Item Ambiguity, Subtlety, and Discriminating Power in the California Psychological Inventory

John A. Johnson
Pennsylvania State University
College Place
DuBois, Pennsylvania 15801
(814) 371-2800

Presented at the 57th Annual Meeting
of the Eastern Psychological Association
at New York, April, 1986
ITEM AMBIGUITY, SUBLTLE, AND DISCRIMINATING POWER IN THE CALIFORNIA PSYCHOLOGICAL INVENTORY. John A. Johnson, Pennsylvania State University.

The ability of 24 CPI items to discriminate between 12 subject groups (total N=672) was regressed on four indices of ambiguity and five indices of subtlety. Items receiving highest Likert scale ratings of ambiguity and drawing the greatest consensus on response meaning best predicted discriminating power. This suggests that item responding is an unconscious projection of one's social identity rather than a simple report of actual behavior.
Item Ambiguity, Subtlety, and Discriminating Power in the California Psychological Inventory

Topical Session Preference: PERSONALITY (Personality Measurement)

Problem or Major Purpose

This paper describes the joint impact of item ambiguity and item subtlety on item validity. **Item ambiguity** is uncertainty about the denotative reference or meaning of an item. **Item subtlety**, on the other hand, is uncertainty about the psychological implications of an item response (i.e., how the scorer interprets the response). Subtle items are not necessarily ambiguous; a respondent can be perfectly clear on the meaning of an item yet uncertain about the psychological significance of response options. **Item validity** has many possible meanings; here it is defined as **discriminating power**, i.e., the degree to which different groups of subjects show different patterns of endorsement frequency for that item.

Operational Definitions of Item Ambiguity

Johnson (1984) has provided a review of different operational definitions of item ambiguity and subtlety. Those definitions employed in the present study are described below and identified with the acronyms used by Johnson.

**Response inconsistency (RESINC)**—the percentage of subjects in a particular group changing their response to an item upon retesting—was suggested as an index of ambiguity by Benton (1935).

According to Fricke (1957) and Hanley (1962), ambiguous items are difficult to answer accurately; this results in random responding and therefore a tendency toward **balanced endorsement frequencies** (BAEF; roughly 50% responding "True" and 50%, "False"). BAEF has been found to correlate highly with RESINC in many studies (Goldberg, 1963).
Goldberg (1963) agrees that item ambiguity can be operationalized by response inconsistency; however, he does not see any intrinsic relationship between endorsement frequency and ambiguity. He therefore regards the correlation between RESINC and BALEF as a confound. Goldberg's statistical ambiguity index (AMBDEX) adjusts response instability by endorsement frequency.

Johnson (1984) found that direct Likert scale ambiguity ratings (AMBRAT) were more reliable than RESINC, BALEF, or AMBDEX, but suggests that final judgment on the "best" index should come after studies of item ambiguity and validity.

Item Ambiguity and Validity

Intuitively, ambiguous items seem to be less valid than nonambiguous items, because ambiguity makes providing an objective, veridical response difficult. This attitude is reflected in the work of Goldberg (1963; 1968). An alternative interpretation (Elias, 1951) is that an ambiguous item functions like a mini-projective test, wherein persons project their personality onto the item; this implies a positive correlation between ambiguity and validity. Underlying these two views about ambiguity are assumptions about whether item responding is a conscious, objective reporting of actual behavior or an unconscious, subjective presentation of a constructed self-image (Johnson, 1981). Empirical studies using different indices of ambiguity are needed to help resolve the theoretical issues.

The first major empirical study of item ambiguity and validity (Benton, 1935) showed no relationship between ambiguity and item discriminating power. Eisenberg (1941) found a negative relationship between item ambiguity and item discriminating power, but his results are confounded by the fact that the
criterion groups are defined by the same items scored for ambiguity. Two other studies (Gordon, 1953; Isard, 1956), using unconfounded external criteria, found a positive relationship between item ambiguity and item discriminating power.

The first two research questions, then, are: (1) Which operationalization of ambiguity best predicts item discriminating power, RESINC, BALEF, AMBDEX, or AMBRAT? and (2) Is ambiguity positively or negatively related to discriminating power?

Operational Definitions of Item Subtlety

Item subtlety can be assessed by the percentage of nonexperts who provide an incorrect rating (INCRAT) of the psychological meaning (keying) of the item (Duff, 1965). A second possible operational definition is semantic instability (SEMINS), the percentage of subjects changing their interpretation of item meaning over time. A third index of subtlety is the percentage of a group choosing the nonmodal (NONMOD) interpretation. NONMOD measures disagreement on the "lay" social meaning of an item response, irrespective of the standard, "professional" keying for the item. Whether the modal interpretation is incorrect (MODINC), i.e., when the modal interpretation fails to correspond to the item keying, is a fourth index of item subtlety. Finally, one can present subjects with items and their keyed meanings and gather subtlety ratings (SUBRAT).

