Guides for the Selection and Construction of Social Scales and Indexes

In: Handbook of Research Design & Social Measurement

By: Delbert C. Miller & Neil J. Salkind

Pub. Date: 2011
Access Date: August 20, 2019
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780761920465
Online ISBN: 9781412984386
DOI: https://dx.doi.org/10.4135/9781412984386
Print pages: 327-346

Guides for the Selection and Construction of Social Scales and Indexes

Scaling techniques play a major role in the construction of instruments for collecting standardized, measurable data. Scales and indexes are significant because they provide quantitative measures that are amenable to greater precision, statistical manipulation, and explicit interpretation. Before constructing a new scale, however, it is important to conduct a very careful survey of the literature to ascertain if an appropriate scale already is available to measure the dependent or independent variables in a given study. The general rule is this: The available scale should be used if it has qualities of validity, reliability, and utility (in that order of priority). With such a scale, comparative and accumulative research is possible. The need to develop a new scale can almost be considered a disciplinary failure unless the variable represents a factor never before considered as open to measurement. This discussion begins, therefore, at the point at which the literature has not revealed an appropriate scale and the researcher decides to construct an index or scale.

How does one “think up” a number of indicators to be used in empirical research? This question is answered by Paul F. Lazarsfeld and Morris Rosenberg (1962) as follows:

“

The first step seems to be the creation of a rather vague image or construct that results from the author's immersion in all the detail of a theoretical problem. The creative act may begin with the perception of many disparate phenomena as having some underlying characteristic in common. Or the author may have observed certain regularities and is trying to account for them. In any case, the concept, when first created, is some vaguely conceived entity that makes the observed relations meaningful. Next comes a stage in which the concept is specified by elaborate discussion of the phenomena out of which it emerged. We develop “aspects,” “components,” “dimensions,” or similar specifications. They are sometimes derived logically from the overall concept, or one aspect is deduced from another, or empirically observed correlations between them are reported. The concept is shown to consist of a complex combination of phenomena, rather than a simple and directly observable item. In order to incorporate the concept into a research design, observable indicators of it must be selected. (p. 15)

”

The terms *indexes* and *scales* are often used interchangeably to refer to all sorts of measures, absolute or relative, single or composite, the product of simple or elaborate techniques of measurement.

Indexes may be very simple. For example, one way to measure morale is to ask the direct question, “How would you rate your morale? Very good, good, fair, poor, or very poor?” This might be refined slightly so that the responses are placed on a numerical scale. Note that there are nine points on the following scale.

very good		*good*		*fair*		*poor*		*very poor*
1	2	3	4	5	6	7	8	9

The basis for construction is logical inference, and the use of a numerical scale requires the assumption of a psychological continuity that the respondent can realistically act upon in self-rating. Face validity usually is asserted for such a scale, although it would be possible to make tests of relations with criteria such as work performance, absenteeism, lateness, amount of drinking, and hours of sleep.

A composite index is one of a set of measures, with each composite index formed by combining simple indexes. For example, morale may be considered as a composite of many dimensions. Four measures can be combined by such questions as the following:

> How satisfied are you with your job?
> How satisfied are you with your company or organization?
> How satisfied are you in your personal life?
> How satisfied are you with your community?

Response choices of *very good, good, fair, poor*, and *very poor* may be offered for each question, with respective weights of 5, 4, 3, 2, and 1. A range from 4 to 20 points is possible. Such a composite index may improve precision, reliability, and validity.

Rigor is introduced as greater attention is paid to tests of validity and reliability. At a certain point, a given means of measurement reaches its limit of improvement, and a more refined technique becomes necessary for greater precision. Many scaling techniques concern themselves with linearity and equal intervals or equal-appearing intervals. This means that the scale follows a straight-line model and that a scoring system is devised, preferably based on interchangeable units and subject to statistical manipulation. This is a major attribute of the Thurstone attitude scaling technique.

Unidimensionality or homogeneity is another desired attribute. A scale that is unidimensional or homogeneous measures only one dimension and not some mixture of factors. This is a prime concern of the Guttman scaling technique. Reproducibility is a characteristic that enables the researcher to predict the pattern of a respondent's answers by knowing only the total scale score. This attribute is built into Guttman scaling techniques.

The intensity of feeling is introduced in the Likert technique. The respondent usually is asked to indicate his or her feelings on a 5-point scale ranging from *strongly agree* to *strongly disagree*. Tests of item discrimination are applied.

There is no single method that combines the advantages of all these techniques. It is therefore important that we understand their respective purposes and the differences between them.

Reference

Edited by: **Lazarsfeld, Paul F.**, & **Rosenberg, Morris.** (Eds.). **(1962).** The language of social research: A

reader in the methodology of social research.Glencoe, IL: Free Press.

Fukuhara, S.Keller, S. D.Kaasa, J. E.Ware, J. E., Jr.Leplege, A.Gandek, B.Sanson-Fisher, R. WAaronson, N. K.Sullivan, M.Alonso, J.Wood-Dauphinee, S.ApoloneG.Bjorner, J. B.Brazier, J.Bullinger, M. Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA Project. *Journal of Clinical Epidemiology*51933–944.(1998).

Lattin, James M. A minimum-cost network-flow solution to the Case V Thurstone scaling problem. *Psychometrika*55353–370.(1990).

Schriesheim, Chester A.Novelli, Luke, Jr. A comparative test of the interval-scale properties of magnitude estimation and Case III scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement*4959–74.(1989).

Yen, Wendy M. The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*23299–325.(1986).

5.18.1 Thurstone Equal-Appearing Interval Scale

Nature: This scale consists of a number of items whose positions on the scale have been determined previously by a ranking operation performed by judges. The subject selects the responses that best describe how he or she feels.

