A MODEL OF ITEM AMBIGUITY IN PERSONALITY ASSESSMENT

LEWIS R. GOLDBERG
University of Oregon
and
Oregon Research Institute

One of the most common criticisms heard from subjects upon taking almost any popular personality inventory is that "so many of the items seem ambiguous." Further questioning will often elicit statements such as "Many of the items I'd have answered one way today and another way tomorrow" or "The items seemed so ambiguous that I was never sure how to respond."

That subjects may become annoyed with the form, content, or response structure of some particular test items would not in itself be cause for grave concern, if the instruments which utilized these items yielded highly satisfactory test scores. However, few psychologists would question the need for improving present-day personality tests, and one might well begin by investigating the apparent difficulties encountered by subjects in responding to individual test items.

In any study of item characteristics, one cannot help but be impressed by the marked inconsistency of response elicited by most personality test items. Evidence of such response instability has

---

1 The author gratefully acknowledges the stimulation and help provided by Paul J. Hoffman, Rolfe LaForge, and others at the Oregon Research Institute. This paper has profited from the critical reading of earlier drafts by Hoffman and LaForge as well as by Donald Fiske, Leonard Rorer, Jerry S. Wiggins and Nancy Wiggins. This project has been supported by two grants from the Graduate School of the University of Oregon, and Grant #G-25123 from the National Science Foundation.
been reported in the psychometric literature since the early 1930's. Table 1 lists some representative figures from this literature showing the percentage of items changed on retest by the average subject for some typical early personality inventories.

Table 2 lists the percentage of items changed by the average subject for six more recently developed personality and "response set" measures (Goldberg, Dufort & Hammersley, in preparation). Approximately 300 college undergraduates were administered a one-hour battery of tests, and 10 random samples from this population were retested over 10 different test-retest intervals. Table 2 lists the average percentage of items changed for each of the six measures for a three-week test-retest interval. Since the amount of item change is obviously related to the number of response options considered (Fiske, 1957), the data are analyzed for different

**TABLE 1**

*Percentage of Items Changed by the Average Subject under Test-Retest Conditions:
Some Representative Figures from the Psychometric Literature*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Personality Test</th>
<th>Test-Retest Interval</th>
<th>Number of Response Options Considered</th>
<th>Average % Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenz (1934)</td>
<td>Bernreuter Personality Inventory</td>
<td>1 month</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Neprash (1936)</td>
<td>Thurstone Personality Schedule</td>
<td>2 weeks</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 weeks</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 weeks</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Benton and Stone (1937)</td>
<td>Landis and Zubin Personal Inquiry Form</td>
<td>immediate</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 day</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 days</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 days</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 days</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 days</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 days</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21 days</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>Farnsworth (1938)</td>
<td>Bernreuter Personality Inventory</td>
<td>1 year</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 years</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 years</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>Eisenberg and Weisman (1941)</td>
<td>Thurstone Neurotic Inventory</td>
<td>23–28 days</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Glaser (1949)</td>
<td>California Test of Personality</td>
<td>1 month</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Chance (1955)</td>
<td>Bell Adjustment Inventory</td>
<td>14 days</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Strong (1962)</td>
<td>Strong Vocational Interest Blank</td>
<td>3 days</td>
<td>3</td>
<td>21</td>
</tr>
</tbody>
</table>
TABLE 2
Percentage of Items Changed upon Retest by the Average Subject as a Function of the Number of Response Categories

<table>
<thead>
<tr>
<th>Scale</th>
<th>Number of Response Options</th>
<th>Percentage Change when Items are Scored as:</th>
<th>Absolute % Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dichotomous</td>
<td></td>
</tr>
<tr>
<td>MMPI: Factor A</td>
<td>2</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>MMPI: Factor R</td>
<td>2</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Berg Perceptual Reaction Test</td>
<td>4</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Rust and Davies Reported Behavior Inventory</td>
<td>3</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Bass Social Acquiescence Scale</td>
<td>3</td>
<td>22</td>
<td>36</td>
</tr>
<tr>
<td>Couch and Keniston Agreement Response Set Scale</td>
<td>7</td>
<td>24</td>
<td>34</td>
</tr>
</tbody>
</table>

numbers of response categories, thereby permitting the direct comparison across measures utilizing different response formats.

Although the findings presented in Tables 1 and 2 indicate a marked degree of response instability on items from both older and more recent personality inventories, an even more significant indication of intra-individual variability becomes evident when one examines the test-retest reliability coefficients for modern personality scales. The majority of these average no more than .70 (for example, Jackson, 1961), indicating that roughly one-third of the total scale variance is error of measurement, if we make the assumption that the personality trait to be measured is relatively invariant. That is, although psychologists might well expect some changes in responses from one situation to another different one, most psychologists concerned with human measurement expect some stability of responses over time when the testing conditions appear to be reasonably constant and the traits to be measured are assumed to be relatively enduring ones. In this paper, it is assumed that some important personality traits are stable over short periods of time, and the discussion will be limited to the psychometric measurement of such stable traits. One possible way to reduce the error now associated with the measurement of these traits is to construct more stable items to use as initial item pools for later scale development.
Parameters of Item Stability

