of 0.5/-.05 represents a moderate degree of association between constructs and it is not unusual to observe correlations of this magnitude between variables in social sciences. Unlike with D_0 where the maximum value of the coefficient (1.0) can be obtained only when observed correlations are equal to zero (an unrealistic case in social sciences) the maximum value of D_m can be obtained when observed correlations equal 0.5/-05. The focus on deviation from the middle value of correlation range is also justified by the fact that weaker correlations probably indicate insufficient association between the measure and the criterion, while higher correlations can be indicate of insufficient discriminant validity (i.e., the variables measure the same construct). At the same time, a fixed target value of 0.5/-.05 provides standard reference points and allows for comparison of the D_m coefficient values across studies. The range of possible values of the D_m coefficient is from -0.5 to 1.0; however, obtaining negative values would be unlikely as this occurs only when extreme deviations of item-criterion correlations from 0.5/-0.5 are observed.

The D_m validity coefficient is subject to some limitations. Because it is based on Pearson's product-moment correlation coefficients between test item scores and criterion variables, it can be properly calculated only when the assumptions of correlation analysis are met. Namely, linearity, normality and homoscedasticity (Harper, 1965). However, there are a number of alternatives to Pearson's *r* for categorical variables, such as Kendall's τ, Cramer's V, or the phi coefficient.

APPLIED EXAMPLE

Let us consider an example of how D_m can be used for validity evaluation and improvement. The example is based on a simulated dataset generated for this example and intended for illustration purposes only. Suppose a researcher is developing an instrument for measuring organizational commitment. The tentative version of the instrument contains four items, each asking about a respondent's commitment to the organization. The researcher wants to make sure that the instrument measures organization commitment and no item contaminates the construct. To validate the instrument she surveys a small group of university professors. Table 1 provides the original responses to the questionnaire items provided by each of six respondents.

Insert Table 1 about here

With so few respondents the researcher cannot use factor analysis or item response theory approaches with the data. The results she obtains are very mixed and unstable and do not provide basis for any meaningful conclusion. Internal consistency as provided by Cronbach's alpha is also highly suspect given the small number of items and respondents. The researcher had known that her sample size would be too small for these analyses, so she included a set of eight additional questions (to be used as the exogenous variables in the D_m calculation) in her questionnaire (Table 2). She can now calculate the D_m index by comparing the obtained and hypothesized correlations between item scores and the exogenous variables that she hypothesized to be related to organizational commitment.

Insert Table 2 about here

The researcher recognized that the items under consideration might be related not only to organizational commitment, but also to a number of other constructs that are not a focus of her research. For example, it is easy to confuse organizational commitment with unwillingness to move to a different location. In other words, a professor may not be committed to the university at which he is currently employed and so one might think he would be willing to accept a new job offer. However, this particular professor does not want to move because his spouse is happily employed and does not want to move, so his answer to the question: “How likely would you be to consider accepting a job offer from a different university?” would be “1” on the 7-point scale. Thus, even though the question seems to relate to organizational commitment, the example illustrates why this may not be a good item for an organizational commitment test.

To detect the items that may relate to a different construct, the researcher creates a table with hypothesized directions of moderate correlations (0.5 or -.05) and their correlations with the test items on the four exogenous variables that were hypothesized to be positively related to organizational commitment. She also includes a set of hypothesized correlation directions with four other exogenous variables that were hypothesized to be related to a potentially contaminating construct (unwillingness to move). The researcher calculates the correlations between the four items on the commitment survey and each of the eight additional questions (Table 3).

Insert Table 3 about here

Figure 5 depicts the pattern of correlations. The two thick lines show the hypothesized moderate correlations (0.5 or -.05) between the test items and each of the exogenous variables (dashed for organizational commitment, solid for unwillingness to move). As can be seen, items 1, 3 and 4 (light dashed lines) resemble the profile hypothesized for organizational commitment, while item 2 (light solid line) approaches the profile hypothesized to relate to unwillingness to move. Thus, it appears that item 2 does not fit the organizational commitment construct.