Utility: This scale approximates an interval level of measurement. This means that the distance between any two numbers on the scale is of known size. Parametric and nonparametric statistics may be applied. See part 6 for more information about such statistics.

Construction:

The investigator gathers several hundred statements conceived to be related to the attitude being investigated.
A large number of judges (50-300) independently classify the statements in 11 groups, ranging from *most favorable* to *neutral* to *least favorable*.
The scale value of a statement is computed as the median position to which it is assigned by the group of judges.
Statements that have too broad a spread are discarded as ambiguous or irrelevant.
The scale is formed by selecting items that are evenly spread along the scale from one extreme to the other.

Examples

?

Duggan, Ashley, Hess, Brian, Morgan, Deanna, Kim, Sooyean, & Wilson, Katherine. (1999, April). *Measuring students' attitude toward educational use of the Internet*. Paper presented at the annual conference of the American Educational Research Association, Montreal, Canada.

Student attitudes toward the Internet were investigated in a study designed to develop an instrument that would provide a quantitative measure of the attitudes undergraduates have toward educational uses of the Internet. The study also investigated some behavioral correlates of student attitudes. The responses of 395 undergraduates to some form of the scale were used to construct the measure. Statements soliciting attitudes toward educational use of the Internet were written in two formats: the Thurstone equal-appearing interval scale and the Likert-type summated rating scale. These pilot scales were administered with a social desirability response scale to ensure that students did not respond to scale items in a socially desirable manner. The final form, administered to 188 students, was an 18-item Likert-format "Attitude Toward Educational Uses of the Internet" (ATEUI) scale that yielded a high internal consistency. Several behavioral correlates lent some credence to the scale's construct validity. Favorable attitudes were associated with (a) keeping track of valuable educational Internet sites; (b) sharing information found on the Internet with friends; (c) choosing classes that use the Internet; (d) greater frequency of Internet use, both in general and for educational purposes; (e) a greater number of reasons for using the Internet in education; and (f) a greater number of Internet features used. There were no differences between men and women or in class standing in ATEUI responses. Future research that considers using the ATEUI should continue to obtain new behavioral correlates ofthe domain.

?

Fukuhara, S., Keller, S. D., Kaasa, J. E., Ware, J. E., Jr., Leplege, A., Gandek, B., Sanson-Fisher, R. W, Aaronson, N. K., Sullivan, M., Alonso, J., Wood-Dauphinee, S., Apolone G., Bjorner, J. B., Brazier, J., & Bullinger, M. (1998). Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA Project. *Journal of Clinical Epidemiology, 51*, 933-944.

The similarity in meaning assigned to response choice labels from the SF-36 Health Survey (SF-36) was evaluated across countries. Convenience samples of judges (range = 10 to 117; median = 48) from 13 countries rated translations of response choice labels, using a variation ofthe Thurstone method of equal-appearing intervals. Judges marked a point on a 10-cm line representing the magnitude of a response choice label (e.g., *good* relative to the anchors of *poor* and *excellent)*. Ratings were evaluated to determine the ordinal consistency of response choice labels within a response scale, the degree to which differences between adjacent response choice labels were equal interval, and the amount of variance that was due to response choice label, country, judge, and interaction between response choice label and country. Results confirmed the hypothesized ordering of response choice labels. The percentage of ordinal pairs ranged from 88.7% to 100% (median = 98.2%) across countries and response scales. Examination ofthe average magnitudes of response choice labels supported the "quasi-interval" nature of the scales. Analysis of variance (ANOVA) results supported the generalizability of

response choice magnitudes across countries; labels explained 64% to 77% of the variance in ratings, and country explained 1% to 3%. These results support the equivalence of SF-36 response choice labels across countries. Departures from the assumption of equal intervals, when observed, were similar across countries and were greatest for the two response scales that were recalibrated under standard SF-36 scoring. Results provide justification for scoring translations of individual items using standard SF-36 scoring. Whether these items form the same scales in other countries as they do in the United States was evaluated with tests of scaling assumptions.

?

Lattin, James M. (1990). A minimum-cost network-flow solution to the Case V Thurstone scaling problem. *Psychometrika, 55*, 353-370.

Presents an approach for determining unidimensional scale estimates that are insensitive to limited inconsistencies in paired comparisons data. The solution procedure, shown to be a minimum-cost network-flow problem, is presented in conjunction with a sensitivity diagnostic that assesses the influence of a single pairwise comparison on traditional Thurstone scale estimates. When distortion was indicated in the data, the network technique appeared to be more successful than Thurstone scaling in preserving the interval scale properties of the estimates.

?

Schriesheim, Chester A., & Novelli, Luke, Jr. (1989). A comparative test of the interval-scale properties of magnitude estimation and Case III scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement, 49*, 59-74.

Differences between recommended sets of equal-interval response anchors derived from scaling techniques using magnitude estimations and Thurstone Case III pair-comparison treatment of complete ranks were compared. Differences in results for 205 undergraduates reflected differences in the samples as well as in the tasks and computational algorithms.

?

Yen, Wendy M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299-325.

Two methods of constucting equal-interval scales for educational achievement are discussed: Thurstone's absolute scaling method and item response theory. Alternative criteria for choosing a scale are contrasted. It is argued that clearer criteria are needed for judging the appropriateness and usefulness of alternative scaling procedures.

5.18.1.1 Readings on Thurstone Equal-Interval Scales

Duggan, Ashley, **Hess, Brian**, **Morgan, Deanna**, **Kim, Sooyean**, & **Wilson, Katherine.(1999, April).** Measuring students' attitude toward educational use of the Internet. Paper presented at the annual conference of the American Educational Research Association, Montreal, Canada.

