Tailored Interviewing: An Application of Item Response Theory for Personality Measurement

Wagner A. Kamakura  
*Owen Graduate School of Management*  
*Vanderbilt University*

Siva K. Balasubramanian  
*College of Business Administration*  
*University of Iowa*

Despite recent technological advances in mass survey administration, surveys seeking to measure personality traits still employ procedures identical to the traditional paper-and-pencil scales, in which every respondent is asked exactly the same items, regardless of his or her trait level.

We present an empirical application of personality measurement in which the number and sequence of scale items are tailored to the trait level of each respondent and show that the total number of questions asked of each respondent could be reduced substantially, with a measurable and controllable increase in the standard error of measurement.

Contemporary surveys seeking to measure respondent attitudes or personality traits of interest invariably administer a scale (e.g., the Anomia scale used in the annual National Opinion Research Council surveys) using a paper-and-pencil analog in which every respondent has to answer all the scale items. Clearly, this approach fails to take advantage of a priori differences in attitude or trait levels among respondents; therefore, at least some items are redundant for each respondent. Consider, for instance, the following items in a scale to measure attitudes toward abortion with dichotomous (yes/no) response options:

1. Abortion is a terrible sin.
2. Abortion could be legalized if there is an acceptable definition of when life begins.
If the respondent answered "yes" to Item 1 (high negative attitude), there appears to be little value in bothering him or her to respond to Item 2 (moderately positive attitude) as we can clearly predict a "no" response here. In other words, the incremental information provided by Item 2 for this respondent is insignificant; more useful information regarding the respondent's attitude could be obtained by asking an item tapping the highly negative attitude range. For any given multi-item scale, items that lack in informativeness vary across respondents, thus rendering such a fixed scale inefficient; the unproductive increase in survey time caused by uninformative or marginally informative items jeopardize effective survey administration for at least two reasons: (a) increase in survey time and cost, and (b) decrease in respondent motivation to participate in the survey.

The moral of the attitude-measurement example just cited also holds in a trait-measurement context. To avoid items lacking in significant information, the respondent must ideally answer only those items which would be most informative at his or her trait level. Evidently, tailoring scale items to individual trait levels in a survey is feasible only in an interactive environment, where the likelihood of an item being included in the interview depends on the responses to items asked before. This article uses Item Response Theory (IRT) to demonstrate the advantages of interactive computer-based measurement over traditional paper-and-pencil methods. Specifically, we employ a tailored interview procedure which "tailors" the number and sequence of items in the measurement scale according to the trait level of each respondent; further, we show that this procedure greatly decreases the number of items asked of each respondent without a substantial increase in measurement error. Although the idea of tailoring the items in a measurement instrument to the particular individual measured has been studied extensively in educational applications (Green, 1983; Urry, 1977; Weiss, 1982), an important contribution of this article lies in extending this method for the first time to the personality measurement domain.

**TAILORED INTERVIEWING**

State-of-the-art survey tools, such as CRT interviewing (Tyebjee, 1979) and self-interviewing (Sexton, 1986), still employ procedures identical to the traditional paper-and-pencil scale administration, in which every respondent is asked all the items in the measurement scale, regardless of his or her trait level. Further, this approach is also adopted in research on the variation in personality scores engendered by differences in modes of scale administration, such as paper-and-pencil testing versus computer-based testing of the same measurement scale (Koson, Kitchin, Kochen, & Stodolsky, 1970; Rezovik, 1977). This research tradition follows from classical measurement theory, wherein the final trait measure is a linear combination of the observed responses to all the
scale items. Furthermore, comparability among respondents across the trait measure is conditional on their having answered identical sets of items.

In sharp contrast, IRT (on which the tailored interview procedure is based) is not bound by such restrictions. First, this theory dispenses with the need to obtain responses to all the scale items; instead, it recognizes that different items are most informative at different ranges in the trait measure. Hence, if a preliminary estimate of the respondent's trait level can be obtained, then he or she need be asked only those items that would prove to be most informative for this trait level. This estimate is viable because of two important IRT properties: (a) the sample free calibration of item parameters, which allows the development of standard item banks based on any sample of respondents; and (b) the item free measurement property that permits comparison of trait measures across two or more respondents despite their having responded to different sets of items on the same scale. Empirical evidence on these two IRT properties can be found in Wright and Panchapakesan (1969), Wright and Douglas (1977a, 1977b) and Koch (1983).