Insert Figure 5 about here

The researcher calculates the D_m validity index for the 4-item instrument. With all four items, the D_m coefficient equals 0.48. If item 2 is removed, the D_m increases to 0.72 indicating a substantial improvement in construct validity. Based on the findings, it is clear that the instrument would be a cleaner measure of organizational commitment without item 2.

CONCLUSIONS

The assessment of items and how the contribute to a theoretically meaningful construct has been hampered by the need for large sample sizes and many items. In the early stages of scale development, these conditions cannot often be met – particularly with hard-to-reach populations (e.g., clinically depressed individuals, chief executive officers, airplane pilots, high level athletes, etc.). The D_m index provides a measure of construct validity akin to that of the traditional test criterion-related validity index and can be used in situations where large-sample analyses are not possible. It is very useful insofar as the test item developers must be clear about the construct validity of each item – how it fits within the nomological net of other constructs. A plot of the correlations between items and exogenous variables can assist in providing diagnostic evidence of potentially problematic items. These features should assist in ensuring the tests are constructed of items that meet the expectations of potential users.

References

Booth, A., Johnson, D., & Edwards, J. N. (1983). Measuring Marital Instability. *Journal of Marriage and the Family, 45*(2), 383-394.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment*. Thousand Oaks: Sage.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Ghiselli, E. E., Campbell, J. P., & Zedek, S. (1981). *Measurement Theory for the Behavioural Sciences*. New York: W.H. Freeman.

Harper, A. E., Jr. (1965). Down with the validity coefficient. *Journal of Vocational and Educational Guidance, 11*(3), 75-86.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational & Psychological Measurement, 60*(4), 523-531.

Hollenbeck, J. R., & Whitener, E. M. (1988). Criterion-related validation for small sample contexts: An integrated approach to synthetic validity. *Journal of Applied Psychology, 73*(3), 536-544.

Kline, T. J. B. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation*. Thousand Oaks: Sage Publications.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183-1192.

Piazza, T. (1980). The analysis of attitude items. *American Journal of Sociology, 86*(3), 584-603.

Roberts, L. W., & Clifton, R. A. (1992). Measuring the Affective Quality-of-Life of University-Students - the Validation of an Instrument. *Social Indicators Research, 27*(2), 113-137.

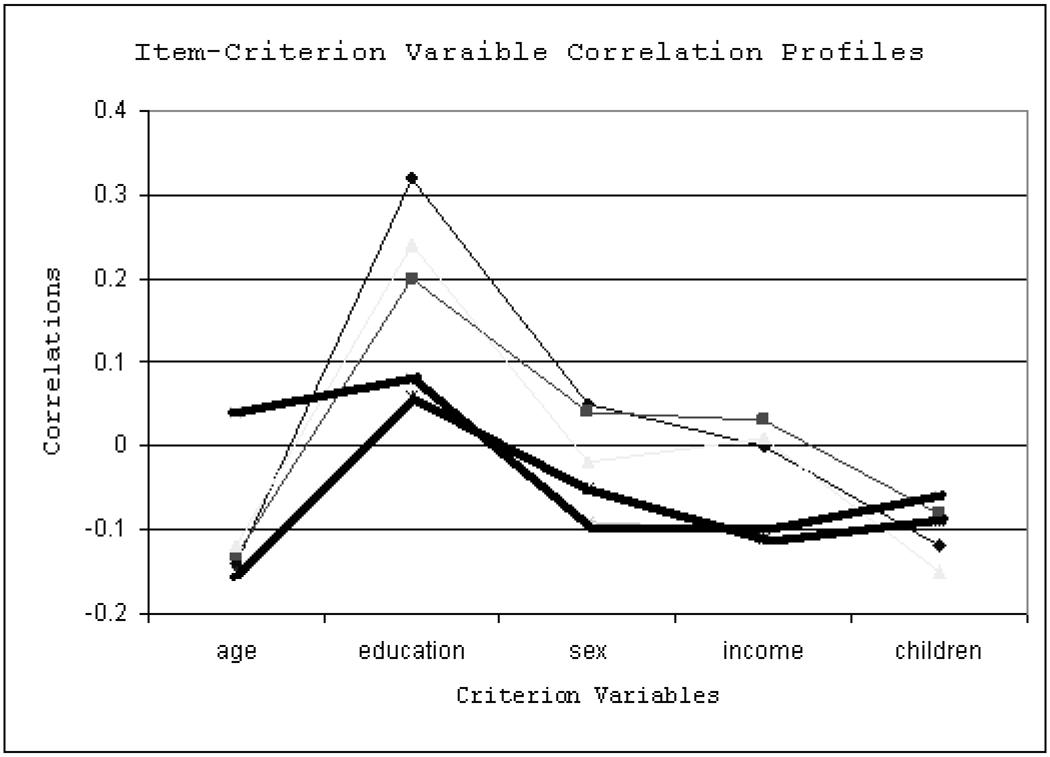
Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*(4), 396-408.