Based on convergent correlations and reliabilities, Johnson (1984) tentatively suggests that INCRAT is the best subtlety index, although he again suggests reserving judgment until validity studies are conducted.
**Item Subtlety and Validity**

Studies of item subtlety by Gynther and his associates (Burkhart, Christian, & Gynther, 1978; Burkhart, Gynther, & Christian, 1978; Christian, Burkhart, & Gynther, 1978; Gynther, Burkhart, & Hovanitz, 1979; Hovanitz & Gynther, 1980; Hovanitz, Gynther, & Marks, 1983) show a clear consensus: subtle items are less valid than more obvious items. Hovanitz, et al. caution, though, that their results pertain only to the use of the MMPI with nonclinical samples. She suggests (personal communication, March 15, 1984) that subtlety may be less important when measuring normal traits with neutral social desirability. The present study extends Gynther, et al.'s work by using an inventory of normal personality rather than psychopathology and by using multiple operationalizations of item subtlety.

The two principle research questions here are: (1) Which, if any, operationalization of item subtlety will predict item discriminating power in an inventory of normal personality? and (2) Will item subtlety predict item discriminating power beyond what can be predicted by item ambiguity?

(4) Subjects

The author's personality test archives include item responses on the California Psychological Inventory (CPI; Gough, 1975) from 12 groups of subjects tested between 1950 and 1985. Subjects are all male except where indicated. The groups are: 44 consulting engineers, 69 incarcerated murderers, 19 students tested at Johns Hopkins University during the 1960's, the 45 research scientists, 66 Berkeley engineering students, and 100 Air Force officers described by Hogan (1969), 40 inner-city black teenagers, 49 Baltimore County police officers (2 female), 101 parents of intellectually precocious youth (about half female), 22
accelerated high school students (mostly female), 34 Hopkins evening students tested during the 1970's (sex unknown), and 83 students from a branch campus of the Pennsylvania State University, tested in 1984 (about half female). The total N for the study was 672.

(5) Procedure

Measures

The present study used a subset of Johnson's (1984) indices of item ambiguity and subtlety. The procedures used by Johnson for obtaining his indices are described below.

Johnson selected 24 items representing the content of the four major factors of the California Psychological Inventory (Megargee, 1972). He obtained RESINC, BALEF, and AMBDEX values for these items from Goldberg and Rorer's (1964) monograph, which contains CPI item-retest-statistics for 382 college students. He had a group of 65 students to rate the clarity of meaning of his 24 items on a 1-5 Likert scale; the average inverse of these values yielded AMBRAT scores.

Johnson presented the 24 CPI items to his group of 65 students and to a separate group of 42 students and asked them to identify which of the four major themes in the CPI the item tapped. INCRA (incorrect rating) values in the present study were computed by averaging the INCRA values from Johnson's two samples. The group of 42 students was retested after two weeks; the percentage changing their interpretations of item response meaning yielded SEMINS (semantic instability) scores.

NONMOD (disagreement on the "lay" social meaning of an item response) and MODINC (whether the modal score was incorrect) were computed by averaging values for these two indices across Johnson's two samples. Finally, Johnson gathered
SUBRAT scores by presenting a group of 10 educational administrators and staff with the 24 CPI items with their keyed meanings and asking for subtlety ratings on a 1-5 Likert scale.

Analyses

For each of the 24 CPI items originally studied by Johnson (1984), 12 indices of discrimination (DISCRM) were computed by subtracting the endorsement frequency for the total group (P) from the endorsement frequency for the group in question (p) and dividing by the square root of P times Q over 672. DISCRM is essentially a "critical ratio" from statistics (Benton, 1935; Duff, 1965). A principle-components factor analysis of the 24 by 12 matrix of DISCRM values indicated the presence of a general factor of item discriminating power, so factor scores were generated from the first unrotated component.

This DISCRM factor score was then correlated with all of the ambiguity and subtlety indices described above, first as zero-order correlations, and then correlations partialling out the possible influence of item variance and social desirability. Item variance was computed from the 672 subjects in the present study; social desirability values were taken from the Goldberg and Rorer (1964) monograph. Finally, DISCRM scores were regressed on the best ambiguity and subtlety predictors to ascertain whether ambiguity and subtlety contribute independently to the prediction of validity.

Results, Implications, and Conclusions

Partial correlations controlling for the effects of item variance and social desirability did not differ significantly from the zero-order correlations; therefore only the latter will be presented. The full matrix of correlations is presented in Table 1.
Table 1 indicates that the RESINC, BALEF, and AMBDEX indices of ambiguity all fail to predict item validity. Only the AMBRAT index correlates significantly, and in a positive direction (r= .45). This replication of the positive correlation between ambiguity and discriminating power found by Gordon (1953) and Isard (1956) supports the notion that items are like mini-projective tests. That is, idiosyncratic interpretations contribute to validity because they reflect the test-taker's projected personality.

