FRANK M. ANDREWS and RICK CRANDALL**

# THE VALIDITY OF MEASURES OF SELF-REPORTED WELL-BEING*

(Received 15 December, 1975)

ABSTRACT. Using a new analytic approach, construct validity estimates are developed for proposed social indicators of self-reported well-being. Two separate investigations are reported: the first involves data on six aspects of well-being each assessed by six methods from 222 adults in one geographic area; the second, a partial replication and extension, involves a more limited set of indicators measured on a sample of 1297 respondents representative of all American adults.

The results provide evidence that perceptions of well-being can be measured by single questionnaire or interview items using any of four formats with validities in the range of 0.7 to 0.8 and with correlated method effects contributing less than 10% of the total variance. Two other formats, however, were markedly less valid. These findings are important in view of past criticisms of 'subjective' social indicators as lacking in validity, and the findings can guide current efforts to develop new ways to assess the quality of life.

Methodologically, the article illustrates the feasibility and utility of deriving parameter estimates of structural equation models of multimethod-multitrait data using Joreskog's LISREL algorithm. The possibility of deriving validity estimates in this way, even when the data include correlated errors, opens new and important opportunities to precisely assess the amount of error variance in much social science data.

Increasing public concern about 'quality of life' and 'individual well-being' has stimulated a growing body of research in the area now called 'social indicators' (e.g., Andrews and Withey, 1974, 1976; Campbell *et al.*, 1976; Executive Office of the President, Office of Management and Budget, 1973; Wilcox *et al.*, 1972). This concern with measurement of life quality holds great significance for social scientists for two broad reasons. It has the potential of generating massive bodies of data which can become a rich resource for basic research on a wide variety of social phenomena. In addition, because of the potential impact on policy decisions, the social indicators area represents a vehicle by which social scientists' skills and perspectives can be brought to bear on important social problems.

At the heart of the social indicator movement are the indicators themselves. Selecting and developing indicators for subsequent monitoring is no simple task, and a wide variety of criteria need to be considered. One such criterion is the matter of validity: how well do the indicators measure what they are intended to indicate? However appealing a set of indicators might be in

*Social Indicators Research* 3 (1976) 1–19. *All Rights Reserved*
*Copyright* © 1976 *by D. Reidel Publishing Company, Dordrecht-Holland*

the eyes of public policymakers or private citizens, a set of invalid indicators can be expected to result in bad data, poor decisions, and eventual discredit to social scientists and the social indicator movement.

The concern of this paper is with the validity of a broad class of social indicators: self-report measures of individual well-being. The paper has both substantive and methodological orientations. It demonstrates how some newly developed techniques can be applied to derive estimates of construct validity, and it shows that a class of social indicators which has been criticized in the past for low validity[1] can be measured with moderate accuracy.

Concerns about validity of measurement have always been of interest to at least some scientists, and it has been widely granted that at least moderate validity is a necessary characteristic of any scientific measure. However, often only lip service is paid to serious investigation of validity and measurement topics. Writing for sociologists, Bohrnstedt and Carter have underscored the importance of validity concerns:

> Except for a few noted exceptions, sociologists seem to be blatantly unconcerned with the problems of measurement error . . . it is measurement error which produces the most serious distortions in our regression estimates. . . . our plea is for sociologists engaged in substantive research to confront the unreliability of their measurement instruments. . . . we do not feel it is either unrealistic or unreasonable to expect sociologists to recognize explicitly the error existent in their instruments and to take this error into account in their analyses. (Bohrnstedt and Carter, 1971, pp. 142–143)

Past work in sociology and psychology has clarified certain theoretical issues relevant to validity, reliability, and causality. Cronbach and Meehl (1955) distinguished construct validity (the relationship between an observed measure and an unobserved theoretical construct) from several other types of validity involving observed criteria. Heise and Bohrnstedt (1970) noted that the variance of a measure can be partitioned into three portions: valid variance (that which reflects what the measure is intended to measure), correlated error variance (that which reflects influences other than those the measure was designed to tap and which affect other measures as well), and residual variance. Heise and Bohrnstedt also noted that the validity of a measure depends on the proportion of its variance which is valid, while its reliability depends on the sum of the valid and correlated error proportions.

These perspectives seem particularly applicable for assessing the validity of measures of perceived well-being. People's feelings and perceptions are internal subjective states, of great importance to the person who holds them, but are not necessarily linked in a one-to-one relationship with any externally observable behaviour or set of life conditions. The absence of suitable validity criteria requires an assessment of *construct* validity. Furthermore, any feasible approach to assessing perceptions of large numbers of people for a substantial number of life aspects is likely to generate some common errors among the indicators, and hence their construct validity must be evaluated in the presence of correlated errors. This problem of estimating construct validity when non-independent errors are present is, of course, not restricted to measures of perceived well-being, but applies to a wide range of social science data.

An investigation of construct validity depends on both: (a) a network of relationships among a set of observed measures, and (b) a series of theoretical assumptions about the relationships of a set of hypothetical constructs (i.e., unobserved variables) to one another and to the observed measures. Direct estimates of construct validity have been relatively few. Use of a multitrait-multimethod design for this purpose was first formally suggested by Campbell and Fiske (1959). It represented an important advance at the time, but was unable to provide precise estimates of construct validity. Several later suggestions and counter-suggestions for methods of quantifying this approach were made in the psychological literature (Conger, 1971; Jackson, 1969, 1971), but the most precise use of this technique seems to be possible only when it is wedded to methods of structural analysis.