Common sense considerations, augmented by the reflections of tested subjects, would suggest that a crucial component of item stability is certainly item ambiguity. Ambiguity has been traditionally defined as doubtfulness or uncertainty in the meaning of a stimulus; it often denotes a stimulus whose meaning is open to various interpretations. Ambiguity can be measured subjectively by rating methods or objectively as some function of either of two indices: (a) inter-individual variability in the meaning of a stimulus (e.g., Broen, 1960) or (b) intra-individual variability in meaning over repeated administrations of the stimulus (i.e., interpretative instability). It is apparent that indices utilizing either of these two measures will usually be highly correlated, for a stimulus which elicits disparate meanings among a group of persons will typically elicit unstable meanings for an individual over time. The relationship between inter-individual and intra-individual response variability in personality assessment has been documented by Mitra and Fiske (1956) and Fiske (1957). The relationship between intra-individual variability and ambiguity has such strong rational appeal that investigators have been led to use intra-individual response variability (item instability) as a direct measure of item ambiguity.

However, there is a serious error in such a practice. It has been demonstrated repeatedly that items of extreme endorsement frequencies (for example, items to which 90–100 per cent of a population choose the same alternative) tend to be responded to more consistently (i.e., with a smaller percentage of subjects who change their responses) over repeated administrations of the item than more balanced items (i.e., items to which a population's choices approach a 50–50 split between alternatives). In other words, items at the extremes of any attribute continuum have been shown to elicit less change in responses than those reflecting positions in the middle range of the attribute (Frank, 1936; Hertzman & Gould, 1939; Eisenberg & Wesman, 1941; Mitra & Fiske, 1956; Crockett, Bates & Caylor, 1958). This finding, of the relative stability of extreme items, has occurred in virtually all studies, regardless of the assessment instrument utilized. It seems directly analogous to findings from the area of ability testing, where items at the average
difficulty level for a population tend to be less stable over retesting
than items of extreme ease or difficulty. Figure 1 is an example of
the typical curvilinear relation found between stability and endorse-
ment frequency.\(^2\)

Since items seen by subjects as highly “ambiguous” can dichoto-
mize the attribute continuum at many points, it is desirable to
eliminate the effect of differences in endorsement frequency from
an index of item ambiguity. That is, the practice of utilizing the
percentage of subjects changing their responses to the item as an
index of item ambiguity leads one to the position that items

![Endorsement Frequency Graph]

---

\(^2\) The relationship illustrated in Figure 1 does not result solely from that
existing between the standard error of a proportion and the magnitude of
the proportion (in this case, the item’s endorsement frequency). Changes in
endorsement frequency due to sampling errors provide a lower bound for the
degree of instability of any item, but most present-day personality test items
are so unstable that this lower bound does not serve as a useful approximation
to the item’s actual instability; that is, the correlation between the mean
shift in endorsement frequency over two administrations of an item and the
percentage of subjects changing their responses to the item is virtually zero
(for example, \(r = .13\) [and \(.09\)] for 95 male [and 108 female] college students
administered the MMPI and then retested after four weeks).
dichotomizing an attribute continuum at some extreme point are all less ambiguous than items which more evenly dichotomize the continuum. An alternative proposal is to formulate an index of item ambiguity which is based upon the percentage of response change but which takes into account the effects of item imbalance.

The Response to a Monotone Item

To more fully understand how such an index might be constructed, let us consider the response of a subject to a test item. It is important to distinguish between two major forms of items, called “monotone” and “non-monotone” items by Coombs (1952) and Green (1954). The defining characteristic of a “monotone” item is that as the magnitude of the attribute measured by the item increases, the probability of any single response category (“True,” “False”; “Yes,” “No”; etc.) increases monotonically with it. That is, the item dichotomizes the attribute continuum such that there is only one boundary between “Yes” and “No” regions. The item “I am tall” is an example of a monotone item; as height increases in the sample, the probability of a “True” response increases monotonically with it. Now consider a “non-monotone” item—for example, “I am of medium height.” In this case, as with all non-monotone items, the probability of a particular response is not monotonically related to the underlying attribute continuum. The item divides the continuum into three or more regions; at least two boundaries are required to separate “No” from “Yes” regions.

Although some factual items, as well as many attitude items, are non-monotonic, the majority of our present-day personality test items appear to be monotones in form. Take the hypothetical item, “I am a shy person,” or—a similar item—the adjective “Shy” administered in an adjective check list. For persons who see themselves as forceful and extroverted, these items are easily checked “False.” As progressively more retiring sorts of persons are selected, one begins to find subjects hesitating in responding to the item, as if they were asking themselves, “Just how shy does a person have to be before he considers himself ‘shy’?”. Eventually one finds subjects responding “True” to this item, and as progressively more shy individuals are selected, one would probably find that the ease
with which they answer the item in the "True" direction increases.

