Appendix C:
Constructing Indexes: Recommendations for Improving Measurement

Mahoney, Thombs, & Howe, (1995) described how both qualitative and quantitative methods (e.g., factor analysis) can be combined to develop a psychometrically sound scale. It is hoped that the present paper will extend that discussion. More specifically, the purpose of this paper is to (a) compare and contrast indexes and scales; (b) review recommended index construction guidelines using a comprehensive "real world" example so as to improve measurement; and (c) discuss the roles of dimensionality, item scoring, item analysis, factor analysis, reliability, and validity in index development.

Indexes and Scales: The Commonalities

Babbie (1990, p. 149) has observed "[a]n examination of the substantive literature based in survey data shows that indexes are used much more frequently than scales...[i]ronically, however, the methodological literature contains little, if any, discussion of index construction, though discussions of scale construction abound." Neither Kerlinger (1986), Crocker and Algina (1986), Udinsky, Osterlind, and Lynch (1981) draw the distinction between indexes and scales as does Babbie. However, all describe a type of scale that Babbie labels as an index and do report differences in the scoring procedure and the amount of information a score provides as compared to the other types of scales.

While other authors may not draw his distinction, Babbie (pp. 148-149) does provide a useful starting point for discussing indexes and scales. Babbie (1990), Kerlinger (1986), and Crocker and Algina (1986) describe the three principal types of "scales" which are:

Babbie (1990)	**Kerlinger (1986)**	**Crocker & Algina (1986)**
Index	Summated Rating Scale	Subject-centered [Scaling] Method
Thurstone Scales	Thurstone Equal-Interval	Stimulus-centered [Scaling] Methods
Guttman Scaling	Guttman Scale	Response-centered [Scaling] Approach

With respect to the construction of any one of the above, Babbie, Kerlinger, and Crocker and Algina all outline similar development processes and score interpretations. According to Babbie both indexes and scales: (a) are both usually ordinal measures of variables, i.e., produce ordinal data; (b) rank-order respondents in terms of (a) specific variable(s) and (c) are composite measures of responses to several items designed to tap various dimensions or levels of a variable.

In constructing either an index or scale, it is recommended that a theory be used as the basis for item writing and interpretation. A theory provides not only provides a basis for the above but also helps to establish an index's or scale's dimensionality. A unidimensional index or scale would measure a single variable. An index or scale composed of four subtests each assessing a unique variable dimension would be multidimensional, but each subtest must be unidimensional respecting its unique variable dimension. Thus a theory or variable composed of four aspects or dimensions would require an index or scale consisting of four subtests. Each subtest would address a separate variable dimension and be mutually exclusive of the other. Such an index or scale would yield four scores.

Indexes and Scales: The Differences

Indexes

Definition. According to Babbie, "good indexes provide an ordinal ranking of respondents on a given variable" (p. 165). With indexes, response patterns are not used in scoring; all the subject does is indicate his or her position on a continuum of interest. Indexes

are constructed by simply adding responses to several items designed to measure a variable or a variable dimension (also called level). The chief advantage of an index is the relative ease (as compared to a scale) of construction.

Weighting responses. Crocker and Algina (1986, p. 50) point-out, "item scoring weights are assigned by an arbitrary decision of the scale [index] developer, and typically all items are weighted equally." Scoring options can range from "0" (does not possess variable characteristic or "1" (possess variable characteristic to the traditional Likert Scale, e.g., "strongly disagree (1)" to "strongly agree (5)". If an index developer uses a five point Likert Scale which ranges from "strongly disagree (1)" to strongly agree (5)", all a subject would do is indicate his or her position on that continuum with respect to the variable of interest by circling any number between 1 and 5.

Interpretation. Crocker and Algina (p. 50) state, "[l]ittle or no attention is given to possible differences between the items to which the subject is required to respond because the sole purpose is to 'scale' [i.e., locate the subject on a continuum of interest]." Babbie (1990) agrees. Thus a conservative interpretation of an index score is placement on a continuum of interest, only. However, Kerlinger (1986, p. 454) argues that summated rating scales (indexes) allow for an expression of intensity. Referring to indexes, Crocker and Algina (p. 50) note, "scaling models have not been developed to scale the properties of the scores derived." They also point out, "[t]he total [index] score is computed by simply *assuming* that point values assigned to each possible response option form a numeric scale with the properties of order or with order and equal units. It is the property of order which allows for the ranking of subjects along a continuum of interest. The property of equal distance allows for the determination of

intensity a subject holds to a position on the continuum. The key points are these: (a) any suggestion of intensity of feeling in indexes is restricted to those items employing Likert type response options and (b) any assertion of intensity is inferred. With indexes, intensity can not be empirically established as can be done with a scale.