Work by Blalock (1964) and Duncan (1966) signaled the start of broad interest by sociologists in structural analysis. In a subsequent series of proposals it was shown that unmeasured constructs were incorporable within path models (e.g., Blalock 1970; Costner, 1969; Heise, 1969; Land, 1970). After it became evident that methods involving path models and factor analytic techniques were both special cases of structural equation models (Goldberger, 1972; Goldberger and Duncan, 1973), only one development remained. Joreskog (1969, 1970, 1973) developed a powerful maximum-likelihood technique for simultaneously estimating parameters for observed and unobserved variables in a structural model which allowed error components to be correlated.

In this paper we describe two applications of this new methodology. The first, using data from a somewhat restricted set of respondents, develops construct validity estimates for more than 30 measures designed as possible social indicators of perceived well-being. The second assesses the generality of the results of Investigation 1 by repeating a portion of the analysis on data representative of all American adults and extending the analysis to additional measures.[2]

## INVESTIGATION 1

### *Population*

The data come from 222 adults living in the Toledo, Ohio, area and up to three raters nominated by each of these respondents. The respondents do not constitute a probability sample, but they closely resemble both an actual probability sample of Toledo and a national sample of American adults with respect to age, sex, race, marital status, and employment. They tend to be somewhat more educated and to have modestly higher family incomes than is typical for Americans generally. These respondents were paid $25 to answer a 640 item questionnaire about their perceived quality of life. Administration took about 3 hours and was handled at local churches during July, 1973.

### *Measures*

The basic design of this investigation involves multimethod-multitrait data. For each respondent, data are available for six aspects of well-being (i.e., 'traits'), each of which was assessed by six methods.

The 'traits' are the respondents' affective evaluations of: (1) "Your house or apartment," (2) "Your independence or freedom – the chance you have to do what you want," (3) "The way you spend your spare time – your non-working activities." (4) "The way our national government is operating," (5) "Your standard of living – the things you have, like housing, car, furniture, recreation and the like," and (6) "Your life as a whole."

The six methods are:

(1) A rating by the respondent using a scale with seven categories labeled 'delighted', 'pleased', 'mostly satisfied', 'mixed (about equally satisfied and dissatisfied)', 'mostly dissatisfied', 'unhappy'. and 'terrible'.[3] Respondents were told "... we want to find out how you feel about various parts of your life, and life in this country as you see it. Please include [indicate] the

feelings you have now – taking into account what has happened in the last year and what you expect in the near future . . . How do you feel about _______?" This scale will be referred to as the Delighted-Terrible (or the D-T) Scale.

(2) A 7-point non-verbal scale consisting of stylized faces. Each 'face' consisted of a circle with two eyes (which did not change) and a mouth which varied from a 'smile' of almost a half circle to a similar half-circle upside-down for a 'frown'. Respondents were told: "Here are some faces expressing various feelings . . . Which face comes closest to expressing how you feel about your _______?" This is the Faces Scale.

(3) The next scale was drawn as a ladder with nine rungs.[4] The top rung was labeled "Best I could expect to have" and the bottom rung was labeled "Worst I could expect to have." Respondents were told: "Here is . . . [a] picture of a ladder. At the bottom of this ladder is the worst situation you might reasonably expect to have. At the top is the best you might expect to have. The other rungs are in between . . . Where on the ladder is your _______? On which rung would you put it?" This is the Ladder Scale.

(4) The fourth scale consisted of nine circles. Each was divided into eight 'slices' and each slice contained either a '+' or '–'. Circles were ordered so they contained progressively more pluses and fewer minuses. Respondents were told: "Here are some circles that we can imagine represent the lives of different people. Circle 0 has all minuses in it, to represent a person who has all bad things in his or her life. Circle 8 has all pluses in it, to represent a person who has all good things in his or her life. Other circles are in between. Which circle comes closest to matching how you feel about _______?" This is the Circles Scale.

(5) The last self-rating was a social comparison technique modified from one used by Holmes (1971; Holmes and Tyler, 1968). Respondents were told:

> Now let's compare your life and some aspects of it with the lives of six *people you know well.* It does not matter to us who these people are, but for your convenience, write down the initials of each person in the boxes provided below. (Think of real people you meet from time to time.)
>
> Under each set of initials put a 'B' if you think that on the whole your life (or in later questions, some aspect of it) is better *for you* than that person's would be.
>
> Put an 'S' if yours seems about the same for you as that person's would be.
>
> Put a 'W' if yours seems worse for you than that person's would be.

Questions read' "Compared to this persons's________, for me, my _______is:_______." A respondent's score was the mean of his answers, compared to up to six people, counting each B = 3, S = 2, and W = 1.

(6) The sixth method used ratings by others. The respondents provided names of others who knew them well. These people were mailed a short questionnaire. After appropriate follow-ups, data from an average of 2.3 raters were available for each respondent. Raters were told:

> These questions all concern how you *think* the person listed at the bottom of the letter feels about aspects of *his or her* own life . . . Tell how you think the person feels.