Being a bit more precise, for an hypothetically "honest" individual the decision to respond "True" or "False" to a dichotomously-scored monotone item is assumed to be a joint function of: (a) the perceived boundary established by the item on the attribute continuum, and (b) the individual's perceived position on this continuum. If the item is perceived as having a boundary on one side of the individual's own position, one alternative is checked; if the item is seen as having a boundary on the other side of the individual's position, the other alternative is checked.8

Figure 2 is an attempt to graph an hypothetical frequency distribution of subjects on the attribute continuum of "perceived shyness" and to indicate where the item "Shy" in an adjective check

---

8 As Edwards (1957) and his associates have frequently warned, persons may differ in their tendencies to admit socially undesirable aspects of themselves; consequently, an "hypothetically honest" individual may be rare. Social desirability considerations, however, can probably best be conceptualized as influencing a subject's perception of his own position on the attribute continuum (and/or the degree he is willing to report an "honest" appraisal of his position) and thus may leave the essential components of this more simple model unchanged.
list might fall on this underlying continuum. Note that 70 per cent of the population sees this item as establishing a boundary above the point on the attribute continuum where they see themselves and thus respond “False” to the item; the remaining 30 per cent perceives the item as establishing a boundary below the point where they see themselves and thus respond “True.” For those individuals who see themselves far from the boundary, B, (i.e., in regions A or C of this distribution), the decision in regard to this item would be an easy one to make, since the item’s boundary should be perceived as quite distant from the individual’s position. Individuals close to B, however, should face more difficulty since the boundary lies closer to their own position on the attribute continuum.

Those individuals who encounter the most difficulty in making a decision upon the first presentation of an item should be the persons most likely to change their response to the item upon its subsequent presentation. Assuming the usual normal frequency distribution of subjects on most attribute continua, items whose boundaries are seen by the great majority of a population as quite distant from their own (i.e., extreme items) will be responded to without difficulty for most of the population upon the first presentation of the item and will be changed only by some members of that very small subsection of the population which is itself very extreme on the attribute continuum. On the other hand, items with boundaries falling near the middle of the attribute continuum (i.e., balanced items) will find many individuals having difficulty in responding to them and many more individuals changing their responses upon retesting.

A Model of Item Ambiguity

Obviously the boundary established by an item is not perceived at the same point on the attribute continuum by everyone in the population. More realistically the boundary of an item can be conceived as occupying a “band” on an attribute continuum corresponding to its perceived position by different members of a population. One could conceptualize the “equivocality-band” of an item as the range of disagreement in regard to the item’s boundary on an attribute continuum. If every individual in the population sees the item’s boundary as falling at the same position, the item could be said to have a minimum equivocality band; if some individuals perceive the item’s boundary as being extreme in one direction while
other individuals perceive it as extreme in the other direction, the item could be said to have a maximum equivocality band.

It is assumed that there is a strong relationship between the degree to which different persons agree in their positioning of the boundary of an item on an attribute continuum (upon the first administration of the item) and the extent to which individuals are consistent in positioning the boundary upon successive presentations of the item. That is, it is assumed that items of broad equivocality bands (as defined by different members of the population on the first administration) will be items of great intra-individual variability in their positioning over repeated administrations of the items and can therefore be considered as having wide "ambiguity bands."

Previously, it has been hypothesized that the closer an individual perceives his position\(^4\) on an attribute continuum to be to the boundary of an item, the more difficulty he will have in responding to the item upon its first administration and the more likely will he be to change his response upon its readministration. In addition, it seems reasonable to assume that the wider the ambiguity band of an item (i.e., the greater the intra-individual variability in boundary positioning) the less stable should be the item. Conversely, the narrower the ambiguity band of the item, the more stable it should be.

In summary, then, the stability of any item (defined in the traditional way as the percentage of individuals who give a consistent response to the item over repeated administrations) depends upon (a) the narrowness of the ambiguity band of the item (i.e., how specifically localized is the perceived position of the item's boundary on the attribute continuum over time by the average individual) and (b) the frequency density at the item's boundary point. That is, there should exist two kinds of "stable" and two kinds of "unstable" items: an item can be stable because it is highly specific and precisely localized (narrow ambiguity band) and/or an item can be stable because it is extreme and few subjects are found near the

\(^4\) Obviously, one could hypothesize an "ambiguity band" for an individual's perceived position on an attribute continuum, and then response stability could be seen as a joint function of two interacting ambiguity bands. Moreover, items and persons might also be conceptualized as points in multi-dimensional space, rather than as points on a single underlying attribute continuum. The rationales for these more complex models have not yet been completed.
item's boundary. Conversely, an item may be unstable because it has a broad ambiguity band (i.e., it is perceived as being in many spots on the trait continuum by the average person on different occasions) and/or because it lies in the middle of the attribute continuum (where there are many individuals clustered).

Figure 3 is an attempt to depict these notions graphically for three items, "U," "V," and "W," located on the same attribute continuum. Note that each item is conceived as having an ambiguity band, corresponding to its perceived position on the attribute continuum by the average person on different occasions. Item U is graphed as having a narrow ambiguity band; that is, it is perceived on most occasions as being in about the same position on the attribute continuum. Item W is graphed as having a much broader ambiguity band and as being a more extreme item. Item V is graphed as also being relatively extreme but having an ambiguity band similar to item U. Note that most persons would respond "True" to item V and "False" to item W, and that they could make these choices rather easily.

The shaded areas above each item represent the proportion of individuals who would have difficulty making a decision for the
particular item. These are the persons most expected to change their responses upon repeated administrations of the item. As can be seen from Figure 3, although item W has a wide ambiguity band, owing to its extreme position on the attribute continuum it elicits about the same percentage change as item U. Item V, with the same width ambiguity band as item U, elicits the least percentage change.