Scales

Definition. Scales, according to Babbie (p. 148) are "constructed through the assignment of scores to *response patterns* among the several items comprising the scale..[a] scale differs from an index in that it takes advantage of any *intensity structure* that might exist among the individual items." Babbie (p. 165) illustrates "intensity structure" using an example from the Bogardus Social Distance Scale. In this scale are five items designed to assess the extent to which a subject would accept Albanians. The first item, "Are you willing to permit Albanians to live in your country?", is the softest and is weighted "1." The fifth item, "Would you let your child marry an Albanian?", is the hardest and is weighted "5." Logically, a respondent who would let his or her child marry an Albanian would also let Albanians live next door, in the neighborhood, in the community, or in the country. A score of "5" tells not only how close a subject would allow Albanians, but also the intensity to which the subject holds to his or her position as well as predict his or her response patten. Such a scale yields ordinal data as a subject (or stimulus or both) can be ranked; however, no statement about how much closer is the "neighborhood" from the "community" and so on, can be made.

Thurstone scales. The Thurstone scale is described by Babbie (pp. 166-167); and is referred to as the "Thurstone equal appearing interval scale" by Kerlinger (pp. 454-455). Crocker and Algina (pp. 50-55) describe stimulus-centered scaling where the unit of interest is

the "location of a stimuli (or item) [not the subject] on a psychological continuum." Because, the methodology requires that stimuli or items be placed on a continuum at equal distances apart, it is possible to estimate the distance between stimuli or items on the continuum. Thus, for each stimuli or item, a scale value is derived from specified equations. The process of estimating the distance between items requires the contributions of many item judges, who usually have to examine dozens of items and all possible pairwise comparisons of those items. Thurstone scales are not widely used in survey research given their labor intensity, costs, and long development process (Babbie, 1990, p. 167).

Guttman scales. Babbie (pp. 167-170; Kerlinger, pp. 455-456). Crocker and Algina (1986, p. 55) describe the Guttman scale as response-centered and go on to observe, "[r]esponse data are used to scale subjects on a psychological continuum, based on the strength of the items endorsed (or answered correctly); at the same time, items are scaled in terms of the strength or amount of the trait possessed by the subjects who endorse them" (pp. 55-56). In a Guttman scale, both the items and subjects are scaled. Items tend to be small in number and homogeneous (Kerlinger, 1986, p. 455). Items may be ordered similar to the pattern found in the Bogardus Social Distance Scale and in Thurstone style scales. Allowable response patterns are determined. Response patterns which do not conform to allowable patterns are termed errors. Knowing a subject's score, allows one to predict a subject's response pattern to the stimuli or items presented. The Guttman scale is difficult to construct; can be constructed only after data have been collected (Babbie, pp. 170) and is sample dependent. Udinsky, Osterlind, and Lynch (1981) comment, regarding the Guttman scale, "reaching a satisfactory criterion of scalability is difficult;...the selection of good items is difficult; it is difficult to construct; scoring is easy; it is

very time consuming; [and] it is not recommended for research studies" (p. 172).

Commentary

In view of the differences between the index and the various scales, particularly construction difficulties, costs, and score interpretation, Babbie's distinction makes sense. This is especially true when one considers that many evaluators, surveyors, and decision-makers have little experience in correctly constructing Thurstone or Guttman style scales. Thus, the easier to construct index (also called summated rating scale) provides a practical and powerful alternative.

Babbie (1990) also asserts that the terms, "index" and "scale" are used interchangeably in the survey literature, thus resulting in the loss of their precise technical definitions which has lead to misapplication and misinterpretation. The most common error being the attribution of properties to an instrument, which it simply does not possess. A knowledge of the differences between the two, especially score interpretation, should limit inappropriate attribution.