An important requirement for this adaptive measurement procedure is the availability of parameter estimates\(^\dagger\) for all the items in the measurement scale to be used, namely, an item bank that has already been calibrated on a previous sample. Procedures for the estimation of item parameters are discussed in detail elsewhere (Bock, 1972; Lord, 1974; McKinley & Reckase, 1980; Wright & Douglas, 1977b), and computer programs for this purpose are widely available (LOGLOG: Kolakowski & Bock, 1973; ANCILLES: Urry, 1978; LOGIST: Wingersky, Barton, & Lord, 1982; among others).

Given an item bank with known item parameters, the interview can be tailored to each respondent by first asking a few initial items, estimating the respondent's trait level based on these initial items, drawing the item most informative at that particular trait level, and repeating the process until a given stopping rule is reached.

Lord (1971a, 1971b) discussed several fixed and shrinking step-size procedures that determine the increment (or decrement) in threshold of the item to be asked, after a given response to a previous item. His results (1971b), however, indicate that the procedures depend markedly on the choice of the first item and on the choice of the step size (or of the initial step size for the shrinking-step procedure).

A different approach was suggested by Owen (1969, 1975), using a Bayesian procedure that updates the trait estimate and its standard deviation. In this

\(^\dagger\)Two types of item parameters are required for each item: (a) item-threshold parameter and (b) item-discrimination parameter (see the equation for Birnbaum's, 1966, 2-parameter logistic model in the Appendix). Ideally, the item bank should contain item-threshold values evenly distributed in the \(-2.0\) to \(+2.0\) range, and the item-discrimination parameter values should preferably be above \(.8\) (Urry, 1977).
approach, initial estimates of the trait ($\theta$) and its standard deviation in the population of individuals ($\sigma$) are set to 0 and 1 respectively, assuming a standardized normal prior sampling distribution. Using this prior distribution, Owen developed updating formulas for the trait estimate and its standard deviation (see Jensen, 1974, pp. 758–760). A criterion for the selection of items based on Owen's Bayesian procedure is also elaborated in Jensen (1974), which determines the iterative item selection such that it minimizes the posterior standard error of the trait estimates. Jensen (1974) identified two stopping rules for the tailored procedure: (a) stopping when a minimum standard error of measurement is reached for each respondent, or (b) stopping when a maximum number of items are answered by each respondent.

If the first stopping rule (minimum standard error) is used, a constant standard error of measurement is attained across all respondents. This common standard error of the trait estimates is related to the correlation between the trait measures and their true values as follows:

$$R(\theta, \dot{\theta}) = (1 - \sigma^2)^{1/2} \quad (1)$$

where $\sigma$ is the error variance.

Although this correlation (the reliability of the measurement instrument) plays an important role in traditional measurement approaches, the standard error of measurement estimated by latent trait models provides a more meaningful assessment of measurement error. Although the reliability coefficient refers to a constant (or average) error of measurement over all trait levels, the standard error provided by latent trait models estimates the measurement error at the particular trait level (see Appendix).

The relative efficiency of Owen's procedure vis-à-vis traditional standardized paper-and-pencil tests has been recognized by researchers as being valid on theoretical (Green, 1983; Jensen, 1974) and empirical grounds (Ury, 1974, 1977; Weiss, 1982; Weiss & Kingsbury, 1984). These studies highlight the following advantages of using this Bayesian procedure: (a) the reliability of the trait measures can be controlled, and (b) the number of items required for reliable measurement can be substantially reduced. On the other hand, a drawback of Bayesian trait estimates is that they tend to be regressed toward prior estimates; this may lead to biases, especially when the stopping rule results in a small set of items (Weiss & Kingsbury, 1984). Furthermore, Owen's Bayesian updating formulas are specific to the dichotomous normal ogive model (Lord & Novick, 1968), while logistic latent trait models are most commonly used for item parameter estimation.