Fukuhara, S.Keller, S. D.Kaasa, J. E.Ware, J. E., Jr.Leplege, A.Gandek, B.Sanson-Fisher, R. WAaronson, N. K.Sullivan, M.Alonso, J.Wood-Dauphinee, S.ApoloneG.Bjorner, J. B.Brazier, J.Bullinger, M. Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA Project. *Journal of Clinical Epidemiology*51933–944.(1998).

Lattin, James M. A minimum-cost network-flow solution to the Case V Thurstone scaling problem. *Psychometrika*55353–370.(1990).

Schriesheim, Chester A.Novelli, Luke, Jr. A comparative test of the interval-scale properties of magnitude estimation and Case III scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement*4959–74.(1989).

Yen, Wendy M. The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*23299–325.(1986).

5.18.2 Likert-Type Scale

Nature: This is a summated scale consisting of a series of items to which the subject responds. The respondent indicates agreement or disagreement with each item on an intensity scale. The Likert technique produces an ordinal scale that generally requires nonparametric statistics.

Utility: This scale is highly reliable when it comes to a rough ordering of people with regard to a particular attitude or attitude complex. The score includes a measure of intensity as expressed on each statement.

Construction:

The investigator assembles a large number of items considered relevant to the attitude being investigated and clearly either favorable or unfavorable.
These items are administered to a group of subjects representative of those with whom the questionnaire is to be used.
The responses to the various items are scored in such a way that a response indicative of the most favorable attitude is given the highest score.
Each individual's total score is computed by adding his or her item scores.
The responses are analyzed to determine which items differentiate most clearly between the highest and lowest quartiles of total scores.
The items that differentiate best (at least six) are used to form a scale.

Examples

?
Belmonte-Serrano, M. A., Beltran, Fabregat J., & Paz, Furio M. (1996). A comparative study of

HAQ questionnaires—versions of 20 and 8 items—with Likert scale and visual analogue scale in rheumatoid arthritis patients. *Revista Espanola de Reumatologia, 23*, 83-88.

The purposes of this study were to perform a cross-validation of the Spanish versions of the disability scale of the Stanford Health Assessment Questionnaire (HAQ) versus Pincus's reduced version of 8 items (MHAQ) and to compare the visual analogue scale (VAS) against the verbal ordinal type (Likert) as instruments to assess pain and global health in rheumatic patients. A questionnaire with the 20 items of the HAQ and both types of scales was given to rheumatoid arthritis patients. The value of the 8 items of the MHAQ was obtained from the original data, as a subset of the HAQ itself. The HAQ and MHAQ showed an excellent internal consistency (alpha = .9) and good criterion and construct validity. The HAQ and MHAQ were highly and significantly correlated, both regarding total scores (r = .88, $p < .001$) and for each of the subscales (r values ranging from .57 to .87, $p < .001$). The mean score for the MHAQ was 31.2% lower than that obtained with the HAQ (p < .001). The study found a significant, moderate, and similar correlation of the HAQ and the MHAQ with the Steinbrocker functional capacity and both types of scales measuring pain and health status. The study of the pain and global health scales showed good correlation between Likert and VAS types for both variables (r = .77 and $r = .50$, $p < .001$). Likert and VAS scales had a highly linear relationship (p < .0001). In general, it was found that the MHAQ is as suitable as the HAQ; however, the marked difference of mean scores between the two questionnaires makes it impossible to use them indistinctly. The significant correlation and high linearity found between Likert and VAS scales, both for pain and for global health status, suggests that it is possible to use the equivalency values of the Likert scales as a surrogate for missing values in the VAS scales.

?

Cheung, K. C., & Mooi, L. C. (1994). A comparison between the rating scale model and dual scaling for Likert scales. *Applied Psychological Measurement, 18*, 1-13.

Problems relating to existence of interval scales, dimensionality of a trait, and patterns of item response functions were approached through contrasting scaling methods: item response theory modeling and dual scaling. Similarity of the methods was established through a study of 326 female junior college students in Singapore who completed a Likert-type scale.

?

Gu, Yongqui, & others. (1995). How often is often? Reference ambiguities of the Likert-scale in language learning strategy research. *Occasional Papers in English Language Teaching*, 5,19-35.

This article, based on personal experience, examines the ambiguities of the Likert-type 5-point scale in learning strategy elicitation. Four parallel questionnaires consisting of the same batch of 20 items taken from the Oxford scale (1990) were administered among a group of 120 tertiary level, non-English majors in China. Questionnaire 1 used the Oxford scale without specifying dimensions of reference. Questionnaire 2 told the respondents to choose their answers by comparing with their peers in the same grade. Questionnaire 3 asked them to select their present behavioral frequency as compared with their own past learning experience in secondary

schools. In questionnaire 4, subjects were told to check off the relevant frequency of a behavior by comparing its frequency of occurrence with that of other language skills. Results showed that out of the 20 items used, 13 were significantly different among the four questionnaires. Methodological implications for questionnaire research are discussed in the article, and suggestions for future research are proposed.

?

Hassan, Abdel Moneim Ahmed, & Shrigley, Robert L. (1984). Designing a Likert scale to measure chemistry attitudes. *School Science and Mathematics, 84*, 659-669.

This article brings together the principles of designing Likert-type attitude scales and demonstrates the procedure through the development of a chemistry attitude scale. Design procedures and validity and reliability of the scale are each discussed, with a data summary for the 20-item chemistry scale.

?

Smith-Sebasto, N. J., & D'Costa, Ayres. (1995). Designinga Likert-type scale to predict environmentally responsible behavior in undergraduate students: A multistep process. *Journal of Environment Education, 27*, 14-20.

Describes an attempt to develop a reliable and valid instrument to assess the relationship between locus of control of reinforcement and environmentally responsible behavior. Presents a six-step psychometric process used to develop the Environmental Action Internal Control Index (EAICI) for undergraduate students. Contains 54 references.