In actual practice, the percentage of individuals who change their responses to a given item can be readily determined by test-retest procedures, as can the item's endorsement frequency. The remaining problem is to ascertain the width of the item's ambiguity band (a distance along the assumed attribute continuum), given these two indices.

Derivation of an Index of Ambiguity

The preceding discussion provides the rationale behind the index of item ambiguity. For every two administrations of any item pool, two indices are immediately available for each item: (a) a measure of item endorsement frequency (actually two such measures are available, one for each administration of the item; the average endorsement frequency over both administrations of the item is the index presently utilized\(^5\)), and (b) a measure of item instability (defined as the percentage of subjects in the population who change their responses upon a second presentation of the item). The problem is to derive an index of ambiguity, based upon item instability, which corrects for item endorsement frequency. Only the monotonic case will be treated here.

Figure 4 graphs the geometric rationale behind this ambiguity index, called, for short, Amtdex.

For Figure 4:

\[ E \] = Endorsement frequency (shaded area above)

\[ A \] = Amtdex (index of ambiguity; width of ambiguity band)

\[ X_n \] = Scale value (\(E^{th}\) percentile) on the attribute continuum which cuts from the normal curve an area equal to the average endorsement frequency of the item

\[ f(X_n) \] = Ordinate of the normal curve erected at \(X_n\)

---

\(^5\) The geometric mean is being considered as an alternative to the arithmetic mean utilized in the preliminary index.
$X_1 \epsilon X_2$ = Scale values on the attribute continuum such that $X_2 - X_1 = A$

$\beta = \frac{1}{2}A = X_2 - X_E = X_E - X_1$

$I$ = Average percentage of change (index of instability) taken as equal to the area under the normal curve between $X_E - \beta$ and $X_E + \beta$

The index of instability ($I$) for any item is represented by that portion of the normal curve between $X_E - \beta$ and $X_E + \beta$ and as here defined is equal to the percentage of persons changing their responses to the item upon its second presentation. $X_E$ and $f(X_E)$ can be readily found from normal curve tables entered for the area cut off the normal curve equal to the item’s endorsement frequency, averaged over both administrations of the item. The problem is to solve for the distance, $A$, between points $X_1$ and $X_2$, each equidistant from $X_E$.

The solution involves integrating the equation for the normal curve between $-\infty$ and $X_E + \beta$ and subtracting the integral between $-\infty$ and $X_E - \beta$, setting this value equal to $I$ and solving for $\beta$. Although no formal solution exists, $\beta$ can be readily found by interpolation from the table of values of the unit normal curve.
In practice, a simpler approximation is merely to define the area of the rectangle:

\[ I = A \cdot f(X_e) \]

as equivalent to the associated area under the normal curve. Then

\[ A = \frac{I}{f(X_e)}. \]

This approximation will slightly underestimate Ambdex when the endorsement frequency is 50 per cent, and it will slightly overestimate Ambdex for extreme endorsement frequencies. It will be most accurate when \( X_e \) falls directly under the curve’s inflection points.

Table 3 lists some representative values of Ambdex for various item endorsement frequencies and various percentages of subjects changing their responses.

Rust (1961) reported that Ambdex was highly related to an estimate of \( r \)-tetrachoric in a population of 169 male undergraduates tested on a 160-item adjective check list at the end of their freshman and senior years at Yale. N. Wiggins (1961) has shown that Ambdex and \( r \)-tetrachoric are mathematically related, though not identical, statistics.