The tailored interviewing algorithm to be used in the next section overcomes these limitations by using a maximum likelihood procedure for the trait estimation at each stage of the interviewing process. Maximum-likelihood estimation (MLE) of logistic models is known to provide asymptotically unbiased and
efficient results, making the final trait estimates less sensitive to the prior estimates as in the Bayesian procedures. Furthermore, algorithms are already available in the literature for the MLE of dichotomous (Birnbaum, 1968; Wood, Wingersky, & Lord, 1976) and multichotomous-response models (Bock, 1972; Samejima, 1969). Details on the MLE of traits based on Birnbaum's (1968) 2-parameter logistic model are presented in the Appendix.

Because the MLE procedure is asymptotically unbiased, the interviewing algorithm will result in asymptotically unbiased trait estimates regardless of the sequence of questions asked, unless all items are answered positively (or all negatively) in every step of the process. In such cases, the MLE procedure will assign extreme trait estimates to the respondent. The same is true, however, for most other measurement techniques; if the respondent answered all items in the same extreme fashion, we can only assume that this respondent falls in the upper/lower bounds of the trait range covered by the scale.

In summary, the implementation of tailored interviewing as a procedure for trait measurement would involve: (a) development of an item bank through the estimation of item parameters from a calibration sample, (b) choice of a trait estimation procedure (e.g., the MLE procedure used in the following section, the Bayesian updating procedure suggested by Owen, 1969, or the difficulty of the last item asked per Lord, 1971b), (c) development of an item selection algorithm (e.g., fixed or shrinking step rule per Lord, 1971b, Jensen's, 1974, posterior standard error index, or the maximum information criterion per Lord, 1980), and (d) choice of a stopping rule to terminate an interview (e.g., minimum level of measurement error acceptable, the maximum number of items to be asked, or some combination of the preceding criteria).

METHOD

Empirically testing the tailored interviewing procedure poses a dilemma with respect to two issues: test realism and test comparability. If the test involved a trait measurement context, it is desirable to collect the empirical data on traits exactly through the steps outlined by the tailored procedure in order to add realism to the test. However, because the primary objective is to compare the usefulness of the tailored procedure vis-à-vis traditional paper-and-pencil methods, it is necessary to orient data collection in a manner that facilitates such a comparison. The traditional method of measurement is possible only if each respondent answers all the scale items; the data base used in our empirical test meets this requirement.

For this empirical test, we used the true–false responses to a 54-item scale from a national survey of 2,500 high school students. This scale is part of the California Psychological Inventory (which includes 18 constructs measured by 480 items) and measures socialization, which indicates the "degree of social
maturity, integrity and rectitude which the individual has attained" (Gough, 1975, p. 10). Relative to the other CPI scales, this scale has been more widely studied. For instance, 26,824 subjects were tested in 10 countries in a single study of this scale (Megargee, 1972); several studies have measured this trait as a unidimensional construct and established the ability of this scale to differentiate groups characterized by different levels of socialization. In addition, the reliability and validity of this scale has been tested and confirmed for more than 24 samples with diverse characteristics, covering more than 15,000 subjects (Gough, 1975).

The tailored interview process was implemented through a computer simulation procedure which can "interview" each respondent such that only the responses to the items selected by the interview algorithm were drawn from each respondent's database. This simulation accurately depicted the sequence of operations in a real life tailored interview. One may argue that the simulation could be biased, because the respondents, in fact, had been asked other related items not used by the tailored interview. However, this bias was minimized by the fact that the socialization scale items were embedded in a large set of items, in a random order.

A calibration sample was first chosen to develop the item bank. Our calibration sample consisted of a random subsample of 1,000 students from the database. Their responses to the 54 items were used to estimate the item parameters for Birnbaum's (1968) 2-parameter logistic model. In a preliminary estimation attempt, 10 of the 54 items were deleted because of their extremely low item discrimination. The final item bank thus consisted of the 44 items with the parameter estimates listed in Table 1.

Table 1 shows that many of the selected items have discrimination levels below the minimum suggested by Urry (1977). These results are understandable, because the items are not measuring abilities in a testing context (as in educational applications) wherein the answers to an item can be clearly segregated into either correct or incorrect categories; instead, their purpose is to capture responses germane to the socialization trait, for which the distinction between true and false responses is, relatively speaking, less clearly defined.