5.18.2.1 Readings on Likert-Type Scales

Belmonte-Serrano, M. A.Beltran, Fabregat J.Paz, Furio M. A comparative study of HAQ questionnaires—versions of 20 and 8 items—with Likert scale and visual analogue scale in rheumatoid arthritis patients. *Revista Espanola de Reumatologia*2383–88.(1996).

Cheung, K. C.Mooi, L. C. A comparison between the rating scale model and dual scaling for Likert scales. *Applied Psychological Measurement*181–13.(1994).

Gu, Yongqui, others. How often is often? Reference ambiguities of the Likert-scale in language learning strategy research. *Occasional Papers in English Language Teaching*519–35.(1995).

Hassan, Abdel Moneim AhmedShrigley, Robert L. Designing a Likert scale to measure chemistry attitudes. *School Science and Mathematics*84659–669.(1984).

Smith-Sebasto, N. J.D'Costa, Ayres. Designinga Likert-type scale to predict environmentally responsible behavior in undergraduate students: A multistep process. *Journal of Environment Education*2714–20.(1995).

5.18.3 Guttman Scale Analysis

Nature: The Guttman technique attempts to determine the unidimensionality of a scale. Only items meeting the criterion of reproducibility are acceptable as scalable. If a scale is unidimensional, then a person who has a more favorable attitude than another should respond to each statement with a favorableness score equal to or greater than that of the other person.

Utility: Each score corresponds to a highly similar response pattern or scale type. It is one of the few scales where the score can be used to predict the response pattern to all statements. Only a few statements (5 to 10) are needed to provide a range of scalable responses. Note the analysis presented in Table 5.12, which shows how 14 subjects responded with a "Yes" to several statements and how scores reflect a given pattern of response.

Table 5.12 PATTERNS OF RESPONSES IN GUTTMAN SCALE ANALYSIS

Respondent	*Item 7*	*Item 5*	*Item 1*	*Item 8*	*Item 2*	*Item 4*	*Item 6*	*Item 3*	*Score*
7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	7
9	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	7
1	Yes	Yes	Yes	—	Yes	Yes	—	Yes	6
10	Yes	Yes	Yes	Yes	Yes	Yes	—	—	6
13	Yes	Yes	Yes	Yes	Yes	Yes	—	—	6
3	Yes	Yes	Yes	Yes	Yes	—	—	—	5
2	Yes	Yes	Yes	Yes	—	—	—	—	4
6	Yes	Yes	Yes	Yes	—	—	—	—	4
8	Yes	Yes	Yes	—	—	Yes	—	—	4
14	Yes	Yes	Yes	Yes	—	—	—	—	4
5	Yes	Yes	Yes	—	—	—	—	—	3
4	Yes	Yes	—	—	—	—	—	—	2
11	—	—	—	—	Yes	—	—	—	1
12	Yes	—	—	—	—	—	—	—	1

Construction:

Select statements that are felt to apply to the measurable objective.
Test statements on a sample population (about 100).
Discard statements with more than 80% agreement or disagreement.
Order respondents from those having the most favorable responses to those having the fewest favorable responses. Order from left to right.
Order statements from those having the most favorable responses to those having the fewest favorable responses. Order from left to right.
Discard statements that fail to discriminate between favorable respondents and unfavorable respondents.
Calculate coefficient of reproducibility.

a. Calculate the number of errors (favorable responses that do not fit pattern)

b. Reproducibility = 1-(number of errors/number of responses)
c. If reproducibility equals .90 or greater, a unidimensional scale is said to exist.

Score each respondent by the number of favorable responses or response patterns.

Examples

?

Burgin, Robert. (1989). Guttman scale analysis: An application to library science. *Library and Information Science Research, 11*, 47-57.

Outlines the general techniques of Guttman scale analysis and briefly describes its uses in social science research. To illustrate the potential application to library science, a Guttman scale of restrictiveness in dealing with overdue books was developed, and data from a 1986 survey of public libraries were fitted into the scale.

?

Kramer, Deirdre A. (1983, November). *A developmental investigation of relativistic and dialectical thought*. Paper presented at the Annual Scientific Meeting of the Gerontological Society, San Francisco.

Post-formal operational thought is characterized by both relativism and dialecticism. To examine age differences across adulthood in relativistic and dialectical thought, and to determine whether formal operations are necessary but not sufficient for these forms of thought, 20 young (mean age, 19.6), 20 middle-aged (mean age, 46.2), and 20 older (mean age, 68.5) adults were administered three cognitive tasks. The Ammons Quick Test was administered to determine the presence of comparable verbal intelligence. Subsequently, subjects were administered four formal operations tasks: separation of variables; three measures of coordination of two frames of reference; and two lifelike dilemmas, to which they were asked to react. Reactions to the dilemmas were placed into four categories of thought (formalistic-mechanistic, relativistic, awareness of contradictions, and integration of contradictions into a dialectical whole). Analysis of the results showed that older adults scored significantly higher on the Quick Test than young adults, with middle-aged adults falling between those two groups. On formal operations tasks, performance was intact across adulthood. On the lifelike dilemmas, older adults showed significantly less rejection and more acceptance of relativistic and dialectical thought. Guttman scale analysis showed that formal operations were necessary but not sufficient for dialectical thought. The findings provide potential support for the hypothesis that dialectical thought is post-formal operational.

?

Lange, A., Kooiman, K., Huberts, L., & Van Oostendorp, E. (1995). Childhood unwanted sexual events and degree of psychopathology of psychiatric patients: Research with a new anamnestic questionnaire (the CHUSE). *Acta Psychiatrica Scandinavica, 92*, 441-446.