Items then followed in the format: "I think he or she feels _______ with his/her_________." Raters answered using the Delighted-Terrible Scale described above.[5]

### *Analysis*

The general form of the structural model used in this investigation appears in Figure 1, together with the results of applying the model to one subset of the measures. The analytic task is to estimate a set of parameters for this model which will come as close as possible to accounting for the observed covariations among measures, given what is assumed to be their linkages to the unmeasured causal variables (the hypothetical constructs) and the presumed linkages among these unmeasured variables.

The theoretical assumptions which make it seem reasonable to interpret some of the obtained parameters as estimates of validities and method effects are portrayed in the model or incorporated in the constraints imposed on the parameter values. Note that the variance of each observed measure (represented by the rectangles) is assumed to derive from three distinct sources: the respondents' 'true feelings' about the relevant aspect of well-being (shown in the circles on the left), the sensitivity of the particular method of measurement to effects of biases and/or halo, i.e., correlated errors or 'method effects' (shown in circles on the right), and a residual.

For each of the circles on the left, direct linkages are provided to all of the measures intended to tap an identical aspect of well-being but *not* to measures intended to tap other aspects. This helps to determine the 'meaning' of the circles on the left – i.e., they come to mean what the observed measures to which they are linked have in common, which in this case is feelings about a particular (identically defined) aspect of well-being. Although

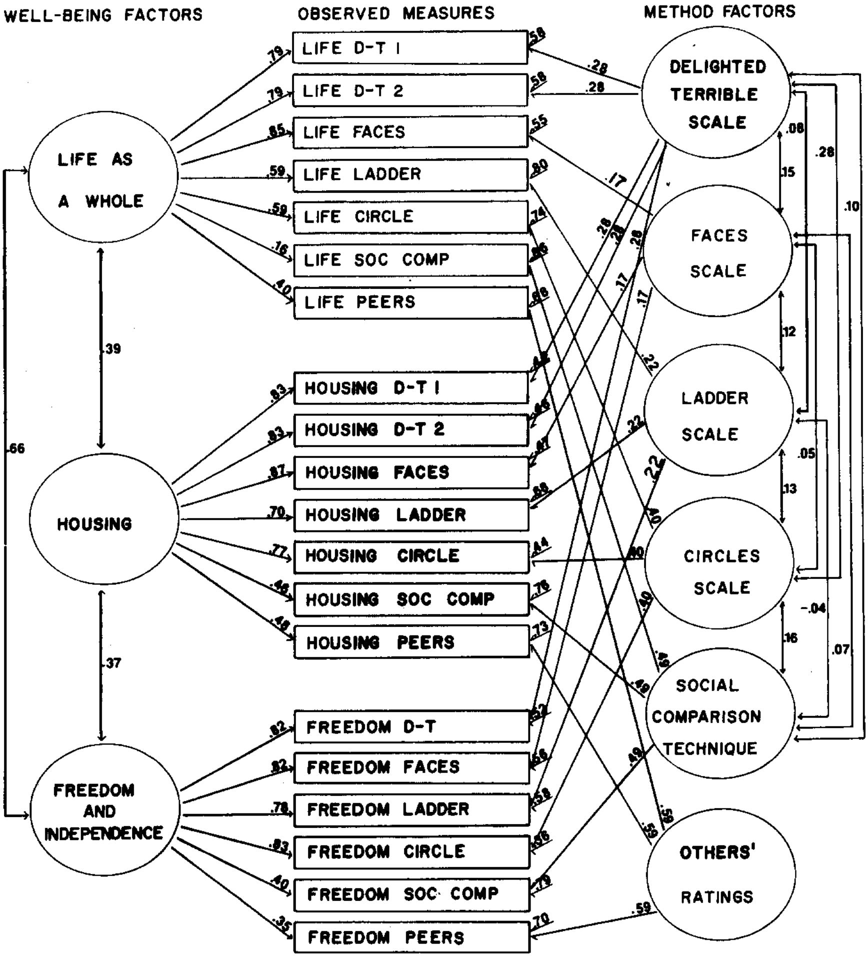

Fig. 1. General form of structural model used in Investigation I and parameter estimates for one sub-set of the measures (Analysis I).

there are no direct linkages between true feelings about one aspect of well-being and measures of other aspects, the model does provide direct linkages between the circles on the left. This incorporates our expectation that feelings about one aspect of life will be related to feelings about other aspects.

A similar set of theoretical assumptions governs the specification of linkages involving the circles on the right, intended to represent the method effects. It is assumed that a method can have a direct effect only on measures that use that method, but there may be relationships among certain of the method factors themselves. (Because measures derived from Others' Ratings were obtained from a source totally separate from that used for the remaining methods, this factor was constrained to be independent of the others.)

In addition to the pattern of linkages, certain constraints imposed on the parameter values help to ensure that the circles on the left and right take on the meanings intended for them. The most significant constraint involves linkages from the right-hand circles (the method factors) to the measures. The magnitude (whatever it might be) of the linkages from any one factor had to be the *same* for all the measures using the same method.[6] This forces the circle to mean something which has equal applicability to all of the measures using the same method, which is in accord with our expectations about the nature of a method effect. One minor constraint was imposed on the parameter estimates linking the circles on the left (the well-being factors) to the observed measures: estimates for measures referring to the same aspect of well-being and based on the same method should be equal. (Two such cases appear in Figure 1: two measures of Life-as-a-whole based on the D-T Scale, and two measures of Housing using this Scale.) For the analysis shown in Figure 1, there were no constraints on the parameter estimates between any pair of linked circles.