---

2The 10 deleted items also had the lowest item-total correlations and did not load on the main factor of a clearly unidimensional factor solution. Nevertheless, it is possible that they might be tapping some other aspect of the socialization construct, decreasing the validity of the scale. This hypothesis was tested via confirmatory factor analysis. We compared a unidimensional factor model with a two-factor structure in which the 10 deleted items measured a second latent variable. The LISREL results support the unidimensional assumption (AGFI = 0.57, RMSR = 0.145) over the two-factor model (AGFI = 0.42, RMSR = 0.171). Although these results support the unidimensional assumption over the hypothesis that the deleted items measure a second dimension of the socialization construct, they do not rule out the possibility that the deleted items represent more than one additional dimension of the construct. Unfortunately, testing this more complex structure would require more specific assumptions regarding the relationship between the deleted items and the additional latent variables.
<table>
<thead>
<tr>
<th>Item #</th>
<th>$a_i$</th>
<th>$b_i$</th>
<th>$X^2$ (5 Degrees of Freedom)</th>
<th>Item #</th>
<th>$a_i$</th>
<th>$b_i$</th>
<th>$X^2$ (5 Degrees of Freedom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.51778</td>
<td>-2.32328</td>
<td>11.8</td>
<td>32</td>
<td>1.61609</td>
<td>-0.04692</td>
<td>36.7*</td>
</tr>
<tr>
<td>2</td>
<td>0.98340</td>
<td>-3.84417</td>
<td>2.7</td>
<td>33</td>
<td>0.91927</td>
<td>-1.79514</td>
<td>10.7</td>
</tr>
<tr>
<td>4</td>
<td>1.02439</td>
<td>-3.23646</td>
<td>6.8</td>
<td>36</td>
<td>0.20858</td>
<td>-0.10946</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>1.09218</td>
<td>-3.53337</td>
<td>8.3</td>
<td>37</td>
<td>0.36784</td>
<td>-0.24478</td>
<td>7.5</td>
</tr>
<tr>
<td>6</td>
<td>0.63791</td>
<td>0.14004</td>
<td>19.0*</td>
<td>38</td>
<td>1.25565</td>
<td>-3.10646</td>
<td>18.8*</td>
</tr>
<tr>
<td>9</td>
<td>1.18954</td>
<td>-2.77291</td>
<td>4.4</td>
<td>39</td>
<td>0.64405</td>
<td>0.02201</td>
<td>20.7*</td>
</tr>
<tr>
<td>10</td>
<td>0.53300</td>
<td>-0.35519</td>
<td>26.1*</td>
<td>40</td>
<td>0.36228</td>
<td>-0.14482</td>
<td>18.0*</td>
</tr>
<tr>
<td>11</td>
<td>0.67146</td>
<td>-0.20778</td>
<td>9.7</td>
<td>41</td>
<td>1.37310</td>
<td>-0.29941</td>
<td>20.7*</td>
</tr>
<tr>
<td>12</td>
<td>0.78589</td>
<td>-2.07451</td>
<td>8.6</td>
<td>42</td>
<td>0.91923</td>
<td>-4.05825</td>
<td>5.5</td>
</tr>
<tr>
<td>13</td>
<td>0.29853</td>
<td>-0.60510</td>
<td>14.6</td>
<td>43</td>
<td>0.70872</td>
<td>-1.74641</td>
<td>6.2</td>
</tr>
<tr>
<td>14</td>
<td>0.81111</td>
<td>-2.14415</td>
<td>15.8*</td>
<td>44</td>
<td>0.45625</td>
<td>0.14694</td>
<td>13.0</td>
</tr>
<tr>
<td>15</td>
<td>0.11307</td>
<td>0.84011</td>
<td>8.4</td>
<td>45</td>
<td>1.01619</td>
<td>-1.55598</td>
<td>4.9</td>
</tr>
<tr>
<td>16</td>
<td>0.33990</td>
<td>-0.11313</td>
<td>2.2</td>
<td>46</td>
<td>0.85490</td>
<td>-0.97005</td>
<td>4.4</td>
</tr>
<tr>
<td>17</td>
<td>0.82135</td>
<td>-1.85909</td>
<td>14.3</td>
<td>47</td>
<td>0.89011</td>
<td>-1.63302</td>
<td>6.5</td>
</tr>
<tr>
<td>18</td>
<td>1.23033</td>
<td>-2.29962</td>
<td>7.5</td>
<td>48</td>
<td>1.21487</td>
<td>-2.81977</td>
<td>3.1</td>
</tr>
<tr>
<td>19</td>
<td>0.71177</td>
<td>1.27865</td>
<td>11.9</td>
<td>49</td>
<td>0.52321</td>
<td>-1.67885</td>
<td>12.9</td>
</tr>
<tr>
<td>20</td>
<td>1.15761</td>
<td>-1.52316</td>
<td>13.8</td>
<td>50</td>
<td>0.87027</td>
<td>-3.08324</td>
<td>13.4</td>
</tr>
<tr>
<td>21</td>
<td>0.67014</td>
<td>-1.27795</td>
<td>4.8</td>
<td>51</td>
<td>1.26370</td>
<td>-0.32632</td>
<td>11.2</td>
</tr>
<tr>
<td>23</td>
<td>1.44845</td>
<td>-1.42669</td>
<td>12.6</td>
<td>52</td>
<td>1.60846</td>
<td>-1.80344</td>
<td>17.6*</td>
</tr>
<tr>
<td>25</td>
<td>1.15956</td>
<td>-2.12076</td>
<td>13.4</td>
<td>53</td>
<td>0.88647</td>
<td>-3.41716</td>
<td>6.3</td>
</tr>
<tr>
<td>26</td>
<td>0.76896</td>
<td>-2.25243</td>
<td>7.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1.27036</td>
<td>0.20517</td>
<td>13.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.89625</td>
<td>-3.22044</td>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.70811</td>
<td>-2.37455</td>
<td>9.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant at the $p < .05$ level.