By means of a recently constructed anamnestic instrument, the Childhood Unwanted Sexual

Events (CHUSE) questionnaire, the incidence of childhood experiences with sexual threat and/or abuse was investigated among 152 female psychiatric patients. The construction and applicability of the questionnaire are described. A Guttman scale analysis showed a unidimensional construct (severity of the sexual abuse) for the CHUSE. Within this psychiatric population, sexually abused women reported significantly more psychopathological symptoms than nonabused women. The correlation between severity of the abuse and severity of the psychopathological symptoms was investigated. The use of questionnaires concerning sexual abuse was compared with the more common interview techniques. Suggestions for future research are given.

?

Stempel, Guiod H., III. (1982). A Guttman scale analysis of the Burger Court's press decisions. *Journalism Quarterly, 59*, 256-259.

Analyzes Supreme Court votes on 47 press-related cases and shows that they form a scalable universe that is unidimensional. The author suggests that the dimension is political predisposition.

?

Von Korff, M., Ormel, J., Keefe, F. J., & Dworkin, S. F. (1992). Grading the severity of chronic pain. *Pain, 50*, 133-149.

This research develops and evaluates a simple method of grading the severity of chronic pain, for use in general population surveys and studies of primary care pain patients. Measures of pain intensity, disability, persistence, and recency of onset were tested for their ability to grade chronic pain severity in a longitudinal study of primary care back pain **(n** = 1,213), headache **(n** = 779), and temporomandibular disorder pain **(n** = 397) patients. A Guttman scale analysis showed that pain intensity and disability measures formed a reliable hierarchical scale. Pain intensity measures appeared to scale the lower range of global severity, whereas disability measures appeared to scale the upper range of global severity. Recency of onset and days in pain in the prior 6 months did not scale with pain intensity or disability. Using simple scoring rules, pain severity was graded into four hierarchical classes: Grade I, low disability-low intensity; Grade II, low disability-high intensity; Grade III, high disability-moderately limiting; and Grade IV, high disability-severely limiting. For each pain site, Chronic Pain Grade measured at baseline showed a highly statistically significant and monotonically increasing relationship with unemployment rate, pain-related functional limitations, depression, fair to poor self-rated health, frequent use of opioid analgesics, and frequent pain-related doctor visits both at baseline and at 1-year follow-up. Days in pain was related to these variables, but not as strongly as Chronic Pain Grade. Recent onset cases (first onset within the prior 3 months) did not show differences in psychological and behavioral dysfunction when compared to those with less recent onset. Using longitudinal data from a population-based study **(n** = 803), Chronic Pain Grade at baseline predicted the presence of pain in the prior 2 weeks as well as Chronic Pain Grade and pain-related functional limitations at 3-year follow-up. Grading chronic pain as a function of pain intensity and pain-related disability may be useful when a brief ordinal measure of global pain severity is required. Pain persistence, measured by days in pain in a fixed time period, provides useful additional information.

5.18.3.1 Readings on Guttman Scales

Burgin, Robert. Guttman scale analysis: An application to library science. *Library and Information Science Research*1147–57.(1989).

Gordon, Raymond L.(1977). Unidimensional scaling of social variables.Riverside, NJ: Free Press.

Kramer, Deirdre A.(1983, November). A developmental investigation of relativistic and dialectical thought. Paper presented at the Annual Scientific Meeting of the Gerontological Society, San Francisco.

Lange, A.Kooiman, K.Huberts, L.Van Oostendorp, E. Childhood unwanted sexual events and degree of psychopathology of psychiatric patients: Research with a new anamnestic questionnaire (the CHUSE). *Acta Psychiatrica Scandinavica*92441–446.(1995).

Lin, Nan.(1976). Foundations of social research.New York: McGraw-Hill.

Stempel, Guiod H., III. A Guttman scale analysis of the Burger Court's press decisions. *Journalism Quarterly*59256–259.(1982).

Von Korff, M.Ormel, J.Keefe, F. J.Dworkin, S. F. Grading the severity of chronic pain. *Pain*50133–149.(1992).

5.18.4 Scale Discrimination Technique

Nature: This technique seeks to develop a set of items that meet the requirements of a unidimensional scale, possess equal-appearing intervals, and measure intensity. Aspects of the construction of Thurstone's equal-appearing intervals, Likert's summated scales, and Guttman's scale analysis are combined in this technique, developed by Edwards and Kilpatrick.

Utility: Three distinct advantages of separate scaling techniques are combined. The interval scale quality of the Thurstone technique can be achieved. The discriminability between respondents and the addition of an intensity measure are derived from the Likert technique, and unidimensionality from the Guttman technique. Caution: Item analysis will eliminate items in the middle of the scale.

Construction:

The investigator selects a large number of statements that are thought to apply to the attitude being measured.
Items that are ambiguous or too extreme are discarded.
The statements are given to judges, who evaluate the favorableness of each statement and place it in 1 of 11 categories.
Half of the items with the greatest scatter, or variance, are discarded.

Scores are assigned to the remaining items as the median of the judges' scores.
The statements are devised in the form of a summated scale and given to a new set of judges.
An item analysis is performed to determined which questions discriminate best between the lowest and highest quartiles.
Twice the number of items that are wanted in the final scale are selected. From each scale interval, the statements that discriminate best are selected.
These statements are divided in half, and the halves are submitted to separate test groups.
Coefficients of reproducibility are determined for each test group; those that are .90 or above are used.