Constructing Indexes: Recommendations

While the following recommendations can be validly extended to scales, we will limit ourselves to the use of the term, index, so as to avoid confusion. Babbie (1990, pp. 151-165), as well as others, offers several recommendations for constructing indexes. The following topics will be discussed: (a) item writing and selection, (b) index scoring and missing data, (c) examining item interrelationships (bivariate and Multivariate), and (d) index reliability and validity examination.

Item Writing and Selection

Babbie (p. 151) argues, "[t]he first criterion for selecting items to be included in the index is **face validity** (logical validity)." However, Pedhazur and Schmelkin (1991) would argue that

the first step in index construction is the identification of an applicable theoretical model addressing the phenomena of interest. Once this has been done, the relevant literature should be reviewed and digested. Often, in reviewing the relevant literature, a theoretical model is found as well as an already constructed index. If this is the case, and the index possesses the desired content and psychometric properties, then it may be used. It is essential that appropriate copyright requirements are met. When it is necessary to develop an index, a basis exists to move forward, as the relevant literature has been reviewed and theory identified.

While face validity is not a psychometric concept, Babbie's advice does have merit. Items should look like they belong to the index. For example, an item addressing an aspect of math anxiety should probably not be included on a racial sensitivity index. It is a good idea to write more items than are probably needed for the final index version; this will allow the selection of items which contribute to variance. Crocker and Algina (1986, p. 80) have suggested the following guidelines for item writing:

1. Put statements or questions in the present tense.
2. Do not use statements that are factual or capable of being interpreted as factual.
3. Avoid statements that can have more that one interpretation.
4. Avoid statements that are likely to be endorsed by almost everyone or almost no one.
5. Try to have an almost equal number of statements expressing positive and negative feelings.
6. Statements should be short, rarely exceeding 20 words.
7. Each statement should be a proper grammatical sentence.
8. Statements containing universals such as *all, always, none, and never* often introduce ambiguity and should be avoided.
9. Avoid use of indefinite qualifiers such as *only, just, merely, many, few, or seldom*.
10. Whenever possible, statements should be in simple sentences rather than complex or compound sentences. Avoid statements that contain "if" or "because" clauses.
11. Use vocabulary that can be understood easily by the respondents.
12. Avoid use of negative (e.g., *not, none, never*).

These recommendations apply to both Likert style responses as well as dichotomous

(e.g., agree/disagree formats). Another excellent discussion about item writing is found in Kubiszyn and Borich (1993, pp. 74-119). The criteria for constructing test items advanced by Osterlind (1989) and the test item editorial guidelines as outlined by Oosterhof (1990) are commended for review.

Items should be unidimensional and contain the degree of specificity intended. For example, there are six variable in Cross' Chain of Response Model (1981, p. 124) which attempts to explain adult participation in learning activities. Items designed to tap the academic self-efficacy dimension should not materially (e.g., $\geq$ .30) correlate with items designed to tap the educational attitudes dimension.

Index Scoring & Missing Data

Weighting response options. The most common method of weighting index items is equally. Thus, each response option is given a numerical value which is then summed across all the items comprising the unidimensional index or index subtests. For example, the following scheme might be used: strongly disagree, 1; disagree, 2; no opinion, 3; agree, 4; and strongly agree, 5. Each option (e.g., disagree and agree) are weighted equally; the different numbers are used only to identify the particular response option. This has been shown to produce satisfactory results (Pedhazur and Schmelkin, 1991, p. 125). The item scoring strategy for a scale is more complex (Babbie, 1990, p. 158). Regardless of the scoring scheme selected, it should provide respondents with a full range of response options, given the theory under consideration and the nature of the sample or population under investigation.

Dimensionality and scoring. The establishment of dimensionality for an index or index subtest is essential as, "the validity of summing scores on items is predicated on them tapping the

same dimension" (Pedhazur & Schmelkin, p. 124). Pedhazur & Schmelkin (p. 125) go on to advise that it is often best to present a score as an average for the subtest (i.e., divide the total score by the number of items. This is most helpful if Likert style options have been utilized as the average is easier to interpret. While differential item weighting is possible, such is not recommended as the strategy is "not worth the trouble" (Pedhazur & Schmelkin, p. 626; Nunnally & Bernstein, 1994, p. 332). Babbie (1990, p. 159) expresses similar sentiment. If a multidimensional index is being utilized, it makes no sense to derive a total (i.e., index) score which would be meaningless (Pedhazur & Schmelkin, 1991, p. 69). If for example, four subtests are used, then four separate scores should be computed and reported (Pedhazur & Schmelkin, p. 69, 125).

