Ambiguity as Moderator

Item Ambiguity as a Moderator of Correlations between Observer Ratings and Self-Reports
John A. Johnson
Pennsylvania State University, DuBois Campus
Abstract
This study demonstrates for the first time the empirical relationship between item ambiguity and the validity of scales on the California Psychological Inventory (CPI). Likert ratings and Goldberg's AMBDEX measure of item ambiguity were used to sort items in the CPI into high- and low-ambiguity subscales. The low-ambiguity subscales predicted Q-sort observer ratings of personality as well as the full scales in a sample of 45 research scientists. High-ambiguity subscale correlations were consistently lower.

Ambiguity as Moderator

Item Ambiguity as a Moderator of Correlations between Observer Ratings and Self-Reports

Using a small sample of items from Gough's (1975) California Psychological Inventory (CPI), Johnson (1984, 1986) has drawn two conclusions regarding the measurement of personality item ambiguity and the relationship between item ambiguity and personality scale validity. He first concluded that simple, direct Likert scale ratings of item ambiguity are more reliable and valid than more complex mathematical indices of ambiguity, such as Goldberg's (1963) AMBDEX. AMBDEX, which estimates item ambiguity by adjusting response inconsistency for endorsement frequency, demonstrated a rather low ($r = .48$) degree of consistency across samples and failed to show convergent correlations with other measures of item ambiguity, including Likert ratings (Johnson, 1984).

Johnson's (1986) second (and counter-intuitive) conclusion was that item ambiguity is positively related to item validity—i.e., ambiguous items are more valid than less ambiguous items. Item validity has many possible meanings; Johnson chose to define it as discriminating power, i.e., the degree to which different groups of subjects show different patterns of endorsement frequency for that item. Johnson found a nonsignificant negative relationship between AMBDEX and discriminating power, but a significant positive relationship between Likert ratings and discriminating power ($r = .45, p < .05$). Johnson interpreted this finding as supporting the contention (Elias, 1951; Johnson, 1981) that an ambiguous item functions like a mini-projective test, wherein persons project their personality onto the item. Because projective tests must be ambiguous,
this explains the positive correlation between ambiguity and validity.

Johnson (1986) himself notes two important limitations of his research that render his conclusions tenuous. First, he sampled only 24 of 480 total CPI items, opening the possibility of sampling errors. Second, by defining item validity as discriminating power, he ignored the content of the items and the personality characteristics of the groups he tested. He suggests a more meaningful analysis would examine the moderating effect of ambiguity on the correspondence between self-report scales and observer ratings of personality. The present study carries out Johnson's suggestions by assessing Likert ratings and AMBDEX values for all 480 CPI items, and examining the moderating effect of these ambiguity indices on self- versus observer measures of personality.

Method

Subjects

Three sets of subjects were used. The first set, tested by Lewis R. Goldberg in the 1960's, included 108 female and 95 male General Psychology students and 179 paid volunteers from a freshman dormitory at the University of Oregon. Subjects completed the CPI twice, allowing Goldberg and Rorer (1964) to calculate AMBDEX values for all 480 CPI items. These AMBDEX values reported in that monograph were used in the present study.

The second set of subjects included 37 female and 32 male students enrolled in an introductory psychology class taught by the author. These subjects were instructed to rate each CPI item's ambiguity on a scale from 1 (perfectly clear) to 5 (very ambiguous). Students received extra credit for participating in the study.
The third set of subjects were 45 research scientists assessed by Gough and Woodworth (1960) at the University of California's Institute of Personality Assessment and Research. These subjects completed the CPI and were observed by 8-10 trained observers who recorded their impressions with Block's (1961) 100-item California Q-set. CPI item responses and composite Q-ratings were reanalyzed for the present study as described below.

Scoring

The research scientists' CPI item responses were first scored along seven scales representing the major dimensions assessed by the Hogan Personality Inventory (HPI; Hogan, 1986). These "CPI-HPI" scales had been created by Hogan and the present author by selecting a "core" of items whose content was clearly and obviously relevant to each dimension, and then adding other items demonstrating significant empirical relationships to one of the "cores" (see Hogan, Carpenter, Briggs, & Hansson, 1985; Hansson, Hogan, Johnson, & Schroeder, 1983; and Johnson, 1983). The seven scales are as follows:

- Intellectance (how bright, clever, and cultured one appears);
- Adjustment (absence of anxiety, depression, guilt, and self-doubt);
- Prudence (conscientiousness, responsibility, and respect for rules and authority);
- Ego Control (self-control, orderliness, planfulness);
- Ambition (energy, initiative, perseverance, leadership qualities);
- Sociability (affiliative tendencies) and
- Likeability (tolerance, even-temperedness, cooperativeness).

Kuder-Richardson Formula 20 reliability estimates for the scales range
from .66 for Ego Control to .89 for Intellectance. The CPI-HPI scales correlated in a meaningful fashion with observer ratings, academic achievement, job performance, and other criteria across many subject samples. In some cases, the CPI-HPI scales outpredicted the original CPI scales.

Next, each CPI-HPI scale was divided at the median AMBDEX and median Likert-rating value of ambiguity. This yielded 28 additional scale scores—two high- and two low-ambiguity subscales for each dimension according to the two indices of ambiguity.

Finally, the 100 Q-set items were sorted rationally by two independent raters into the same seven HPI dimensions. Those items upon which both raters agreed and which also showed significant empirical correlations with the CPI-HPI scale scores in the scientist sample were retained as the final "Q-HPI" scales. Two of these Q-HPI scales (representing Adjustment and Prudence) contained only two items and were consequently not used in further analyses.