Furthermore, the item thresholds among these 44 items are more concentrated in the low threshold range, instead of spreading evenly in the ($-2$ to $+2$) range as desired. In simple terms, most of those items seem to be easy to agree with, even by subjects who score low on the socialization trait. This fact (which was beyond our control), as elaborated later, increases measurement error at the high end of the trait continuum, because there are less items to tap this trait range.

The chi-square statistic listed in Table 1 measures the lack-of-fit between the estimated Item Characteristic Curve and the observed response frequencies at 10 trait levels. For all except 9 items, the hypothesis of fit could not be rejected with a .05 probability of Type I error.

With this item bank developed, the next step was to measure the socialization levels of the remaining 1,500 subjects, using the tailored interviewing procedure. Because the data base already contained the actual responses to all the 44 items for all the 1,500 subjects (from which a "reference measure" of each subject's...
socialization trait could be determined), our research objective centered on comparing these reference measures with the measures of socialization obtained through tailored interview procedures (which use only a fraction of the original set of items). The efficacy of tailored interview procedures will be firmly established if the magnitude of accuracy loss entailed by these, vis-à-vis the reference measures, is: (a) not very significant, and (b) considerably outweighed by the cost economies obtained through the use of a smaller set of items in the interview process.

The steps entailed in the tailored interviewing algorithm are as follows:

1. Three items, respectively representing high, middle, and low threshold levels, are selected a priori; initially, the respondent answers these items only.
2. The responses to the three initial items are used to compute a preliminary MLE of the respondent’s trait \( \theta \), using equations (A3) to (A6) in the Appendix.
3. The information function for each of the unused items is evaluated at the current estimate of \( \theta \), and the item with the highest information value is used.
4. The response to the new item and all previous answers are used to compute a new MLE of the respondent’s trait.
5. steps (3) and (4) are repeated until the stopping rule is reached.

Three stopping rules were tested for the tailored interview algorithm. In the first (MAXIT), the interview proceeded until a maximum of 15 items were asked of the respondent. In the second stopping rule (MINSTD), the interview was stopped when a minimum standard error of \( \sigma = .40 \) (arbitrarily specified) was attained for the subject’s trait estimate. The third stopping rule (MIXED) was a combination of the first two. A minimum of 10 items were asked of each respondent, but the interview was not terminated until the minimum standard error \( \sigma = .4 \) was attained.

As a basis for comparison, trait estimates were also obtained by using all 44 items (ALL) as in a typical scaling procedure.