### Table 3

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Ambdex as a Function of Endorsement Frequency and Response Instability</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00 .01</td>
<td>.00 .03 .05 .07 .09 .11 .13 .15 .17 .19 .21 .23 .25 .27 .29 .31 .33 .35</td>
</tr>
<tr>
<td>.02 .04</td>
<td>.04 .06 .08 .10 .12 .14 .16 .18 .20 .22 .24 .26 .28 .30 .32 .34 .36 .38</td>
</tr>
<tr>
<td>.05 .07</td>
<td>.07 .09 .11 .13 .15 .17 .19 .21 .23 .25 .27 .29 .31 .33 .35 .37 .39 .41</td>
</tr>
<tr>
<td>.08 .10</td>
<td>.09 .11 .13 .15 .17 .19 .21 .23 .25 .27 .29 .31 .33 .35 .37 .39 .41 .43</td>
</tr>
<tr>
<td>.11 .13</td>
<td>.12 .14 .16 .18 .20 .22 .24 .26 .28 .30 .32 .34 .36 .38 .40 .42 .44 .46</td>
</tr>
<tr>
<td>.14 .16</td>
<td>.15 .17 .19 .21 .23 .25 .27 .29 .31 .33 .35 .37 .39 .41 .43 .45 .47 .49</td>
</tr>
<tr>
<td>.17 .19</td>
<td>.18 .20 .22 .24 .26 .28 .30 .32 .34 .36 .38 .40 .42 .44 .46 .48 .50 .52</td>
</tr>
<tr>
<td>.20 .22</td>
<td>.21 .23 .25 .27 .29 .31 .33 .35 .37 .39 .41 .43 .45 .47 .49 .51 .53 .55</td>
</tr>
<tr>
<td>.23 .25</td>
<td>.24 .26 .28 .30 .32 .34 .36 .38 .40 .42 .44 .46 .48 .50 .52 .54 .56 .58</td>
</tr>
<tr>
<td>.25 .27</td>
<td>.26 .28 .30 .32 .34 .36 .38 .40 .42 .44 .46 .48 .50 .52 .54 .56 .58 .60</td>
</tr>
<tr>
<td>.27 .29</td>
<td>.28 .30 .32 .34 .36 .38 .40 .42 .44 .46 .48 .50 .52 .54 .56 .58 .60 .62</td>
</tr>
<tr>
<td>.29 .31</td>
<td>.30 .32 .34 .36 .38 .40 .42 .44 .46 .48 .50 .52 .54 .56 .58 .60 .62 .64</td>
</tr>
<tr>
<td>.31 .33</td>
<td>.32 .34 .36 .38 .40 .42 .44 .46 .48 .50 .52 .54 .56 .58 .60 .62 .64 .66</td>
</tr>
<tr>
<td>.33 .35</td>
<td>.34 .36 .38 .40 .42 .44 .46 .48 .50 .52 .54 .56 .58 .60 .62 .64 .66 .68</td>
</tr>
<tr>
<td>.35 .37</td>
<td>.36 .38 .40 .42 .44 .46 .48 .50 .52 .54 .56 .58 .60 .62 .64 .66 .68 .70</td>
</tr>
</tbody>
</table>

Percentage of Subjects Changing Response on Second Administration
Figure 5. Index of ambiguity vs. item endorsement frequency for 82 males administered 160 adjectives.

Since Ambdex was constructed specifically to provide an index of ambiguity which would be independent of item endorsement frequency, a first test of the proposed statistic was to ascertain its relationship with item endorsement frequency for various subject groups and various item pools. Figure 5 shows one scatterplot illustrative of the lack of relationship between endorsement frequency and Ambdex (eta = .01). The data plotted in Figure 5 come from a double administration of an 160-adjective check list (three week test-retest interval) to 82 University of Oregon male undergraduates from an Introductory Psychology course. Note that the attempt to detach ambiguity from item endorsement frequency appears to have been successful for this sample. Rust (1961) found both Ambdex and \( r \)-tetrachoric unrelated to item endorsement frequency in his sample.

The Use of Ambdex for Item Selection

Personality test development has been marked by diligent interest in the principles of scale construction and marked neglect of the properties of the stimuli from which the scales are normally con-
structured. It may well be that our most powerful strategies for item grouping are never able to overcome the defects present in the initial item sample. Therefore, our best hope for more valid assessment tools may well lie in the increased understanding of basic item properties.

In the history of test development, item construction has typically been considered a relatively inconsequential task. Although the direct investigation of personality test item properties has a long history (e.g., Bain, 1931; Smith, 1933; Lentz, 1934; Benton, 1935; Pintner & Forlano, 1938), early investigators—while spotting some of the more salient problems in item construction and interpretation—soon reached impasses that discouraged further explorations. Meanwhile, test developers were relying far too heavily on rationally-constructed items, building face-valid scales and assuming their relevance to the measurement task at hand. The wave of pure empiricism which swept American assessment circles in the decade following World War II was a healthy force in placing the discipline of personality assessment on firmer methodological grounds. During this period there emerged explicit rationales for minimizing the importance of such item properties as item structure and content. One of the earliest of these rationales was presented by Paul Meehl in his influential defense of pure empiricism in inventory construction (Meehl, 1945); a more extreme position with respect to item content has been taken by Irwin Berg (1959).

The rather basic property of item stability (response consistency over repeated administrations of the item) seems not to have been exempted from the general neglect of item characteristics, and the reliability of the basic units of personality measurement has not received the systematic attention it deserves. Were one to inquire as to how to insure a given degree of reliability in a personality scale (other than simply varying the length of the test), one could not find a complete answer in the literature.

As was pointed out earlier, investigators in the 1930's were interested in improving the stability of their scales through an examination of item properties, but they encountered an apparent enigma—namely, that the most discriminating items (e.g., between neurotics and normals) were the very items which were responded to most inconsistently in the general population. That is, they found that the least stable items were the most useful predictively. In retrospect, the enigma is not difficult to understand. The items which
were most useful were balanced items, and the items which were most stable were extreme items. Thus, early investigators (e.g., Bain, 1931; Benton, 1935; Frank, 1936; Hertzman & Gould, 1939; Lentz, 1934) were faced with the choice of having stable scales (composed of extreme items) which were non-differentiating, or scales composed of items with some predictive efficiency but low temporal stability.

This paradox may arise primarily from the strong relationship between inter-individual and intra-individual variability. Isard (1956) has found that the most discriminating forced-choice items in predicting college achievement are those whose alternatives elicit the most inter-individual variability in their preference ratings. Since inter-individual response variability is essential for personality assessment, it may be necessary to purchase inter-individual variability in stimulus “meaning” (equivocality) to achieve this end.