It would seem reasonable to use the model shown in Figure 1 to estimate the validity and error components of the measures because: (a) it incorporates our theoretical expectations about how various phenomena influence the observed measures; (b) serious alternative theories have not come to our attention; and (c) the model in fact fits the data rather well (as will be described shortly). We shall interpret the parameter estimates associated with the linkages between the observed measures and the circles which represent true feelings about well-being as *construct validity coefficients*. The estimates associated with the linkages between the measures and the circles which represent the method factors are interpreted as *method effect coefficients*.

Maximum-likelihood estimates of the parameters were obtained by application of the LISREL computer program (Joreskog and Van Thillo, 1972) to three overlapping subsets of the measures. The repeated applications kept the computing task within feasible bounds and had the effect of providing three independent estimates for each of the method effects and two independent estimates for some of the validity coefficients. Analysis 1 is shown in Figure 1. Analysis 2 included assessments of Freedom and independence, Standard of living, Spare time activities, and the National government by all of the methods. Analysis 3 included assessments of the Life-as-a-whole and Housing by all of the methods. The models applied in Analyses 2 and 3 were highly similar to that shown in Figure 1.[7]

The ability of the LISREL program to generate parameters which reproduce the observed correlations among the measures using models like that in Figure 1 was uniformly good. For Analysis 1, which is typical, the estimated correlations (of which there are 190) among the measures showed a mean deviation from the observed correlations of 0.055. In no case was the discrepancy more than 0.19.[8]

### *Results*

Figure 1 shows the results from the analysis of one subset of the measures. Table I brings together the complete set of validity and method effect coefficients from all three analyses.[9]

As shown in the table, three of the methods – the D-T Scale, the Faces Scale, and the Circles Scale – produced data with median validity coefficients approximating 0.8. Data obtained using the Ladder Scale had slightly lower median validity – 0.7. And the Social Comparison Technique and Ratings by Others showed validities of about 0.4. It was no surprise to find that the respondents' feelings were assessed less accurately by other people than by the respondents themselves, but the low validity of the Social Comparison Technique had not been expected.

Although there was nothing in any of the three analyses to require that the validity estimates be consistent for different aspects of life quality when assessed by the same method, it is reassuring to observe that they turn out to be similar. Nearly all are within ± 0.10 of the median validity value for the method. Thus the analyses include within themselves a series of internal replications which provide further support. Note, also, the generally close agreement of independent estimates of the parameters that were included in more than one analysis.

On the basis of these results, one can infer that single-item measures using the D-T, Faces, or Circles Scales to assess any of a wide range of different aspects of perceived well-being contain approximately 65% valid variance.[10]

TABLE I

Estimated validity and method effect coefficients using data from Investigation I

(Values in parentheses are replications from independent analyses involving overlapping subsets of the well-being measures)

| Measurement Method: | D–T Scale | Faces Scale | Ladder Scale | Circles Scale | Social Comparison | Others' Rating |
|---|---|---|---|---|---|---|
| Validity Coefficients: | | | | | | |
| Housing | 0.83 (0.85) | 0.87 (0.85) | 0.70 (0.70) | 0.77 (0.83) | 0.46 (0.45) | 0.48 (0.51) |
| Spare Time | 0.69 | 0.77 | 0.82 | 0.73 | 0.44 | 0.37 |
| National Government | 0.87 | 0.81 | 0.85 | 0.85 | [a] | 0.38 |
| Standard of Living | 0.79 | 0.77 | 0.70 | 0.80 | 0.38 | 0.37 |
| Freedom or Independence | 0.82 (0.84) | 0.82 (0.83) | 0.78 (0.79) | 0.83 (0.82) | 0.40 (0.46) | 0.35 (0.31) |
| Life-As-A-Whole | 0.79 (0.78) | 0.85 (0.80) | 0.59 (0.59) | 0.59 (0.58) | 0.16 (0.15) | 0.40 (0.42) |
| Median Validity | 0.82 | 0.82 | 0.70 | 0.80 | 0.42 | 0.38 |
| Method Effects[b]: | | | | | | |
| Analysis 1 (Figure 1) | 0.28 | 0.17 | 0.22 | 0.40 | 0.49 | 0.59 |
| Analysis 2 | 0.23 | 0.27 | 0.29 | 0.30 | 0.48 | 0.55 |
| Analysis 3 | 0.29 (0.25) | 0.27 | 0.00 | 0.28 | 0.50 | 0.52 |
| Median method effect | 0.27 | 0.27 | 0.22 | 0.30 | 0.49 | 0.55 |

[a] Data not obtained.

[b] Method effect parameters were constrained to be equal across aspects of well-being in each analysis (with one exception, as indicated, in Analysis 3).

Assessing the same aspects of well-being by either the Social Comparison Technique or through Others' Ratings resulted in only about 15% valid variance. The Ladder Scale fell in between, and produced about 50% valid variance. Clearly, these differences are substantial.