Dichotomously scored items. Of special note, for indexes with dichotomous item scoring (e.g., correct/incorrect, agree/disagree, yes/no, etc.) there is less variance than in Likert style response options. To adjust for this, very large samples are recommended (Kline, p. 127). For dichotomously scored items on achievement tests (e.g., correct/incorrect), Kline does recommend, "that the correct answer should be obtained by 20-80 percent of the sample. This rules out undiscriminating items" (Kline, 1994, p. 127).

Ordinal or interval data? Given that indexes produce ordinal data, there is debate as to whether or not such data may be treated as interval and hence, subjected to inferential parametric statistical testing. Crocker and Algina (p. 63) have recommended that if meaningfullness is enhanced by application of parametric testing (parametric statistics require at least interval level data), then do so; if not, then treat data as ordinal and apply nonparametric inferential statistics. Pedhazur and Schmelkin (1991, p. 27) agree.

Missing data. With respect to missing data, there are several strategies that can be employed. The simplest method is to just ignore missing data and include only those items for which data are available. This will work for most indexes provided the number of missing data points is not too large. Other strategies include: (a) using only instruments where all items have scores, (b) substituting the mean item score where an item score may be missing, or (c) inserting the median item score where missing. Unless a statistical procedure requires a particular strategy, the choice is largely up to the index designer. It is recommended that missing item level scores be ignored unless there is a lot of missing data as such a strategy uses all existing raw data provided by respondents.

Examining Item Interrelationships: Bivariate

According to Babbie, items which have a logical relationship should have an empirical one. Items which are used to classify or assess a variable or variable dimension should be related (i.e., correlated with each other). The purpose of this stage in the analysis is to assess how items behave and to generate information to assist in item retention, revision, and deletion. There are two commonly used strategies to examine item bivariate relationships: item/index (or subtest) correlations using item analysis and interitem correlations. Before examining bivariate item relationships, item descriptive statistics should be studied, giving one a "feel" for the data.

Item analysis. According to Pedhazur and Schmelkin (1991, p. 124), "[item analysis] is of limited usefulness so far as determination of dimensionality of a scale [index] is concerned." Item analysis assumes dimensionality, whether or not present. If such an assumption is false (as will be seen), then a faulty analysis will occur. According to Kline (p. 127), "Nunnally (1978) advocates that item analysis be used to make the first item selection and then the items be

factored." Kline has concluded, "item analysis usually gives similar results to factor analysis" when a theory is employed.

An item analysis example. Following is an example of a research project which attempted to develop a short self-assessed competency index for health education specialists. The authors theorized a four dimensional model consisting of clinical, program planning, program evaluation, and program management dimensions. A 30 item instrument, consisting of four subtests and utilizing a five point Likert style response option set was developed. The range of allowable responses were "Very Incompetent (1)", "Incompetent (2)", "No Opinion/Unsure (3)", "Competent (4)", and "Very Competent (5)." The instrument was reviewed by a panel of five judges experienced in providing direct health education and continuing professional education services. Once collected, data were subjected to both item and factor analysis. Item analysis results are presented in Table 1 and factor analysis results are presented in Table 2.

Item level descriptive statistics. Presented in Table 1 are item means, variance, and item-index ("scale" on many item analysis programs) correlations. Most item means tend to hover close to 4.0, indicating that respondents saw themselves as competent across most items. However, there were several topics where respondents seemed less sure of their competence as several means were close to 3.0, e.g., items 14, 17, 18, and 30. It was interesting to note that the highest item mean was 4.21, suggesting that respondents saw "room to grow" across most of the topics listed. It is clear that most respondents answered the items as the maximum number of possible respondents was 141.