In a very significant paper, Broen (1960) specified the conditions under which changes in equivocality should lead to increased item discriminating power. Normally the test constructor seeks to (a) maximize inter-individual response variability (often, as Broen notes, by sanctioning equivocality) and simultaneously (b) minimize intra-individual variability of response (and consequently of stimulus meaning—i.e., ambiguity). However, since items of high equivocality are typically items of high ambiguity, the paradox is complete. To assess stable components of personality, it would appear necessary either to (a) use the average values of repeated measures (an expensive and time-consuming process) or (b) develop an item technology to the point where items can be constructed which elicit both high inter-individual variability and low intra-individual variability.

The search for both stable and discriminating items leads to another enigma, namely, that typically the more an item reflects some aspect of behavior that is directly observable and easily identifiable the more stable is the item; on the other hand, the more an item reflects some attitude, value, or other “internal” state of mind the more unstable is the item. Consequently, at present, inventory constructors could be faced with the dilemma of choosing stable items reflecting attributes of relatively little personological importance or selecting items of decreased triviality but increased
intra-individual variability. For this reason, the investigation of stable biographical inventory questions (e.g., Rust, 1961; Owens, Glennon & Albright, 1962) for utilization in personality inventories is certainly worth continued effort.

Another approach lies in the general investigation of item properties and the development of objective indices to guide initial item selection (prior to actual scale construction). Ambdex is suggested as a possible index to be used to minimize item ambiguity. It may also be useful to develop an objective index of equivocality (utilizing the dispersions of item boundary placements inter-individually) to provide a parallel measure with Ambdex. As Broen (1960) has illustrated, in some prediction situations the test constructor may wish to minimize both equivocality and ambiguity, thus loading his initial item pool with low Ambdex and low equivocality items. For other situations, however, in which increased equivocality is useful, one might select for an initial item pool those items having high equivocality/Ambdex ratios.

With the exception of Horn (1950), the psychometric literature discloses no opponents to decreasing intra-individual variability, at least until the hypothesized stability of the personality trait itself is attained. Unfortunately, there is less agreement on the extent of personality trait variability (e.g., Cattell, 1957; Fiske, 1961; Secord & Backman, 1961). A general review of the literature on variability can be found in Fiske and Rice (1955). Some important empirical findings on the stability of diverse traits over long intervals of time (as measured by early psychometric scales) are reported by Kelly (1956).

Although there is general concordance on the importance of intra-individual response stability in psychometric assessment, the traditional neglect of item properties has turned research attention away from the systematic investigation of item stability. With one very notable exception (Bills, Vance & McLean, 1951) no major personality inventory has been constructed by explicitly utilizing item stability as a criterion for item selection. There are, however, recent indications of a renewed interest in the stimulus characteristics of items (Owens, Glennon & Albright, 1962; Hanley, 1962; Edwards & Walsh, 1963) and a trend toward analysis of item parameters associated with specific response patterns. The breadth of explanatory power generated by scaling the single item parameter
of social desirability has been repeatedly documented by Edwards (1957). The distinction between content and style (Jackson & Messick, 1958) and the formulation of item properties as mediators of components of strategic, method, and stylistic variance in assessment (Wiggins, 1962) are significant steps forward. These recent theoretical contributions, taken together with the important demonstrations that systematic variation of item properties results in systematic variation in response patterns (Buss & Durkee, 1957; Buss, 1959; Hanley, 1959; Stricker, 1960; Eliott, 1961; Goldfried & McKenzie, 1962; Aiken, 1962), suggest that personality test item construction is progressing from an art to a science. The present paper aims to channel a part of this trend to the area of item stability.

**Summary of the Ambiguity Model**

The preceding discussion has focused exclusively upon the practical uses for an ambiguity index. However, considerations of the role of response variability and stimulus ambiguity in personality assessment also uncover some significant theoretical issues, and the ambiguity model provides the potential opportunity for explaining a large number of relatively diverse empirical findings. Consequently, the model will now be summarized in its present preliminary form, with the hope that it will stimulate independent investigations which in turn may help clarify the ambiguities now present in the ambiguity model. The current model applies solely to monotone items, scored dichotomously; the more general case is being developed. Brief descriptions of the major postulates and theorems of the model follow.

**Postulate 1:** The closer the perceived position of a person is to the boundary established by the item on the relevant attribute continuum (as perceived by the person), the more difficult will be his decision for that item. Some manifestations of difficulty include: (a) response in a middle, or "?," category if this is allowed in the response format (or refusal to respond—leaving the item blank—if a middle category is not provided), (b) latency of response (greater latency indicating greater difficulty), and (c) ratings of item difficulty (or, conversely, ratings of low judgmental confidence). Consequently, the following predictions follow from this postulate:
(P:1a) More items whose boundaries are perceived as close to a subject's position than items perceived as further away will be placed in a middle, or "?" category (if this is allowed in the response format) or will be left blank if a middle category is not provided.

(P:1b) The latency of response should be greater for items whose boundaries are perceived as closer to a subject's position than for those perceived as further away.

(P:1c) A person should report more difficulty in responding (and consequently less confidence in his response) to items whose boundaries are perceived as closer to himself than to those perceived as further away.

(P:1d) The three manifestations of difficulty should be interrelated.