Table I also shows the magnitudes of the method effects. For the D-T, Faces, and Circles Scales, roughly 8% of the total variance could be attributed to method effects, whereas about 25% of the total variance was due to method effects when using the Social Comparison Technique, and 30% when using Ratings by Others. For the Ladder Scale, about 5% of the total variance was due to the method.

The relatively high method effects in measures obtained from Others' Ratings is notable but not surprising. Since other people have less direct access to the respondents' feelings than do the respondents themselves, one would expect substantially more 'halo' in others' ratings than in the respondents' own ratings.[11]

Two other types of parameters appear in Figure 1. Concerned that the several method effects might themselves be correlated, in Analysis 1 linkages were introduced between all of the methods involving data obtained directly from the respondent. As can be seen in Figure 1, these parameters turned out to be very close to zero. The second type of parameter involves correlations among the 'true' (unmeasured) perceptions of well-being. Here, of course, we expected substantial relationships, and they did in fact appear. (Although these latter relationships are interesting in their own right, they are essentially irrelevant to the main focus of the present paper, other than as linkages necessary to complete the measurement model.)

Before discussing the implications and potential uses of these results, we shall extend our explorations in Investigation 2.

## INVESTIGATION 2

The purpose of Investigation 2 is to assess the generalizability of the previous findings, which were derived from a questionnaire administered to people drawn from a local population. Of greater interest to developers of social indicators would be estimates of validity and method effects likely to be encountered in typical surveys of national populations. Investigation 2 provides a partial replication on nationally representative data and extends the prior results by examining certain additional aspects of life quality.

*Population and Method*

Self-reports of well-being were collected in May 1972 from 1297 respondents who constituted a probability sample of all American adults living in the 48 coterminous states and outside of institutions. The relevant items were included in a larger multi-purpose sample survey and were administered in an interview format by experienced interviewers in the respondents' own homes. All answers were on the Delighted-Terrible Scale described earlier.

Feelings about life-as-a-whole were assessed by two items, separated by about 15 minutes of intervening material, each of which asked "How do you feel about your life as a whole?" This part of the design provides a close replication of a portion of Investigation 1. Feelings about self, family life, and material well-being were also each assessed by several items. The items were "How do you feel about...

. . . yourself?
. . . the way you handle the problems that come up in your life?
. . . what you are accomplishing in your life?
. . . your children?
. . . your wife/husband?
. . . your marriage?
. . . the income you (and your family) have?
. . . your standard of living – the things you have, like housing, car, furniture, recreation, and the like?"

These items are different from those of Investigation 1, except for the standard of living measure.

The structural model estimated in Investigation 2 appears in Figure 2. With respect to the sources of variation in the observed measures, this model is identical to that used in Investigation 1: it assumes the variance in each observed measure can be apportioned into a "true" component, a "method" component, and a "residual" component. The model differs, however, in three respects: (a) there is only a single method effect (since all measures are based on self-ratings using the Delighted-Terrible Scale), (b) it is explicitly assumed that "true" feelings about life-as-a-whole are the result of 'true' feelings about self-efficacy, family life, material well-being, and an unspecified residual component, and (c) the constructs shown at the left of the figure are not defined by a series of items with identical content (as was the

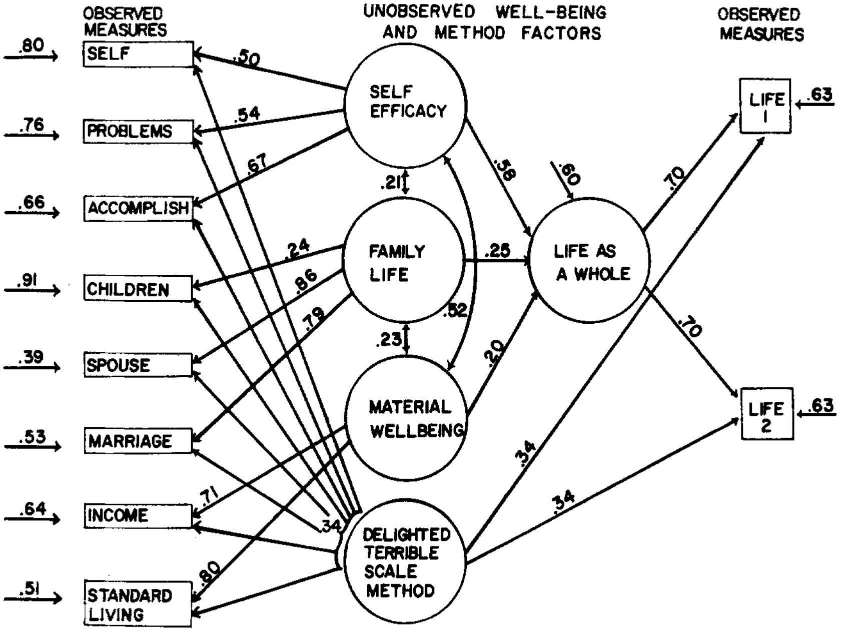

Fig. 2. Structural model used in Investigation 2 and parameter estimates.

case in Figure 1, and as is the case on the right side of Figure 2) but by items with somewhat differing content. For the right side of Figure 2, these differences do not affect our interpretation of the validity- and method-effect linkages which are the focus of this paper. However, for the left side of the figure, the linkages between the measures and the "true feelings" about aspects of well-being are more appropriately interpreted as factor loadings than as validity coefficients. Since the measures to which these factors are linked do not refer to precisely the same aspect of well-being, the meaning of the factor will differ somewhat from that referenced by any particular measure.