Table 1: Item Mean, Variance, Item/Scale Correlation, & N/Item

#	Item	Mean	Variance	Item/Scale Correlation	N/Item
	Clinical Skills/Content Subtest				
1	Health Behavior Theories	3.75	.707	.67	138
2	Selecting Interventions	3.86	.650	.69	139
3	Selecting Educ. Media	4.15	.556	.59	140
4	Community Organizing	3.96	.849	.57	140
5	Teaching Methods	4.21	.593	.69	140
6	Behavior Change Tech.	3.81	.770	.76	140
7	Learning Theories	3.76	.723	.75	140
8	Case Management	3.46	.906	.39	140
	Program Planning Skills/Content Subtest				
9	Establishing Service Priorities	3.78	.778	.62	139
10	Developing Work Plans	3.97	.913	.78	140
11	Developing Staffing Plans	3.76	.980	.79	140
12	Developing Budgets	3.62	1.250	.73	140
13	Planning Document Writing	3.71	.807	.68	140
14	Funding Research	3.18	1.083	.67	139
15	Community Health Analysis	3.44	.782	.53	138
	Program Evaluation Skills/Content Subtest				
16	Program Evaluation Theory	3.55	.881	.75	139
17	Qualitative Evaluation Designs	3.26	.948	.82	140
18	Quantitative Evaluation Designs	3.17	.883	.80	138
19	Sampling Techniques	3.26	.816	.80	138
20	Data Gathering Methods	3.50	.844	.79	138
21	Basic Statistical Procedures	3.32	1.065	.82	139
22	Basic Epidemiology Methods	3.33	1.013	.63	139
23	Report Writing [Evaluation]	4.11	.716	.44	140
	Program Management Skills/Content Subtest				
24	Basic Management Theory	3.67	.649	.72	140
25	Negotiating Techniques	3.59	.784	.71	140
26	Coordinating Resources	4.06	.597	.71	140
27	Using Information Systems	3.61	.799	.57	139
28	Basic Marketing Strategies	3.58	.658	.74	140
29	Working with Power Groups	3.34	.768	.77	140
30	Power Structure Analysis	3.07	.729	.68	139

Examining bivariate item relationships. Referring to Table 1, item/index (also called scale subtest) correlations seem robust except for items 8 and 23. Items which correlate well with the scale or subtest should be retained (Pedhazur and Schmelkin, 1991, p. 124) as "[they] are measuring what most of the [other] items are measuring" (Kline, 1994, p. 137). An item correlation matrix for the management subtest is presented in Figure 1 for illustration. Most of the items correlated well with each other. While some of these correlations are modest they do suggest that the items share some variance. To determine, the percentage of variance shared by the items, square the correlation coefficient.

	24	**25**	**26**	**27**	**28**	**29**	**30**
24	1.00						
25	.553	1.00					
26	.473	.526	1.00				
27	.247	.324	.500	1.00			
28	.555	.507	.518	.274	1.00		
29	.500	.469	.371	.371	.527	1.00	
30	.426	.248	.289	.219	.444	.701	1.00

Figure 1. Program Management Subtest Item Correlation Matrix

Application. Relevant items, given the prevailing theoretical context, with the largest variance should be included in any resulting index or scale. Item variance contributes to subtest reliability (Table 3); generally, the greater the variance, the higher is the reliability coefficient. Thus, items with marginal contributions to variance, relative to other items, could be revised or dropped. Because it is often necessary to delete items, a large pool of items which may be potentially included in an index should be constructed.

Examining Item Interrelationships: Multivariate

Babbie (1990, p. 156) asserts that multivariate relationships between variables or variable

dimensions should be examined before being combined into an index. Pedhazur and Schmelkin (1991, p. 124) have suggested that factor analysis (FA) is an effective procedure for establishing the dimensionality (uni-or multi-) of an index across a variable or set of variable dimensions. FA is also practical; it is often difficult to examine large correlation matrices and arrive at an "overall" conclusion. While only a brief FA discussion is presented here, rather thorough discussions are found in Crocker & Algina, 1986, pp. 287-308; Kline 1993; Kline, 1994; Pedhazur & Schmelkin, 1991, pp. 66-73, 590-694). For dichotomously scored items, Kline (1994, p. 126) suggests that the phi correlation coefficient rather than the Pearson coefficient (as is standard) be used in the initial correlation matrix prior to "running" a factor analysis.

FA is a principal procedure for assessing the internal structure of an index which is said to reflect a construct. If it is established that the internal structure of the index is consistent with the definition of the construct, then there is at least preliminary evidence that the theoretical construct is valid. Accordingly, Pedhazur & Schmelkin (1991, p. 621) have cautioned, "there is little to be gained and much to be lost by applying FA in the absence of any notions regarding the structure of the phenomena studied." More bluntly, they assert, "when you have no theoretical rationale for doing a *FA, DON'T*!" (p. 591).