RESULTS

Our analysis indicates that a switch from the full-scale interview (ALL) to any of the tailored interview procedures (MINSTD, MAXIT, MIXED) will be accompanied by substantial savings in interviewing costs: whereas the full set required asking a total of 66,000 items (1,500 Subjects \( \times \) 44 Items), the MAXIT, MINSTD, and MIXED procedures used only 22,500, 25,446, and 26,267 items, respectively.
This reduction in the total number of items was attained at the expense of a higher measurement error. Figure 1 addresses this issue by comparing the standard error of measurement at each trait level obtained from the full set (ALL) and the MAXIT tailored interviewing procedure. One can conclude that the measurement error increases with the trait level. The major reason for this result lies in the high concentration of low-threshold items in the item bank, as mentioned earlier. Because there were not enough high-threshold items in the item bank, even the best selection of the available items does not provide enough information for an accurate trait estimate. Figure 1 also indicates that the increase in measurement error was, in general, less accentuated than the

**Standard Error of Estimate vs. Trait Level**

*Standard Scale and MAXIT*

![Graph showing standard error of estimate vs. trait level for Standard Scale and MAXIT.*](image)

**FIGURE 1** Standard error of estimate versus trait level. Standard scale and MAXIT.
reduction of 66% in the number of items asked. The increase in error may be compensated by the substantial savings in interviewing time/cost, although an objective assessment of this tradeoff will depend on the value of the information gathered.

The objective of the MINSTD tailored interviewing procedure was to obtain a constant standard error of measurement of \( \sigma = .4 \) for all trait levels. This would produce a scale reliability of \( r(\theta, \hat{\theta}) = .92 \), if the standard error of measurement were in fact constant across the trait continuum. However, as can be seen in Figure 2, the actual standard error obtained was higher than .4 at the high trait levels. This, again, occurred due to the concentration of low-difficulty

![Standard Error of Estimate vs. Trail Level](image)

**FIGURE 2** Standard error of estimate versus trail level. Standard scale and MINSTD.
items in the item bank; for 132 high-trait subjects the standard error of the trait estimate was higher than .4 even after all 44 items had been asked. If a better item pool was available, with more items on the high-trait range, increased overall reliability and accuracy at the high-end of the trait continuum could be attained.

The MIXED procedure combines the other two stopping rules. In this case, the minimum standard error \( \sigma = .4 \) represents an upper bound for the measurement error; subjects who would require less than 10 items for a standard error of .4 were actually measured with a higher accuracy. The remaining subjects were asked as many items as necessary for this minimum standard error of measurement (up to a maximum of 44 items). Due to the less-than-ideal characteristics of the item bank, however, our results are not exactly as expected; for the same reasons discussed before, the minimum standard error was not attained for high-trait subjects, even after using all 44 items. Hence, the measurement error at this extreme was considerably higher than the upper bound of .4 (see Figure 3).

The final trait estimates obtained from each procedure (ALL, MAXIT, MINSTD, and MIXED) across the 1,500 respondents in the validation sample are compared in the correlation matrix in Table 2.

**DISCUSSION**

Using the full set (ALL) as a basis of comparison, Table 2 reports higher correlations for the MAXIT procedure, although this procedure required more items (25,466) than the MINSTD (22,500). The correlations between the tailored-interviewing and the full-set estimates should not be mistaken for reliability coefficients, because the former represents only correlations between estimates obtained by different methods, whereas the latter comprise the (unobservable) correlations between trait estimates and their true values. Note that these correlations are attenuated because the basis of comparison (ALL) also contains a measurement error. Nevertheless, they show a reasonable agreement between the tailored-interviewing estimates and the ones obtained from the full set of items, indicating that the loss of information due to the reduced set of items is not substantial. In fact, the considerable cost economies

<table>
<thead>
<tr>
<th></th>
<th>ALL</th>
<th>MAXIT</th>
<th>MINSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAXIT</td>
<td>.856</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MINSTD</td>
<td>.820</td>
<td>.901</td>
<td></td>
</tr>
<tr>
<td>MIXED</td>
<td>.820</td>
<td>.905</td>
<td>.996</td>
</tr>
</tbody>
</table>
Standard Error of Estimate vs. Trail Level

\textit{Standard Scale and Mixed}

\begin{figure}[h]
\centering
\includegraphics[width=0.6\textwidth]{figure3}
\caption{Standard error of estimate versus trail level. Standard scale and MIXED.}
\end{figure}

provided by tailored interview procedures (asking approximately only one third the original number of scale items) seem to offset this loss of information.