Timmreck, C. W. (1981). *Moderating effect of tasks on the validity of selection tests.* Unpublished doctoral dissertation, University of Houston, Houston, TX, USA.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84*(3), 608-618.

Figure 1.

Correlations between Item Scores and Criterion Variables: Reproduction of the Example from Piazza's (1980) Article



Note: Reproduced with permission from the American Journal of Sociology (University of Chicago Press, permission grant # 61697, Feb 7, 2006).

Figure 2.

Correlations between Item Scores and Criterion Variables: High P^2*, Low Validity*

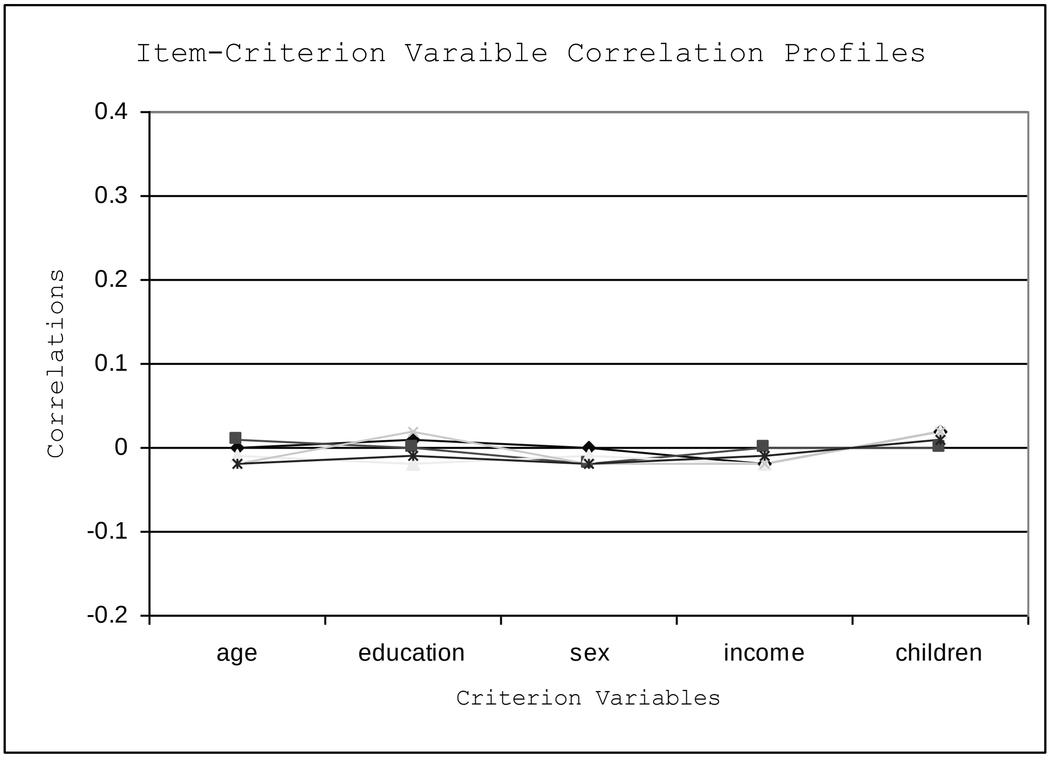


Figure 3.

Correlations between Item Scores and Criterion Variables: Reproduction of Piazza's Example on the -1 to 1 Scale

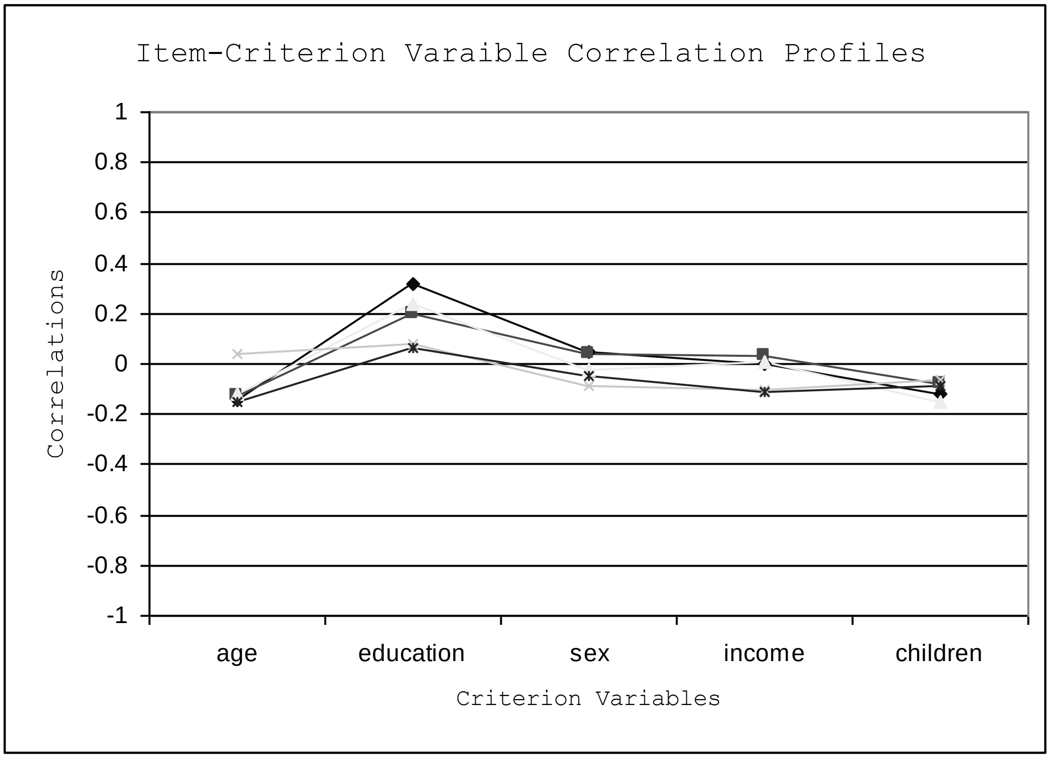


Figure 4.