In our example, the variable "health education competence" was theorized to have four dimensions which required four separate unidimensional subtests that when combined produced a multidimensional index. If subtest unidimensionality had been achieved, then items would load (i.e., aggregate) on their corresponding factor (in this FA application, we may think of each subtest as a factor) as expected. Failure of items to aggregate as expected, indicates: (a) subtest unidimensionality has not been achieved (called misspecification), (b) scores can not be

meaningfully computed, and (c) revision is needed.

Extracting factors. Once the variable (e.g., construct, etc.) dimensions have been operationally defined and items written to describe the dimension, data are collected and entered into a computer program such as SPSS or SAS. Once entered, the method of factor extraction must be specified. Virtually all FA software programs require that the number of factors to be investigated (extracted) be specified. Pedhazur and Schmelkin (1991, p. 595) recommend that the number of factors to be extracted be "consistent with the theoretical formulation." The theory underlying the example suggested four factors; thus, four factors were specified. While Kline (1994, p. 29) describes eigenvalue usage as an alternative procedure, the Pedhazur and Schmelkin recommendation is endorsed. It is often practicable to conduct the initial extraction and view the resulting scree plot before specifying the number of factors to be extracted prior to rotation.

Rotating factors. Once, the number of factors are specified, the method of factor rotation is selected. When there is reason to believe that the factors are uncorrelated ($\leq .3$) use an orthogonal rotation such as varimax (Nunnally & Bernstein, 1994, p. 501), otherwise use an oblique rotation, e.g., direct oblimin. Most FA output presents a factor pattern matrix, a factor structure matrix, and a factor correlation matrix. For an orthogonal rotation, the pattern matrix consists of factor loadings whereas the structure matrix contains the correlations between each variable (e.g., item) and factor (Crocker & Algina, 1986, p. 293; Pedhazur & Schmelkin, 1991, p. 602; Nunnally & Bernstein, 1994, 470). When an orthogonal rotation is used, the factor pattern and factor structure matrices are identical. However, with oblique rotations, the two matrices can (and usually) do differ materially. For an oblique rotation, interpret the pattern

matrix as done in an orthogonal rotation. The factor correlation matrix presents the correlations among factors. When inter-factor correlations exceed ≥ .7, an FA is not recommended (Nunnally & Bernstein (1994, p. 501). However, Mahoney, Thombs, and Howe suggest an inter-factor correlation ceiling ≥ .8.

Factor loadings: General. Presented in Table 2 is a FA summary using varimax rotation with principal axis factoring as the method of extraction with four factors specified where there are three columns: factor, loading, and communality. Once the factor analysis is "run", the factor (i.e., item) loadings on each factor are inspected. The criteria for a meaningful factor load varies, the most common being .3 (Crocker & Algina, 1986, p. 299), .4, or .5 (Pedhazur & Schmelkin (1991, p. 68). With respect to index construction (regardless of rotation type), factor loadings above a cutoff (e.g., .4) maybe included in the next iteration of the index or subtest. The absolute value of a factor loading is of interest (Norusis, 1994, p. 69). If an item loads on more than one factor, then the item is grouped with the factor yielding the highest factor loading (Kline 1994, p. 132). However, when an item loads significantly on more than one factor (e.g., a loading ≥ .4), refrain from using the item or variable (Kline, 1993, p. 143).

Table 2: Item Factor Loadings and Communalities

Factor 1: Evaluation Research		**Loading**	**Communality**
3.	Selecting Educational Media	.4749	.3386
16.	Basic Program Evaluation Theory	.7237	.6303
17	Qualitative Evaluation Theory	.7285	.6746
18.	Quantitative Evaluation Theory	.6664	.6006
19.	Sampling Techniques	.7864	.6802
20.	Data Gathering Methods	.7945	.7235
21.	Basic Statistical Procedures	.6904	.6197
22.	Basic Epidemiology Methods	.5017	.3927
23.	Report Writing [Evaluation]	.4158	.3082