As in Investigation 1, LISREL was used to generate estimates of the parameters of the specified model which provide the best reproduction, under the maximum-likelihood criterion, of the variance-covariance matrix observed among the measures. Also as in Investigation 1, all method-effect parameters were constrained to be equal, the two measures of life-as-a-whole were required to have equal validities, and all non-indicated linkages were fixed at zero. Here again the fit was good – the average discrepancy between the observed and estimated correlations among the measures was 0.032, and the maximum discrepancy was 0.09.

### *Results and Discussions*

The results of Investigation 2 are represented by the parameters included in Figure 2.

For the measures of Life-as-a-whole, the validity estimates are 0.70 and the method effects are 0.34. While not identical to results obtained in Investigation 1, both these figures are reasonably close. (As can be seen in Table I, Investigation 1 provided two estimates of the validity of life-as-a-whole measured by self-ratings on the D-T Scale – 0.79 and 0.78 – and four estimates of the method effect – median = 0.27.)

While the differences are small, it is of interest to conjecture why the answers of the national sample respondents might show slightly lower validities and slightly higher method effects than those of the Investigation 1 respondents. One possibility is a difference in education level. The respondents in Investigation 1 included a somewhat larger proportion with college degrees, and a lower proportion of high school drop-outs than did the national sample. A second possible reason is the amount of practice each group had in answering questions about sense of well-being. Prior to encountering the first question asking "How do you feel about your life as a whole?" the Investigation 1 respondents had spent nearly an hour answering other questions dealing with life quality, whereas the Investigation 2 group encountered this item within the first few minutes after the interviewer began asking about perceptions of well-being. A third possible reason may lie in the difference between the paper-and-pencil format of Investigation 1 and the interview format of Investigation 2.

The linkages on the left side of Figure 2, between the observed measures and the several aspects of well-being, while not interpretable as validity coefficients, are also interesting. The Standard of Living item reflects the Material Well-Being construct with just about the same accuracy for these respondents (0.80) as its estimated validity in Investigation 1 (0.79). With the exception of the item about Children, most of the other parameters are estimated at similar or only slightly lower levels than the validities estimated in Investigation 1 for self-report measures using the D-T Scale format. (The low coefficient for feelings about children is probably not attributable to low item validity but a reflection of the fact that feelings about children are not the same as feelings about spouse and marriage.)

In general, the results of Investigation 2 suggest that the estimates of validity and method effects derived in Investigation 1 for measures based on self-reports using the D-T Scale can be generalized to typical national-level survey applications with only slight modification. The validity estimates of Investigation 1 may be slightly high, and the method effects slightly low, but the appropriate adjustments in the coefficients are probably less than 0.1. Although we will present no evidence concerning the generalizability of results for other measurement methods used in Investigation 1, other information available to us suggests that they also are reasonably representative of what can be achieved in national-level survey applications.

## IMPLICATIONS AND CONCLUSIONS

The results of these investigations can be considered from both quantitative and methodological perspectives.

Quantitatively, the results indicate that using any of several self-report methods, the validity of single questionnaire or interview items used to assess perceptions of well-being can be in the range of 0.7 to 0.8 – implying that roughly half to two-thirds of the variance is valid. By appropriate combination of several items which tap the same underlying perception, a composite measure could be formed which would have somewhat higher validity than any of the single items. For example, given validities of 0.80 and reliabilities of 0.74 (i.e., assuming that the variance of single items is composed of 64% valid variance,10% correlated error variance, and 26% residual variance), a five-item scale could be expected to produce an indicator with about 80% valid variance.[12] Thus while one would not want to claim that social indicators of perceived well-being are perfectly valid, it would appear that fairly substantial validities can be achieved through use of appropriate measurement and scale construction techniques. This finding is important to the social indicator movement because perceptual or 'subjective' measures have in the past been criticized as having low validity. Given the 'internal' nature of perceptions, it was previously difficult to demonstrate that such measures behaved as if they represented what they were supposed to represent and to derive quantitative estimates of their validity.

The results obtained here suggest that some methods for assessing perceptions of well-being are much better than others. Of those we investigated, self-ratings on the Delighted-Terrible Scale, the Faces Scale, and the Circles Scale showed the highest validities and were roughly equal with respect to the intrusion of method variance. The Ladder Scale, while slightly less subject to a method effect, also showed slightly less validity. Use of either the Social Comparison Technique or Ratings by Other people involved a very substantial loss of valid variance and a substantial increase in method variance.[13]

From a methodological perspective the results of this paper illustrate the feasibility and utility of applying structural measurement models to derive estimates of validity. The powerful new algorithms now available for estimating the kinds of models we have employed, coupled with appropriately designed social science data, make it possible to partition the variance of observed measures into true (i.e., valid) variance, error variance which is correlated with errors in other measures, and residual uncorrelated error variance.

The present application of this new technique and of the LISREL computor program produced validity estimates which were reasonable given the observed reliability of the measures (Andrews and Withey, 1974), which were consistent within any one measurement method across different aspects of well-being, and which were replicable in different sets of data.[14] While no one set of results can demonstrate the worth of a new procedure, the current findings serve as one such check.