The intercorrelation between the MINSTD and MIXED estimates is artificially high, apparently because a large number of respondents (all individuals who required more than 10 items) had answered exactly the same items. Further, an important advantage of the MINSTD and MIXED procedures stems from a key common feature they contain: The interview session will last until the desired measurement error is achieved (up to the limit of using all the items in the measurement scale). Because an unbiased MLE procedure is used to
estimate the respondent's trait after each answer, the MINSTD and MIXED procedures will obtain a final measure no worse than from the full-scale administration, even if some of the initial interview items are misread, deliberately falsified, or accidentally miss-scored. Unless the data error is consistent (e.g., the respondent falsifies all the answers in the same direction), the standard error of measurement will be high. This results in a longer tailored interview, because incorrect answers affect not only the sequence of items, but also increase the number of items asked in the session. However, the final trait measurement will be affected only to the extent that some of the items were incorrectly answered, not because of the different sequence of items (note that incorrect answers to a full-scale administration also contribute to similar measurement error). Therefore, when the interview session uses MINSTD or MIXED stopping rules, incorrect answers to scale items lead to a longer tailored interview, thus affecting interview time and costs; on the other hand, measurement per se remains unaffected when the comparison is relative to the full-scale administration. Note, however, that the effects of incorrect answers when the MAXIT rule is used are just the opposite: Although the length of the tailored interview remains unaffected, the final measure could be affected in this case.

This research carries several methodological and substantive implications. Although several empirical studies employing IRT provide supportive evidence for the numerous advantages of tailored testing, the potential value of this concept has not been exploited in survey measurement so far. We argue that some recent technological advances in survey administration may facilitate the applicability of tailored interviewing. Specifically, we demonstrated the appropriateness of the tailored interview methodology when traits/attitudes of respondents are sought to be measured on multi-item scales in a survey. In this context, the benefits of the tailored interview procedure could be numerous: (a) it will provide a means to assess measurement error at each trait level, (b) it enables the researcher to actually prespecify the level of error acceptable in the survey design at each trait level, and (c) it saves money and time by asking respondents only a fraction of the total set of items in the measurement scale.

Interestingly, our empirical test reveals that the tailored interview offers advantages even under "less-than-ideal" conditions; for example, the literature specifies several ideal conditions for applications of tailored procedures in mental ability testing (see Ury, 1977). Although our item bank does not meet these ideal conditions, one can argue that these stringent conditions are necessary in educational applications and are less attainable in the context of trait items where the distinction between true and false responses is less clear. In the trait measurement context, it seems necessary to allow for more subtle nuances of agreement/disagreement than permitted by true or false dichotomies, thus leading to Likert-type scales and to the use of more elaborate latent trait models.
It is important to elaborate whether the potential for applying the tailored interview procedure is limited by certain: (a) characteristics of the measurement scale (e.g., unidimensional vs. multidimensional), or (b) characteristics of items in the scale (e.g., dichotomous vs. multichotomous response choices, criterion-referenced vs. factorially derived). Although our empirical results pertain to a unidimensional trait measured through dichotomous scale items, it must be recognized that these were just data characteristics and in no way imply that the tailored technique is restricted to such data. If multidimensional traits need to be measured, several latent trait models are available for this purpose (Mulaik, 1972; Reckase & McKinley, 1985; Samejima, 1974). Moreover, Urry (1977) suggested that the concept of tailored testing works very well under situations involving multidimensional traits. Further, even if the scale items have multichotomous response options (e.g., MMPI items are trichotomous), several latent trait models that accommodate this item property exist (Anderson, 1977; Andrich, 1978; Bock, 1972; Samejima, 1969) and these can be easily adapted to tailored interviewing applications. Finally, most personality scales contain items that are either criterion-referenced (e.g., MMPI, see Greene, 1980, for a discussion) or factorially derived. Inasmuch as scale items developed factorially would tend to be unidimensional (within a given factor) and criterion-referenced scale items may be unidimensional or multidimensional, this discussion suggests that the tailored-interview procedure would work regardless of this item characteristic.