Factor 2: Program Planning and Management	Loading	Communality
4. Community Organizing	.4315	.4731
10. Developing Work Plans	.4547	.5799
11. Developing Staffing Plans	.6048	.6278
12. Developing Budgets	.7794	.7473
13. Planning Document Writing	.5137	.5308
14. Funding Research	.5921	.5635
24. Basic Management Theory	.6327	.6310
25. Negotiating Techniques	.6078	.6447
26. Coordinating Resources	.6725	.5298
27. Using Information Resources	.4254	.3280
28. Basic Marketing Strategies	.6374	.5448
29. Working With Power Groups	.6842	.6022
30. Power Structure Analysis	.5706	.5746

Factor 3: Clinical Skills and Knowledge	Loading	Communality
1. Health Behavior Theories	.7665	.6202
2. Selecting Interventions	.4507	.3892
5. Teaching Methods	.5370	.4554
6. Behavior Change Techniques	.7641	.6848
7. Learning Theory	.7706	.6483

Factor 4: Clinical Management	Loading	Communality
8. Case Management	.7282	.5596
9. Establishing Service Priorities	.7641	.6848
15. Community Health Analysis	.4731	.4657

Factor loadings: Orthogonal rotation. For an orthogonal rotation, a factor loading may be thought of as a correlation between an item and a factor. Thus for Item 3, the factor loading of .4749 explains 22.6% ($.4749^2$) of the variance in Item 3 (Pedhazur & Schmelkin, 1991, p. 602; Kline, 1994, p. 52).

Factor loadings: Oblique rotation. Supposing Item 3 was the product of an oblique rotation, the factor loading interpretation would be different. The loading would interpreted as a partial standardized regression coefficient. The Item 3 loading of .4749 is the effect of Factor 1 on Item 3 partialing out (i.e., controlling for) the other three factors (Crocker & Algina,1986, p.

248; Pedhazur & Schmelkin, 1991, p. 616). Standardized regression coefficients can exceed 1.0 if the factors are highly correlated. The closer to 1.0, the greater the effect of the factor on the variable, the closer to zero the lessor the effect. Crocker and Algina (p. 293) remind us, "a factor loading in an oblique solution is not equivalent to a partial correlation". When a correlation (including a partial) is squared, the percent of the variance explained by the relationship is revealed (Pedhazur & Schmelkin, p. 425).

Commonality estimates. Since, "FA is aimed at explaining common variance (i.e., variance shared by the indicators, items, variables [etc.]" (p. 598), Pedhazur and Schmelkin (1991) define communality (i.e., common variance) as, "the proportion of variance in the indicator [item] that is accounted for by the extracted factors" (p. 600). Crocker and Algina (p. 295) agree. Kline (1994) offers a similar definition and goes on to observe, "[t]he higher the communality the more the particular set of [extracted] factors explain the variance of the variable" (p. 37). Common variance by definition excludes unique (i.e., specific and error) variance; as well, it is always less than 1.0 (Crocker & Algina, 1986, p. 295). In orthogonal rotations, the communality of a variable is obtained by squaring the variable's factor loadings across each extracted factor and summing the products. Thus, for Item 3, the presence of the four extracted factors, combined, explain 33.86% of Item 3's variance. Communality values are computed differently for oblique rotations (Crocker & Algina, p. 295; Pedhazur & Schmelkin, p. 617) but mean the same. Concerning index construction, items with low communality values are usually dropped, since they have little in common with the variable as operationally defined by the extracted factors.

Application: Interpretation. Returning to our example, it is clear that four factors were

identified but not the four expected. Clearly there has been "some" model misspecification and unidimensionality has not been achieved. With respect to factor 1, Item 3 seems not to belong given the content of the other items. If the factor loading cutoff had been .5 instead of .4, Items 3 and 23 would not be presented. Factor 2 is actually a combination of the planning and management skills subtests, with the substitution of Item 4 for Item 9. It would appear that the clinical skills/content subtest was the most poorly specified. Items expected to comprise it were distributed across all four factors. Factor 1 (i.e., the program evaluation subtest) was the most properly specified of the four subtests (see Table 1). Based on these results, it is clear that the index subtests must be revised as no meaningful scores can be computed. This lack of dimensionality would have not been revealed if only item analysis had been conducted.