The results of the present paper provide evidence that perceptions of well-being can be measured with substantial validity. This can be achieved using a variety of methods, for qualitatively different aspects of life, and under conditions typically encountered in national household-interview type surveys. While the present paper applies this methodology to estimate the validity of social indicators of perceived life quality, the approach is not limited to such measures, and we believe it could also be usefully applied to many other types of social science data.

*Institute for Social Research*
*University of Michigan*

## NOTES

*These investigations were part of a larger project titled Development and Measurement of Social Indicators which was directed by Frank M. Andrews and Stephen B. Withey and supported by the National Science Foundation through grants GS-3322 and GS-42015. A comprehensive report of the whole project is presented in Andrews and Withey, 1976. We are indebted to Marita DiLorenzi for assistance with the preparation and processing of the data presented in this article.

** The second author's current address is Institute for Child Behavior and Development, University of Illinois, Champaign, Ill. 61820.

[1] Andrews (1974) discusses these criticisms.

[2] After these investigations were designed, an article by Alwin (1974) appeared which provides an excellent exposition of the general rationale underlying our analyses. An earlier discussion by Boruch and Wolins (1970) approached many of these same issues from the perspective of restricted maximum-likelihood factor analysis. While applicable to our Investigation 1, the factor analytic model seems inappropriate for our Investigation 2.

[3] Three additional off-scale categories were labeled "Neutral (neither satisfied nor dissatisfied)," "I never thought about it", and "Does not apply to me". The few respondents who chose these categories were treated as having missing data in the analyses.

[4] The ladder format is adapted from Cantril (1965).

[5] Complete copies of all scales are available from the first author. Crandall (1976) reports further details of the analysis of the ratings by others. Although the design of our data is based on a 6 x 6 multitrait-multimethod matrix, there are actually 37 measures relevant to a total of 35 cells. The Social Comparison Technique was not applicable for assessing feelings about the national government, hence this cell was empty; and two cells each contained a pair of parallel measures – assessments of Life-as-a-whole and Housing using the D-T Scale.

[6] In the analysis of one subset of measures (later identified as Analysis 3), this constraint could be relaxed for measures employing the D-T Scale without making the model itself under-identified. The result of permitting this relaxation is included in Table 1 and suggests that the equality assumption is not unrealistic.

[7] Only minor differences in constraints distinguished the models applied in Analyses 2 and 3 from that shown in Figure 1. One difference has already been described in the preceding footnote. The other involved fixing all linkages between method factors at zero for Analyses 2 and 3. Since estimates for these linkages are close to zero in any case (see Figure 1), imposition of this additional restriction made little actual difference in the estimates of other parameters (as can be seen by the similarity of the doubly estimated parameters in Table I).

[8] Using a significance test considered by Jorsekog and Van Thillo (1972), these and subsequent deviations from fit could be significant. However, the significance level is inflated by large sample sizes while the estimated parameters do not vary (Joreskog, 1969) and for our present purposes, the absolute size of the deviations is of greater interest than their statistical probability.

[9] All parameter estimates reported in this paper pertain to standardized variables (i.e., with variance = 1.0).

[10] The square of the validity coefficient indicates the proportion of a measure's variance which is valid. (Similarly, the square of the method effect coefficient indicates the correlated error component of a measure's variance.) Although the 7-point satisfaction scale used by Campbell *et al.* |(1976) is not among the measures examined here, it is of interest for comparison purposes to note that in a 1972 national survey 7-point satisfaction was estimated to yield about 58% valid variance while the D-T Scale yielded

63% when both measures were used to assess evaluations of the same thing, life-as-a-whole (see Andrews and Withey, 1976, Chapter 6).

[11] It is interesting to note the substantial similarity across the three analyses in the independent estimates of the method effect for any given method, despite the fact that these were derived in the context of different combinations of 'traits'. This suggests that the equality constraint on the method effect parameters was not unreasonable. This conclusion is reinforced by the finding of similar method effects for the D-T Scale in Analysis 3 (0.29 and 0.25), where the equality constraint was not imposed.

[12] This result is derived by application of Guilford's (1954) formula 14.37.

[13] Though validity is an important criterion when selecting or developing measurement methods, other factors should be considered. Among them are the nature of the resulting distributions (e.g., degree of skew), the explicitness with which scale categories are labeled, and the extent to which the method requires particular verbal skills or is dependent upon particular words having equivalent meanings in different sub-cultures.

Considering these criteria, the Delighted-Terrible Scale is more explicitly labeled than the other valid methods. The Faces Scale is relatively independent of verbal skills. The Circles Scale shares this characteristic and it may be slightly less directed to feelings than the other scales because of its emphasis on concrete "good and bad things in your life". Of the four most valid methods. the Ladder Scale has the least skew and the Faces Scale the greatest. The Ratings by Others are low in validity but have the advantage of independently confirming the self-report method. The wording of the Social Comparison Technique was different than in its original applications. A different operationalization may be more valid.

[14] Consistency of final estimates was also observed when the LISREL algorithm was started from different initial values, which supports the program's ability to generate a unique solution.