Several factors attest to the appropriateness of the tailored interview procedure in Computer-Assisted Telephone Interviewing (CATI) or Computer-Assisted Interactive Interviewing (CAII) situations. First, these interviewing technologies represent widely accepted innovations in survey research and, therefore, the infrastructure for using them is already available. Second, although item banks have been developed for a wide battery of ability tests (see Urry, 1977), similar item banks for trait scales of interest (e.g., MMPI) could easily be developed based on reasonably large calibration samples. Third, the CATI/CAII environment seems far more conducive for tailored interviewing when compared to the computerized adaptive testing (e.g., using computer-based procedures to “tailor test” the mental abilities of students) environment facing educational researchers. The latter environment could potentially create several implementation problems, such as the unavoidable ethical issues which confound the research (e.g., it may be difficult to persuade a subject of the objectivity of his or her ability measure vis-à-vis the ability measure of another subject as long as both did not respond to the same set of test items). Despite such challenges, however, advances in the educational measurement literature represent the state-of-the-art in tailored testing procedures. Because these problems are absent in the CAII context, survey/personality researchers should endeavor to exploit the full potential of tailored interviews.
ACKNOWLEDGMENTS

We are grateful to Professor Robert C. Nichols for providing access to the data base used in this study and to Professor James Rounds for his valuable comments and suggestions.

REFERENCES

Bock, R. D. (1972). Estimating item parameters and latent ability when categories are scored in two or more nominal categories. Psychometrika, 37, 29-51.


Wagner A. Kamakura
Owen Graduate School of Management
401 21st Avenue South
Vanderbilt University
Nashville, TN 37203

Received December 21, 1987
Revised May 10, 1988
APPENDIX

Maximum-Likelihood Trait Estimation
With the Two-Parameter Logistic Model

The probability that a subject $j$ with trait level $\theta_j$ would answer an item $i$ positively (or "correctly") could be computed using Birnbaum's (1968) 2-parameter logistic model as:

$$ P_i(\theta_j) = \frac{1}{1 + \exp(a_i(b_i - \theta_j))} $$

where $a_i$ represents the item discrimination parameter, and $b_i$ represents the item threshold parameter.

Let $v_i = 1$ if the item is answered positively and 0 otherwise.

Then, the log-likelihood function for an observed set of $I$ responses from subject $j$ can be written as:

$$ L = \sum_{i=1}^{I} v_i \ln[P_i(\theta_j)] $$

and the maximum likelihood estimate of $\theta_j$ is the solution for:

$$ \frac{dL}{d\theta_j} = \sum_{i=1}^{I} a_i [v_i - P_i(\theta_j)] = 0 $$

The solution for equation (A2) can be obtained using Newton's gradient search. Let $\hat{\theta}_j = t$ trait estimate at Step $t$ of the gradient search ($\theta$ can be set to 0 at the initial step, assuming the median ability). Then the estimate can be revised iteratively with:

$$ \theta_{j,t+1} = \theta_j + \Delta_t $$

where

$$ \Delta_t = \frac{dL/d\theta_j}{-d^2L/d^2\theta_j} $$

$$ dL/d\theta_j = \sum_{i=1}^{I} a_i [v_i - P_i(\theta_j)] $$
\[
d\frac{L}{d^2\theta_j} = \sum_{i=1}^{I} a_i^2 (1 - P_i(\theta))P_i(\theta)
\]  \hspace{1cm} (A6)

The computations in (A3) and (A6) are repeated iteratively until \( \Delta \) is smaller than a prespecified tolerance limit.

The evaluation of (A6) at the final iteration provides an indication of how steep is the log-likelihood function around the trait estimate, namely, the level of uncertainty regarding the "true" trait. Therefore, an asymptotic estimate of the standard error of the trait estimate \( \theta_j \) can be computed as the inverse of \( \sqrt{d^2L/d^2\theta_j} \) at the final iteration. This value represents the standard error of measurement for the particular subject \( j \) at his or her particular trait level. In contrast, the traditional paper-and-pencil scales based on the classic measurement theory provide only a general assessment of measurement error over the whole trait continuum, across all subjects.