The various subtlety indices show nonsignificant to marginally significant negative correlations with the ambiguity indices and with item discriminating power. This means that there is a trend toward ambiguous items possessing response options whose psychological meanings are the most obvious. Also, there is a trend toward obvious items discriminating more powerfully between groups than subtle items. A more conservative interpretation of these low correlations would be that subtlety is less important in inventories of normal personality, just as Hovanitz suggested. (Note, however, that social desirability did not act as a suppressor variable in the partial correlations, as Hovanitz, et al. might have predicted.)

The best subtlety index predictor of discriminating power is NONMOD (r= -.34). This implies that items on which there is a consensus of meaning are most discriminating, but that consensus need not correspond to the interpreter’s "correct" keying for that item. NONMOD, but no other variable, contributes significantly in a regression equation where DISCRM is the criterion and AMBRAT,
the other predictor \( R = .60 \).

The study contains two major limitations that can be corrected by additional research. First, the CPI item sample, though representative, is small. It is hard to say, therefore, whether the marginality of some of the correlations is due to sample size or actual nonsignificance. Ambiguity and subtlety ratings for all 480 CPI items would be painstaking, yet possible, to obtain. Second, the "validity" data here take the form of discriminating between groups without regard to the content of the items or the psychological traits that might make each group unique. A logical next step would be to look at correlations between full CPI scales and content-relevant criteria (e.g., peer ratings of the same trait) and to examine the moderating influence of the ambiguity and subtlety indices.

In summary, the present study suggests that the most valid (discriminating) items appear to be ambiguous and yet associated with responses on which there is a consensus of psychological meaning (but not necessarily the "correct" meaning according to the standard key for that scale). These data are consistent with Johnson's (1981) view that responses to personality items should not be regarded as "self-disclosures" (i.e., conscious, objective, veridical descriptions of actual behavior) but rather as "self-presentations" (i.e., relatively unconscious, subjective, projections of a socially-constructed identity).
Ambiguity & Subtlety

(8) References


Table 1

Correlations between Discriminating Power, Ambiguity and Subtlety

<table>
<thead>
<tr>
<th>Measure</th>
<th>DISCRM</th>
<th>RESINC</th>
<th>BALEF</th>
<th>AMBDEX</th>
<th>AMBRAT</th>
<th>INCRA</th>
<th>SEMINS</th>
<th>NONMOD</th>
<th>MODINC</th>
<th>SUBRAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISCRM</td>
<td>—</td>
<td>07</td>
<td>19</td>
<td>-13</td>
<td>45</td>
<td>-18</td>
<td>-13</td>
<td>-34</td>
<td>-10</td>
<td>05</td>
</tr>
<tr>
<td>RESINC</td>
<td>—</td>
<td>—</td>
<td>86</td>
<td>55</td>
<td>42</td>
<td>-33</td>
<td>-32</td>
<td>16</td>
<td>-44</td>
<td>04</td>
</tr>
<tr>
<td>BALEF</td>
<td>19</td>
<td>86</td>
<td>—</td>
<td>10</td>
<td>11</td>
<td>-11</td>
<td>-12</td>
<td>16</td>
<td>-42</td>
<td>10</td>
</tr>
<tr>
<td>AMBDEX</td>
<td>-13</td>
<td>55</td>
<td>10</td>
<td>—</td>
<td>—</td>
<td>-39</td>
<td>-40</td>
<td>08</td>
<td>-40</td>
<td>-06</td>
</tr>
<tr>
<td>AMBRAT</td>
<td>45</td>
<td>42</td>
<td>42</td>
<td>11</td>
<td>—</td>
<td>-11</td>
<td>11</td>
<td>16</td>
<td>25</td>
<td>31</td>
</tr>
<tr>
<td>INCRA</td>
<td>-18</td>
<td>-33</td>
<td>-39</td>
<td>-11</td>
<td>-21</td>
<td>—</td>
<td>22</td>
<td>12</td>
<td>13</td>
<td>03</td>
</tr>
<tr>
<td>SEMINS</td>
<td>-13</td>
<td>-32</td>
<td>-40</td>
<td>-12</td>
<td>11</td>
<td>22</td>
<td>—</td>
<td>31</td>
<td>—</td>
<td>18</td>
</tr>
<tr>
<td>NONMOD</td>
<td>-34</td>
<td>16</td>
<td>08</td>
<td>16</td>
<td>12</td>
<td>49</td>
<td>25</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MODINC</td>
<td>-10</td>
<td>-44</td>
<td>-42</td>
<td>-17</td>
<td>-25</td>
<td>83</td>
<td>13</td>
<td>31</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SUBRAT</td>
<td>05</td>
<td>04</td>
<td>10</td>
<td>-06</td>
<td>31</td>
<td>20</td>
<td>-01</td>
<td>03</td>
<td>18</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. Decimal points omitted from all correlation coefficients. Correlations greater than 30 significant at the .10 level; those greater than 34, at the .05 level; those greater than 44, at the .01 level (both one-tailed). Correlations based on an N of 24 items in all cases. Measures based on the following number of subjects: AMBRAT, N=65; SEMINS, N=42; NONMOD, INCRA, MODINC, N=107; SUBRAT, N=10; RESINC, BALEF, AMBDEX, N=382.