Correlations between Item Scores and Criterion Variables: Misleading High P^2

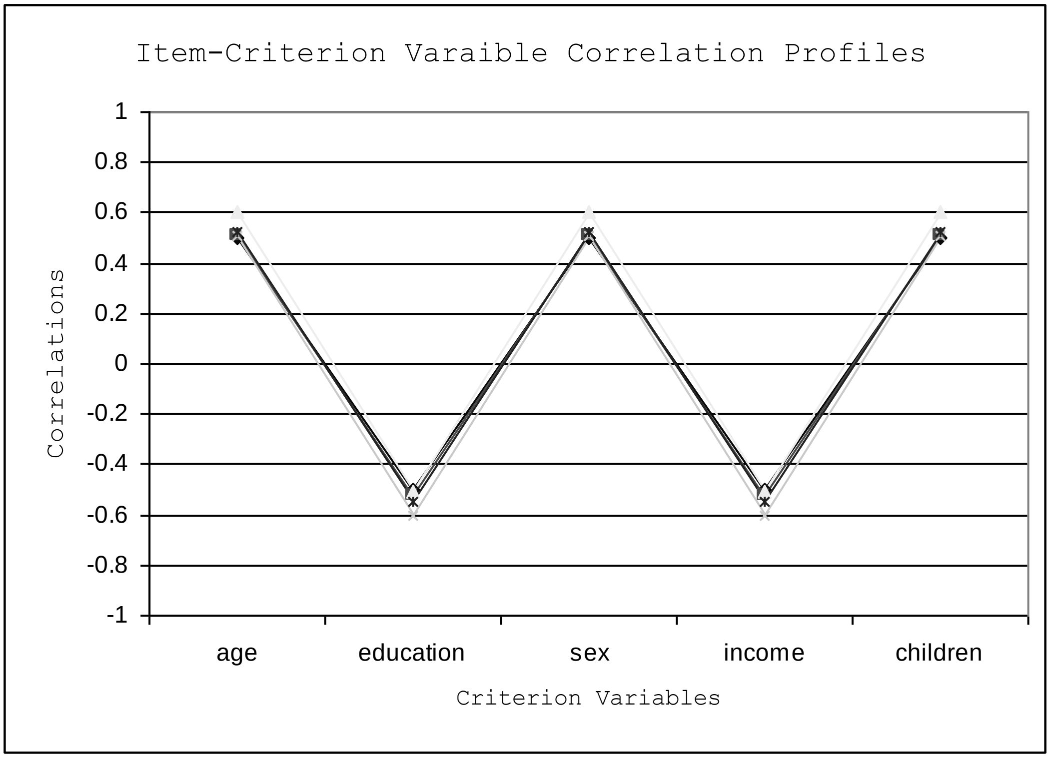


Figure 5.