Examples

?
Crist, D. A., Rickard, H. D., Prentice, Dunn S., & Barker, H. R. (1989). The Relaxation Inventory: Self-report scales of relaxation training effects. *Journal of Personality Assessment, 53*, 716-726.
The development of a self-report measure to assess the effects of relaxation training was examined. A rigorous statistical method of scale construction consisting of a modification of the scale discrimination technique was employed, resulting in a 45-item questionnaire representing three orthogonally derived scales. The three scales—physiological tension, physical assessment, and cognitive tension—demonstrated adequate internal consistency with KR20 reliability coefficients of .89, .95, and .81, respectively. In a second study of predictive validity, 40 individuals were randomly assigned to one of four conditions: relaxation training, tension inducement, pre-post control, or postcontrol. Univariate analysis of variance indicated significant findings for each of the three dimensions of the inventory. The physiological tension scale detected significant increases in tension following tension inducement, whereas the physical assessment scale and cognitive tension scale detected increases in relaxation following relaxation training. Recommendations were made for future research on the inventory.
?
Foster, Don. (1991). *Social psychology in South Africa*. Johannesburg, South Africa: Lexicon.
Most of the basic methods of attitude measurement were developed in the United States during the 1920s and 1930s. They include the social distance scale, Thurstone's equal-appearing interval scale, the Likert form of scaling, and Katz and Braly's assessment of stereotypes. The scale discrimination technique was developed in the 1940s. It was only later that another popular method, the semantic differential, was developed. Most of these methods have been used in South Africa, and each of them are discussed in turn, using, where applicable, a South African example of the method.
?
Killian, Kieran, Watson, Richard, Otis, Joceline, St. Amand, Timothy A., & O'Byrne, Paul M. (2000). Symptom perception during acute bronchoconstriction. *American Journal of Respiratory and Critical Care Medicine, 162*, 490-496.

The hypothesis underlying the study was that some of the variability in symptom intensity seen during acute bronchoconstriction may result from varying intensities of several stimuli, yielding several sensations that can be identified by specific descriptive expressions (symptoms). A total of 232 subjects inhaled methacholine in doubling concentrations to a 20% decrease in FEV1, or 64 mg/ml. The study identified the prevalence of dyspnea, nonspecific discomfort associated with the act of breathing, and 10 specific symptom expressions. Each symptom intensity was rated in Borg scale units. The contribution of the specific symptoms to the intensity of dyspnea is illustrated in the following equation (r = 0.84): Dyspnea = 0.44 + 0.19 Difficult breathing + 0.41 Chest tightness + 0.20 Breathlessness + 0.14 Labored breathing + 0.11 Chest pain. Dyspnea was more intense with bronchoconstriction, baseline pulmonary impairment, weight, and sex (being female). Dyspnea was less intense with age (being older) and as airway responsiveness to methacholine increased ($p < 0.05$ for all factors). Chest tightness and chest pain were at polar extremes on the discrimination scale (i.e., easily discriminated); chest tightness and difficult and labored breathing were not easily discriminated.

?

Velozo, C. A., Magalhaes, L. C., Pan, A. W, & Leiter, P (1995). Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. *Archives of Physical Medicine and Rehabilitation, 76*, 705-712.

The purpose of this study was to determine the construct validity of the Level of Rehabilitation Scale-III (LORS-III) with a special focus on this instrument's capability to discriminate rehabilitation inpatient activities of daily living (ADL)/mobility and communication/cognition ability at admission and discharge. Rasch analysis of existing data sets in the LORS-III American Data System (LADS) was performed. Existing admission and discharge data from 3,056 rehabilitation in-patients (musculo-skeletal injury, cerebrovascular accident, multiple injuries/diseases, brain injury, neuromuscular disorder, and spinal cord injury) were entered into LADS between April 1992, and January 1993. LORS-III consists of 17 measurement areas representing abilities in ADL, mobility, communication, cognition, and memory. Fourteen of the measurement areas are concurrently scored by a nurse and a specified rehabilitation therapist, resulting in a total of 31 items. Consistent with findings reported for other functional status measures, the analysis indicated that the LORS-III consists of two unidimensional scales, an ADL/mobility scale and a communication/cognition scale. Although all scales fit the Rasch measurement model, the ADL/mo-bility scale used at admission was most appropriately targeted to the ability level of the sample. At discharge, the ADL scale generally was too easy because the ability level of the sample moved upward toward functional independence. The communication/cognition scale at both admission and discharge showed a similar “ceiling” effect. These findings indicate the importance of determining the measurement qualities of functional status measures for both admission and discharge ratings. Analyses, such as Rasch, can provide a logical direction for instrument refinement.

5.18.4.1 Readings on Scale Discrimination

Crist, Dwayne A.Rickard, Henry C.Prentice, Dunn StevenBarker, Harry R. The Relaxation Inventory: Self-report scales of relaxation training effects. *Journal of Personality Assessment*53716–726.(1989).

Foster, Don.(1991). Social psychology in South Africa.Johannesburg, South Africa: Lexicon.

Killian, KieranWatson, RichardOtis, JocelineSt. Amand, Timothy A.O'Byrne, Paul M. Symptom perception during acute bronchoconstriction. *American Journal of Respiratory and Critical Care Medicine*162490–496.(2000).

Velozo, C. A.Magalhaes, L. C.Pan, A. WLeiter, P. Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. *Archives of Physical Medicine and Rehabilitation*76705–712.(1995).

5.18.5 Rating Scales

Nature: This technique, based on personal judgments, seeks to obtain an evaluation or a quantitative judgment of personality, group, or institutional characteristics. The rater places the person or object being rated at some point along a continuum or in one of an ordered series of categories. A numerical value is attached to the point or the category.

Utility: Rating scales can be used to assess attitudes, values, norms, social activities, and social structural features.