Application: Revision. A recommended revision strategy is to start the revision process with the extracted factors. A minimum of three to five items need to load on a factor for it to be meaningfully interpretable (Pedhazur & Schmelkin, 1991, p. 626) so some items may need to be revised or added. A return to the relevant literature in the hopes of filling in theoretical gaps may be useful. Remember, to avoid or reduce the adverse effects of misspecification, a thorough knowledge of the theory under consideration is required as is that of related research.

Index Reliability and Validity Examination

Once constructed, an index must be subjected to reliability and validity analysis. Of special note, sample size has a direct effect on reliability and indirectly on validity. The sample from which data are collected should be representative of the population from which it is drawn. With respect to sample size, Kline states, "the more subjects, the better." He also suggests that the subject to variable (in this case, item) ratio be at least 2:1 (1994, p. 74). Nunnally (1978, p.

421) suggests a 10:1 ratio.

Reliability. Cronbach's alpha is recommended (Crocker and Algina 1986, p. 138-139) for internal consistency assessment. The coefficient of stability (Crocker and Algina 1986, pp. 133-134) is used to assess instrument reliability over time. Presented in Table 3 are index subtest descriptive statistics and reliability indices. Subtest alpha coefficients are all above .70 indicating acceptable subtest reliability (Nunnally, 1978, p. 245). The data in Table 3 are based on the item analysis presented in Table 1 where we assumed subtest dimensionality. As we have seen under FA, the assumed dimensionality was not present; so, these subtest data are both worthless and misleading. Application of factor analysis to data drawn from a theory based index can improve measurement accuracy. The reader should also note, that if Nunnally's subject to variable ratio had been followed, there would have been an insufficient FA sample size.

Table 3: Subtest Descriptive Statistics & Reliability Indices

Subtest	# Items	Mean	SD[a]	Media	Alpha	SEMb
Clinical Skills	8	3.87	.535	4.00	.784	.249
Planning Skills	7	3.64	.667	3.71	.814	.288
Evaluation Skills	8	3.44	.693	3.50	.876	.244
Management Skills	7	3.56	.589	3.57	.825	.246

[a]SD =Standard Deviation [b]SEM = Standard error of the mean. The closer to zero, the less estimated measurement error.

Validity. To establish validity, theory is required (Pedhazur and Schmelkin, 1991, p.

181). If the purpose of an index is to predict, then criterion validity must be established (Pedhazur and Schmelkin, p. 33). If the purpose is to explain, then construct validity must be shown, which is the case in most sociobehavioral research (Pedhazur and Schmelkin, p. 52). Excellent detailed discussions are found in Crocker and Algina (1986, pp. 217-305) and Pedhazur and Schmelkin (1991, pp. 32-80). Establishing an index's validity across a theory (e.g., variable with or with out dimensions) requires internal and external validation. If it can be shown that the internal factor structure (as in FA) is consistent with the construct's definition, then initial evidence is available. The next step is external validation. This is more difficult and expensive than internal structure validation. External validation can be done by correlating index scores with those of already established measures. Ideally, the correlation will be statistically significant and high with results running in the same direction.

References

Babbie, E. (1990). *Survey research methods*. Belmont, CA: Wadsworth Publishing Company.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Cross, K. P. (1981). *Adults as learners*. San Francisco: Jossey-Bass.

Kerlinger, F. N. (1986). *Foundations of behavioral research.* New York: Holt, Rinehart, and Winston.

Kline, P. (1993). *The handbook on psychological testing.* London: Routledge.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Kubiszyn, T. & Borich, G. (1993). *Educational testing and measurement*. New York: Harper Collins College Publishers.

Mahoney, C. A., Thombs, D. L., & Howe, C. Z. (1995). The art and science of scale

development in health education research. *Health Education Research, 10,* (1), 1-10.

Norusis, M. J. (1994). *SPSS professional statistics 6.1*. Chicago: SPSS.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Nunnally, J. & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Oosterhof, A. (1990). *Classroom applications of educational measurement*. Columbus, OH: Merrill Publishing.

Osterlind, S. J. (1989). *Constructing test items*. Boston: Kluwer Academic Publishers.

Pedhazur, E. J. & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Udinsky, B. F., Osterlind, S. J., & Lynch, S. W. (1981). *Evaluation resource handbook: Gathering, analyzing, [and] reporting Data*. San Diego: Edits Publishers.