Correlations between Item Scores and Criterion Variables: Example

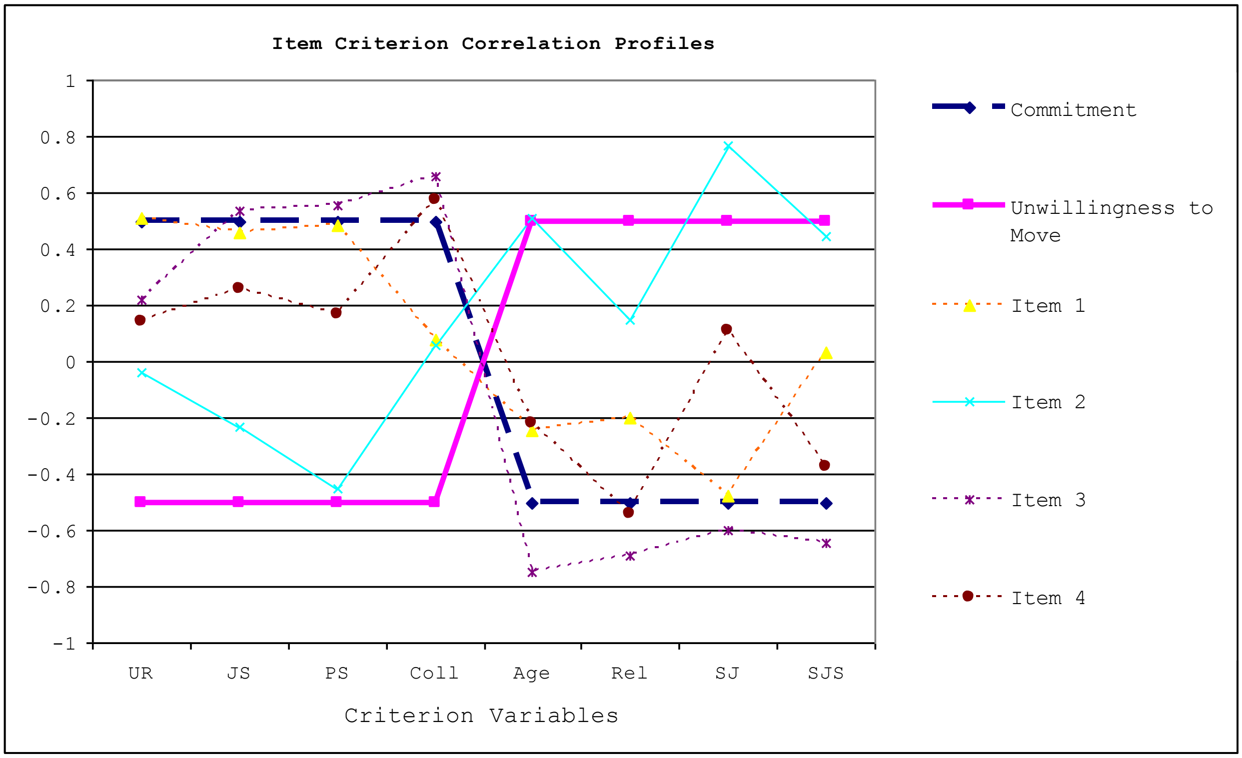


Table 1.

Dataset from the Example

Respondents	Item 1	Item 2	Item 3	Item 4
Professor 1	5	6	6	5
Professor 2	5	4	6	4
Professor 3	4	5	7	5
Professor 4	5	5	4	3
Professor 5	5	5	6	4
Professor 6	3	6	5	5

Table 2.

Responses to Additional Survey Questions

Respondents	Exogenous Theoretically Relevant Variables							
	UR	*JS*	*PS*	*Coll*	*Age*	*Rel*	*SJ*	*SJS*
Professor 1	5	9	8	9	40	1	2	5
Professor 2	4	8	9	6	35	1	1	1
Professor 3	3	7	7	9	30	2	1	1
Professor 4	3	4	5	3	45	4	2	8
Professor 5	2	3	5	2	45	3	2	9
Professor 6	2	2	4	3	50	3	3	7

UR University Rank: High values indicate high rank

JS Job Satisfaction: Satisfaction with job at the current university

PS Pay Satisfaction: Satisfaction with pay at the current university

Coll Collaboration opportunities at the current university

Age Age of the respondent

Rel Number of relatives in the city

SJ Jobs in the city held by close family member (combined score)

SJS Satisfaction of family members with their jobs in the city (combined score)

Table 3.

Hypothesized Direction and Magnitude of Item-Criterion Correlations

Exogenous Theoretically Relevant Variables	Correlations					
	Hypothesized direction of the relationships		Observed Correlations			
	Commitment	Unwillingness to Move	Item 1	Item 2	Item 3	Item 4
UR	0.5	-0.5	0.51	-0.04	0.22	0.14
JS	0.5	-0.5	0.46	-0.23	0.54	0.26
PS	0.5	-0.5	0.49	-0.45	0.56	0.17
Coll	0.5	-0.5	0.08	0.06	0.66	0.57
Age	-0.5	0.5	-0.24	0.51	-0.75	-0.22
Rel	-0.5	0.5	-0.20	0.15	-0.69	-0.54
SJ	-0.5	0.5	-0.48	0.76	-0.60	0.11
SJS	-0.5	0.5	0.03	0.44	-0.65	-0.37