Analyses

Because Johnson (1984) had reported a relatively low cross-sample reliability of AMBDEX for 24 items, the first analysis was a simple attempt at replication by intercorrelating AMBDEX values for Goldberg and Rorer's (1964) three samples.

Johnson (1984) had also reported that AMBDEX and Likert ratings of ambiguity were unrelated for his 24 items. The second analysis examined whether that finding would replicate by correlating AMBDEX values from all 480 items with the corresponding Likert ratings.
The final set of analyses reflect the central concern of the study: Does item ambiguity affect the validity of personality scales (where validity is defined as the correspondence between self-report scores and observer ratings)? This question was addressed by first correlating the full Intellectance, Ambition, Likability, Sociability, and Ego Control CPI-HPI self-report scales with their corresponding Q-HPI observer rating scales. Then, separate correlations were computed for high- and low-ambiguity half-scales, defined by AMBDEX and then by Likert rating values. Differences between high- and low-ambiguity correlations were tested for significance by the procedure suggested by McNemar (1969, p. 158).

Results

Intercorrelations between AMBDEX values for Goldberg and Rorer's three groups were .45, .38, and .49 (all ps < .001), indicating that Johnson's (1984) estimate of cross-sample stability based on 24 items was essentially correct. The magnitude of cross-sample correlations for AMBDEX found in this study also replicates the cross-sample $r$ of .46 reported by Goldberg (1968) in his study of ambiguity in an adjective check list.

AMBDEX values from the three samples correlated significantly with Likert ratings of ambiguity, .32, .24, .29, (all ps < .001). These correlations are comparable to the $r$ of .30 between AMBDEX and ambiguity ratings reported by Goldberg (1968) in his adjective check list study. When the AMBDEX values from the three samples were averaged into a composite set of AMBDEX values, this composite correlated .36 with the Likert ratings. This finding indicates that Johnson's failure to find a
relationship between AMBDEX and Likert ratings may be an artifact of item sampling.

The central findings on item ambiguity and scale validity are summarized in Table 1. This table first shows significant relationships between the full CPI-HPI scales and their corresponding Q-HPI scale correlates, ranging from a low of .30 for Ego Control to a high of .43 for Sociability. When the full scales are divided into high- and low-ambiguity half-scales, only two of 20 correlations lose their significance. For the AMBDEX half-scales, the low-ambiguity scale correlations were higher than the high-ambiguity half-scales in all five cases. For the Likert defined half-scales, low-ambiguity scale correlations were higher in four of the five cases. None of these differences between correlations for either index of ambiguity were significant by McNemar's statistical test, however.

---

Insert Table 1 about here

---

Discussion

The present study has both psychometric and substantive theoretical implications. From a psychometric viewpoint, AMBDEX values appear somewhat (but not totally) unreliable in the sense that these values will differ from sample to sample. Cross-sample correlations of about .45 compare unfavorably to reliability coefficients of about .90 for ambiguity ratings (Goldberg, 1968; Johnson, 1984). AMBDEX is still worth considering, however, particularly if one forms composite AMBDEX values from several samples. The correlation between ambiguity ratings and AMBDEX was higher
for the composite AMBDEX scores than for any of the individual samples.

The substantive implication of this study concerns the impact of item ambiguity on the validity of personality scales. Here we find a clear trend for half-scales composed of relatively nonambiguous items (defined either by AMBDEX values or Likert ratings) to correlate higher with observer ratings of personality than half-scales composed of relatively ambiguous items. In fact, the nonambiguous half-scales showed correlations with the rating criteria at about the same magnitude as the full scales, despite their shorter length and consequent reduced reliability. Although the magnitude of differences between correlations from ambiguous and nonambiguous half-scales did not reach statistical significance, the consistency of differences across scales is striking and merits further attention.

These results call into question Johnson's (1986) suggestion that personality self-report items operate like miniature projective tests and that ambiguous items are therefore more valid than nonambiguous items. The present data suggest just the opposite, that nonambiguous items are more valid. This conclusion could be made stronger with a replication study using an alternative to the Q-set rating scales (adjective rating scales, for example) and by addressing the Prudence and Adjustment dimensions not studied in the present work.

A final comment concerns the extension of the present methodology to other item characteristics. Recently there has been a resurgence of interest in "itemetrics" (Angleitner, John, & Lohr, 1986), which is the measurement of properties of individual items. For example, Werner and Pervin (1986) assessed the area of psychological functioning addressed (cognitive, preferences, feelings, behavioral) in several personality
inventories, and Christian, Burkhart, and Gynther (1978) rated the subtlety of items in the MMPI. Unfortunately, many of these itemmetric research studies have stopped at the measurement of item characteristics without considering how the characteristics affect the validity of scales. An ambitious, yet worthwhile research program would extend the itemmetric-validity methodology of the present study to other item characteristics.
References


Table 1

Self-Peer Correlations for Full, High-, and Low-Ambiguity Scales

<table>
<thead>
<tr>
<th>Q-set Dimensions</th>
<th>Likert Ratings</th>
<th>AMBDEX Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Scale</td>
<td>Ambiguity Level</td>
</tr>
<tr>
<td>Intellectance</td>
<td>.36**</td>
<td>.19</td>
</tr>
<tr>
<td>Ambition</td>
<td>.40**</td>
<td>.31*</td>
</tr>
<tr>
<td>Likeability</td>
<td>.41**</td>
<td>.29*</td>
</tr>
<tr>
<td>Sociability</td>
<td>.43**</td>
<td>.39**</td>
</tr>
<tr>
<td>Ego Control</td>
<td>.30*</td>
<td>.29*</td>
</tr>
</tbody>
</table>

* p < .05

** p < .01 (both one-tailed)