Postulate 2: The closer the perceived position of a person is to the boundary established by the item on the relevant attribute continuum (as perceived by the person), the more unstable will be his response.

Theorem 1: Combining Postulates (1) and (2):
The more difficult is a person's response to an item upon its first administration, the more likely is he to change that response upon retesting. Consequently:

(T:1a) Items placed in a "?" category (or left blank) will be relatively unstable.

(T:1b) Items which elicit long response latencies will be relatively unstable.

(T:1c) Items rated difficult (or given low confidence ratings) will be relatively unstable.

Postulate 3: The distribution of subjects on the attribute continuum is specified. In practice, subjects are assumed to be distributed in approximately normal curve frequencies on most psychological attribute continua.

Theorem 2: From Postulate (3):
Fewer persons will see their own positions as falling close to an extreme item's boundary than to a more balanced item.

Theorem 3: Combining Postulate (1) and Theorem (2):
Items of extreme endorsement frequencies, as compared to more balanced items, will be seen by fewer individuals as difficult items. Consequently:
(T:3a) Balanced items will be placed in a "?" category, or will be left blank, by more individuals than will extreme items.

(T:3b) Balanced items will elicit longer response latencies.

(T:3c) Balanced items will be judged more difficult (or given lower confidence ratings).

Theorem 4: Combining Postulate (2) and Theorem (3):

Items of extreme endorsement frequencies, as compared to more balanced items, will be relatively stable.

Postulate 4: The greater the inter-individual variability in the positioning of an item’s boundary on a relevant attribute continuum (i.e., the more disagreement among persons as to the position of the boundary of an item on the attribute continuum), the greater will be the intra-individual variability of item positioning over time (i.e., the more likely will individuals be to vary in their positioning of the item’s boundary upon repeated administrations of the item). That is, equivocality is related to ambiguity. Since manifestations of intra-individual variability in boundary positioning should include both subjective ratings of ambiguity as well as an objective index of ambiguity (Ambdex), consequently:

(P:4a) Intra-individual variability in item boundary positioning will be directly related to ratings of item ambiguity.

(P:4b) Intra-individual variability in item boundary positioning will be directly related to Ambdex values.

(P:4c) Ratings of item ambiguity will be related to Ambdex values.

(P:4d) Ratings of item ambiguity will be correlated with item equivocality.

(P:4e) Ambdex values will be correlated with item equivocality.

Postulate 5: The more intra-individual variability in the positioning of an item’s boundary, the more unstable will be the response to that item.

(P:5a) Response instability will be correlated with ambiguity ratings.

(P:5b) Response instability will be correlated with Ambdex values.

Theorem 5: Combining Postulate (4) and Postulate (5):

The more inter-individual variability in the positioning of an item’s boundary, the more unstable will be the response to that item.
In summary, response stability can be considered as a positive function of (a) self-item distance (for an individual) and (b) item extremeness (for a group), and a negative function of (c) item ambiguity, (d) item difficulty, and (e) item equivocality.

Table 4 summarizes the postulates and theorems of the model and indicates those which have received empirical test.

Evidence for some of these assumptions can be found in a recent study by Hanley (1962). Using 75 MMPI items, Hanley demonstrated that a measure of response latency and a measure of judgmental confidence were both negatively correlated with extremeness of item endorsement frequency (i.e., latency as well as difficulty were inversely related to item balance). Moreover, Hanley showed that item length, which Strong (1962) found to be related to ambiguity in SVIB items, was also highly related to response latency and difficulty.

A very recent study by Edwards and Walsh (1963) provides additional support for the model. When 176 miscellaneous personality statements were administered twice to 110 male and 111 female college students, item stability was found to be related to the extremeness of item positioning on the social desirability dimension (as independently assessed by ratings from 47 male and 48 female college students) and item stability was negatively related to the variance of these social desirability ratings. Moreover, this study replicated that of Dodd and Svalastoga (1952) in demonstrating that item stability was inversely related to the percentage of subjects utilizing a middle, or "?", category when this response option is permitted. In Dodd and Svalastoga's study, 271 respondents in a poll on national affairs replied by mail to a second poll in which seven questions from the original poll were embedded. For these seven repeated questions, a correlation of .91 was found between the percentage of "don't know" responses on the first poll and the percentage of responses changed between poll and re-poll. Another demonstration of the same theorem has been provided in a study in which consistency of response to an 11-item attitude scale was shown to be inversely related to response latency (Crockett, Bates & Caylor, 1958).