Construction:

The continuum to be measured is divided into an optimal number of scale divisions (approximately five to seven).
The continuum should have no breaks or divisions.
The positive and negative poles should be alternated.
Each trait is introduced with a question to which the rater can give an answer.
Descriptive adjectives or phrases are used to define different points on the continuum.
The investigator should decide beforehand on the probable extremes of the trait to be found in the group in which the scale is to be used.
Only universally understood descriptive terms should be used.
The end phrases should not be so extreme in meaning as to be avoided by the raters.
Descriptive phrases need to be evenly spaced.
During pretesting, the investigator asks respondents to raise any questions they have about the ratings and the different points on the continuum if they are unclear.

Assigned numerical values are used to score.

Examples

?

Linacre, John M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103-122.

Suggests eight guidelines to help an analyst investigate whether rating-scale categories are cooperating to produce observations on which valid measurement can be based. Presents these guidelines in the context of Rasch analysis and illustrates their use.

?

Varner, RoyV, Chen, Y. Richard, Swann, Alan C., & Moeller, Frederick G. (2000). The Brief Psychiatric Rating Scale as an acute inpatient outcome measurement tool: A pilot study. *Journal of Clinical Psychiatry, 61*, 418-421.

Because guidelines for length of stay at psychiatric hospitals may have an unacceptable impact on patient outcome at discharge, a valid measurement tool is needed to evaluate significant patient change during brief hospitalization (typically 7 days) and to provide early prediction of unfavorable short-term outcome. This study examines the utility of the Brief Psychiatric Rating Scale (BPRS) as such a tool. During a 2-month testing period, the BPRS was administered to 87 adults successively admitted to an acute general psychiatric inpatient unit at admission, with administrations at 2 days, 7 days, and weekly thereafter until discharge. Total BPRS scores and four subscores were used in the data analysis, which included paired *t* tests and correlation analyses. Mean BPRS total scores demonstrated significant ($p < .001$) patient improvement at days 2, 7, and 14 of the hospital stay. Changes in subscores and their relationship to eventual outcome varied across diagnostic groups. The BPRS thus appears to be a useful inpatient outcome measure because it is capable of demonstrating significant change during stays of 1 week or less. Subscale scores may provide more specific prediction of change and may help clarify outcome in individual patients who show insignificant change in total score.

?

Ward, M. M., Marx, A. S., & Barry, N. N. (2000). The rating scale preference measure as an evaluative measure in systemic lupus erythematosus. *Lupus, 9*, 696-701.

Preference measures may be useful tools to assess patients' overall health-related quality of life. In a prospective longitudinal observational study of changes in the symptoms and clinical disease activity of 23 patients, the authors studied the validity and sensitivity to change of the rating scale preference measure in patients with systemic lupus erythematosus (SLE) and compared its properties with those of the patient global assessment of SLE activity. Patients were assessed every 2 weeks for up to 40 weeks. Construct validity was assessed by the strength of correlations between changes over time in the rating scale preference measure and patient global assessment and changes in the physician global assessment, Systemic Lupus Activity Measure (SLAM), European Consensus Lupus Activity Measure (ECLAM), the British Isles Lupus

Assessment Group index (BILAG), and Systemic Lupus Erythematosus Disease Activity Index (SLEDAI). Changes in the rating scale were more highly correlated with changes in each of these standards than were changes in the patient global assessment, demonstrating the construct validity of this measure. Sensitivity to change was measured using the 2-week interval of greatest change in either the physician global assessment or the SLE activity measures as standards. The rating scale preference measure was less sensitive to change than was the patient global assessment when tested against four different standards. The sensitivity to change of the rating scale was less than half that of the patient global assessment when either the SLAM or ECLAM was used as the standard. Although these results support the validity of the rating scale as a measure of health-related quality of life in patients with SLE, its limited sensitivity to change may make it less attractive as an endpoint measure in clinical trials.

?

Wright, Benjamin D., & Masters, Geofferey N. (1982). *Rating scale analysis*. Chicago: MESA.

This book discusses the construction of variables and development of measures. It begins by outlining the qualities a number must meet before it qualifies as a measure of something. The basis is the measurement philosophy of G. Rasch. The first requirement for making good measures is good raw material. To achieve the possibility of comparisons, the data must contain the possibility of a single variable along which persons can be measured. Chapter 2 presents a set of data that must be inspected, and techniques for inspecting the data are reviewed. In chapter 3, five different models for measuring are described, each of which was developed for a particular type of data. There are other models in the measurement literature, but these, all members of a family, are used because they meet the standards set for measurement. Chapter 4 shows how to use these models to get results, describing four different estimation procedures: PROX, PAIR, UCON, and CON. The quality control of variables is discussed in chapter 5. Chapters 6, 7, and 8 then illustrate the use of the techniques discussed, using four different data sets that were collected to measure drug use, fear of crime, knowledge of elementary physics, and child development.

5.18.5.1 Readings on Rating Scales

Linacre, John M. Investigating rating scale category utility. *Journal of Outcome Measurement*3103–122.(1999).

Varner, Roy VChen, Y. RichardSwann, Alan C.Moeller, Frederick G. The Brief Psychiatric Rating Scale as an acute inpatient outcome measurement tool: A pilot study. *Journal of Clinical Psychiatry*61418–421.(2000).

Ward, M. M.Marx, A. S.Barry, N. N. The rating scale preference measure as an evaluative measure in systemic lupus erythematosus. *Lupus*9696–701.(2000).

Wright, Benjamin D., & **Masters, Geofferey N.(1982).** Rating scale analysis.Chicago: MESA.

5.18.6 Latent Distance Scales

Nature: Analysis of these scales is based on a probability model that attempts to apply to qualitative data the principles of factor analysis, providing ordinal information. The basic postulate is that there exists a set of latent classes such that the manifest relationship between any two or more items on a questionnaire can be accounted for by the existence of these latent classes and by these alone.