## BIBLIOGRAPHY

Alwin, D. F.: ·1974, 'Approaches to the Interpretation of Relationships in the Multitrait-Multimethod Matrix', In *Sociological Methodology 1973–74* (edited by H. L. Costner), Jossey-Bass, San Francisco.

Andrews, F. M.: 1974, 'Social Indicators of Perceived Life Quality', *Social Indicators Research* 1, 279–299.

Andrews, F. M. and Withey, S. B.: 1974, 'Developing Measures of Perceived Life Quality: Results from Several National Surveys, *Social Indicators Research* 1, 1–26.

Andrews, F. M. and Withey, S. B.: 1976, *Social Indicators of Well-being: Americans' Perceptions of Life Quality*, Plenum, New York, in press.

Blalock, H. M., Jr.: 1964, *Causal Inferences in Nonexperimental Research*, University of North Carolina Press, Chapel Hill.

Blalock, H. M., Jr.: 1970, 'Estimating Measurement Error Using Multiple Indicators and Several Points in Time', *American Sociological Review* 35, 101–111.

Bohrnstedt, G. W., and Carter, T. M.: 1971, 'Robustness in Regression Analysis', in *Sociological Methodology 1971* (ed. by H. Costner), Jossey–Bass, San Fransisco.

Boruch, R. F. and Wolins, L.: 1970, 'A Procedure for estimation of Trait, Method, and Error Variance Attributable to a Measure', *Educational and Psychological Measurement* 30, 547–574.

Campbell, A., Converse, P. E., and Rodgers, W. L.: 1976, *The Quality of American Life: Perceptions, Evaluations and Satisfactions*, Russell Sage Foundation, New York.

Campbell, D. T. and Fiske, D. W.: 1959, 'Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix', *Psychological Bulletin* 56, 81–105.

Cantril, H.: 1965, *The Pattern of Human Concerns*, Rutgers University Press, New Brunswick, New Jersey.

Conger, A. J.: 1971, 'Evaluation of Multimethod Factor Analysis', *Psychological Bulletin* 75, 416–420.

Costner, H. L.: 1969, 'Theory, Deduction, and Rules of Correspondence', *American Journal of Sociology* 75, 245–263.

Crandall, R.: 1976, 'Validation of Self-Report Measures Using Ratings by Others', *Sociological Methods and Research*, 4, 380–400.

Cronbach, L. J. and Meehl, P. E.: 1955, 'Construct Validity in Psychological Tests', *Psychological Bulletin* 52, 281–302.

Duncan, O. D.: 1966, 'Path Analysis: Sociological Examples', *American Journal of Sociology* 72, 1–16.

Executive Office of the President: Office of Management and Budget, 1973. *Social Indicators, 1973.* U.S. Government Printing Office, Washington D.C.

Goldberger, A. S.: 1972, 'Structural Equation Methods in the Social Sciences', *Econometrica* 40, 979–1001.

Goldberger, A. S. and Duncan, O. D. (eds.): 1973, *Structural Equation Models in the Social Sciences*, Seminar Press, New York.

Guilford, J. P.: 1954, *Psychometric Methods* (2nd ed.), McGraw-Hill, New York.

Heise, D. R.: 1969, 'Separating Reliability and Stability in Test-Retest Correlation', *American Sociological Review* 34, 93–101.

Heise, D. R. and Bohrnstedt, G. W.: 1970, 'Validity, Invalidity, Reliability', in *Sociological Methodology 1970* (ed. by E. F. Borgatta and G. W. Bohrnstedt), Jossey–Bass, San Fransisco.

Holmes, D.: 1971, 'Conscious Self-Appraisal of Achievement Motivation: The Self-Peer Rank Method Revisited', *Journal of Consulting and Clinical Psychology* 36, 23–26.

Holmes, D. and Tyler, J.: 1968, 'Direct Versus Projective Measurement of Achievement Motivation', *Journal of Consulting and Clinical Psychology* 32, 712–717.

Jackson, D. N.: 1969, 'Multimethod Factor Analysis in the Evaluation of Convergent and Discriminant Validity', *Psychological Bulletin* 72, 30–49.

Jackson, D. N.: 1971, "Comments on 'Evaluation of Multimethod Factor Analysis', *Psychological Bulletin* 75, 421–423.

Joreskog, K. G.: 1969, 'A General Approach to Confirmatory Maximum Likelihood Factor Analysis', *Psychometrika* 34, 183–202.

Joreskog, K. G.: 1970, 'A General Method for Analysis of Covariance Structures', *Biometrika* 57, 239–251.

Joreskog, K. G.: 1973, 'A General Method for Estimating a Linear Structural Equation System', In *Structural Equation Models in the Social Sciences*, (ed, by A. S. Goldberger and O. D. Duncan), Seminar Press, New York.

Joreskog, K. G. and van Thillo, M.: 1972, LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. Unpublished research bulletin, RB–72–56, Princeton, New Jersey, Educational Testing Service, December.

Land, K. C.: 1970, 'On the Estimation of Path Coefficients for Unmeasured Variables from Correlations among Observed Variables', *Social Forces* 48, 506–511.

Wilcox, L. D., Brooks, R. M., Beal, G. M., and Klonglan, G. E.: 1972, *Social Indicators and Societal Monitoring: (An Annotated Bibliography)* Jossey-Bass, San Francisco.