Postulate (4) of the ambiguity model states that inter-individual variation in the positioning of an item's boundary should be directly related to the average intra-individual variation of boundary posi-
TABLE 4
Postulates and Theorems of the Ambiguity Model

<table>
<thead>
<tr>
<th>Postulate (P) or Theorem (T)</th>
<th>Individual Case</th>
<th>Group Case</th>
<th>Empirical Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>P:1</td>
<td>$D_{ij} = f(d_{ij})$</td>
<td>$D_{.j} = f(d_{.j})$</td>
<td>Dodd &amp; Svalastoga (1952); Edwards &amp; Walsh (1963)</td>
</tr>
<tr>
<td>P:1a</td>
<td></td>
<td>$O_{.j} = f(d_{.j})$</td>
<td></td>
</tr>
<tr>
<td>P:1b</td>
<td>$L_{ij} = f(d_{ij})$</td>
<td>$L_{.j} = f(d_{.j})$</td>
<td></td>
</tr>
<tr>
<td>P:1c</td>
<td>$R_{ij} = f(d_{ij})$</td>
<td>$R_{.j} = f(d_{.j})$</td>
<td></td>
</tr>
<tr>
<td>P:1d</td>
<td></td>
<td>$O_{.j} = f(L_{.j})$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$L_{ij} = f(R_{ij})$</td>
<td>$L_{.j} = f(R_{.j})$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$O_{.j} = f(R_{.j})$</td>
<td></td>
</tr>
<tr>
<td>P:2</td>
<td>$I_{i.} = f(d_{.i})$</td>
<td>$I_{.i} = f(d_{.i})$</td>
<td></td>
</tr>
<tr>
<td>T:1</td>
<td>$I_{i.} = f(D_{i.})$</td>
<td>$I_{.i} = f(D_{.i})$</td>
<td></td>
</tr>
<tr>
<td>T:1a</td>
<td>$I_{i.} = f(O_{i.})$</td>
<td>$I_{.i} = f(O_{.i})$</td>
<td>Dodd &amp; Svalastoga (1952); Edwards &amp; Walsh (1963)</td>
</tr>
<tr>
<td>T:1b</td>
<td>$I_{i.} = f(L_{i.})$</td>
<td>$I_{.i} = f(L_{.i})$</td>
<td>Crockett, Bates, &amp; Caylor (1958)</td>
</tr>
<tr>
<td>T:1c</td>
<td>$I_{i.} = f(R_{i.})$</td>
<td>$I_{.i} = f(R_{.i})$</td>
<td></td>
</tr>
</tbody>
</table>

P:3 Specification of distribution of individuals on underlying attribute continuum (here assumed to be the normal distribution)

T:2 $d_{.i} = f([.5 - p].)_{.i}$
T:3 $D_{.i} = f([.5 - p].)_{.i}$
  $O_{.i} = f([.5 - p].)_{.i}$
T:3a $L_{.i} = f([.5 - p].)_{.i}$ Hanley (1962)
T:3b $R_{.i} = f([.5 - p].)_{.i}$ Hanley (1962)
T:4 $I_{.i} = f([.5 - p].)_{.i}$ Edwards & Walsh (1963)

P:4 $S_{i.} = f(S_{.i})$ Goldberg (present report)

P:4a $a_{.i} = f(S_{.i})$
P:4b $A_{.i} = f(S_{.i})$
P:4c $a_{..} = f(A_{..})$
P:4d $a_{..} = f(S_{..})$
P:4e $A_{..} = f(S_{..})$

P:5 $I_{.i} = f(S_{.i})$ Edwards & Walsh (1963)

P:5a $I_{.i} = f(a_{..})$
P:5b $I_{.i} = f(A_{..})$
T:5 $I_{.i} = f(S_{..})$

Notation: $i =$ individuals; $j =$ occasions; $k =$ items (omitted from the notation)

- $D =$ item difficulty
- $d =$ perceived self-item distance
- $p =$ endorsement frequency (proportion); $[.5 - p] =$ item balance
- $I =$ response instability (proportion of responses changed)
- $S =$ dispersion of item boundary positions
- $L =$ response latency
- $O =$ percentage of individuals omitting item (or placing item in "I" category)
- $R =$ "difficulty" ratings
- $a =$ "ambiguity" ratings
- $A =$ Ambdex values

[a bar over a symbol indicates a mean value taken over the indicated subscript]

tioning over repeated administrations of the item. Evidence supporting this assumption can be seen in Figure 6.

The data plotted in Figure 6 come from a study in which 160
adjectives were administered on ten occasions (approximately one-week interval between administrations) to ten University of Oregon undergraduates; the task of the subjects was to position each adjective on the underlying attribute continuum of "Social Desirability," defined for them in approximately the same manner as that used by Edwards (1957). The correlation between variation in item positioning inter-individually and variation in item positioning intra-individually (over time) was .52.

Summary

A major stumbling block to improved personality scales may arise from defects in the initial item pool utilized for scale construction. One of the most important defects of both early and recent personality test items may be their ambiguity, here considered as a property of the item which tends to elicit marked intra-individual variability in response over repeated test administrations. Conventional measures of item instability (i.e., the percentage of subjects changing their responses to an item upon its repeated administra-
tion) have been shown to be closely related to item endorsement frequency and thus can be shown to have rather severe limitations as indices of ambiguity. This paper attempts to integrate existing knowledge concerning item instability, and a model of item ambiguity and response instability is presented. An item statistic, Ambdex, derived from the model is proposed as a preliminary index of item ambiguity, and its use as an item selection statistic is discussed.

REFERENCES


Dodd, S. C. and Svalastoga, K. “On Estimating Latent from Mani-
fest Undecidenss: The ‘Don’t Know’ Percent as a Warning of Instability among the Knowers.” EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XII (1952), 467–471.


Pintner, R. and Forlano, G. "Four Retests of a Personality Inventory." *Journal of Educational Psychology, XXIX* (1938), 93–100.