Utility: Latent class analysis provides a description of categorical latent (unobserved) variables from an analysis of the structure of the relationships among several categorical manifest (observed) variables. This method is commonly called categorical data analogue to factor analysis.

Construction:

The investigator lists questions believed to be related to the latent attitude.
Answers to questions are dichotomized in terms of positive-negative, favorable-unfavorable, and so on.
The proportion of respondents who demonstrate the latent attitude in each response is calculated.
Items are arranged in terms of their manifest marginals.
The latent class frequencies are computed through inverse-probability procedures.
Response patterns are ranked in terms of average latent position, or an index is used to characterize each response pattern.

Examples

?
Eshima, Nobuoki. (1991). Latent scalogram analysis. *Behaviormetrika, 20*, 1-21.
In scientific research, response structures are often more complex than those assumed under the latent distance model, an extension of the scalogram analysis proposed by Guttman (1950). Scaling models for these response structures are reviewed and compared, then expanded to develop latent scalogram analysis. Model selection procedures in both exploratory and confirmatory contexts are utilized in extracting both linear and branching hierarchical structures. Dynamic interpretation of latent scales is offered from a mathematical viewpoint, and a method for interpreting the proportions of latent scales is proposed. Numerical analysis shows the efficiency of this approach for deriving a simple latent structure based on binary data and for interpreting the extracted structure.
?
Eshima, Nobuoki, & Asano, Chooichiro. (1988). On latent distance analysis and the MLE algorithm. *Behaviormetrika, 24*, 25-32.
Discusses Lazarsfeld and Henry's (1968) proposal of latent distance analysis, in which all the binary items (yes/no) are dominated by a common factor and each individual in a population

responds to each item with a response probability, depending on the level of the latent factor. An algorithm is proposed that does not give any improper solution after transforming the latent response parameters to the logistic form and applying the new method of maximum likelihood estimation (MLE). The actual data analysis is provided.

?

Mellenbergh, Gideon J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*, 223-236.

A general linear latent trait model for continuous item responses is described. The special unidimensional case for continuous item response is K. G. Jöreskog's model of congeneric item response. The correspondence between models for continuous and dichotomous item responses is shown to be closer than usually supposed.

?

Pascual, Leone Juan, & Baillargeon, Raymond. (1994). Developmental measurement of mental attention. *International Journal of Behavioral Development, 17*, 161-200.

Presents a dialectical constructivist model of mental attention and of working memory that is used to explicate research participants' processing in misleading test items. A set of 10 theoretical structural predictions were semantically derived that stipulated relations between mental attentional resources and the varied mental demands of items, as they jointly codetermine probable performance. These predictions were evaluated using a known family of ordered latent class models. A group of 616 children (aged 5-14 years) was tested. Results show that (a) data fit Lazarsfeld's latent distance model, providing initial support for the 10 predictions; (b) the M-power of children (latent mental-power classes), when assessed behaviorally, may increase with age in a discrete manner and have the potential to generate interval scales of measurement; and (c) what statisticians often consider "error of measurement" appears (in part) to be signal, not noise.

?

Windle, Michael, & Dumenci, Leyent. (1998). An investigation of maternal and adolescent depressed mood using a latent trait-state model. *Journal of Research on Adolescence, 8*,461-484.

A study using a latent trait-state model to study aspects of maternal and adolescent depressed mood found that maternal depression was predicted by lower family income, lower family cohesion, lower perceived social support, and higher parental role stress. The study also found that adolescent-trait depression was predicted by lower perceived family support, lower grade point average, more stressful life events, and female gender.

5.18.6.1 Readings on Latent Distance Scales

Eshima, Nobuoki. Latent scalogram analysis. *Behaviormetrika*201–21.(1991).

Eshima, NobuokiAsano, Chooichiro. On latent distance analysis and the MLE algorithm.

*Behaviormetrika*2425–32.(1988).

Guttman, L.(1950). The basis for scalogram analysis. In Edited by: **S. A. Stouffer** with the Social Science Research Council (Eds.), Studies in social psychology in World War II: Vol. 4. Measurement and prediction.Princeton, NJ: Princeton University Press.

Edited by: **Langeheine, Rolf**, & **Rost, Jürgen.** (Eds.). **(1988).** Latent trait and latent class models.New York: Plenum.

Lazarsfeld, Paul F., & **Henry, Neil W(1968).** Latent structure analysis.Boston: Houghton Mifflin.

Mellenbergh, Gideon J. A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*29223–236.(1994).

Pascual, Leone JuanBaillargeon, Raymond. Developmental measurement of mental attention. *International Journal of Behavioral Development*17161–200.(1994).

Windle, MichaelDumenci, Leyent. An investigation of maternal and adolescent depressed mood using a latent trait-state model. *Journal of Research on Adolescence*8461–484.(1998).

5.18.7 Paired Comparisons

Nature: This technique seeks to determine psychological values of qualitative stimuli without knowledge of any corresponding respondent values. By asking respondents to select the more favorable of a pair of statements or objects across a set of several pairs, an attempt is made to order the statements or objects along a continuum. This is sometimes called the forced-choices technique.

Utility: Ordering by paired comparisons is a relatively rapid process for securing a precise and relative positioning along a continuum. Comparative ordering generally increases reliability and validity over arbitrary rating methods.

Construction:

The investigator selects statements that relate to the attribute being measured.
The statements are arranged in all possible combinations of pairs as follows, with the number of combinations equal to N(N-1)/2, where *N* is the number of statements.
Judges are asked to select which statement of each pair is the more favorable.
The proportion of judgments each statement received over every other statement is calculated.
The proportions are totaled for each statement.
The proportions are translated into standardized scale values.
An internal consistency check is applied by computing the absolute average discrepancy.
Statements are presented to respondents, who are asked to indicate favorableness or

unfavorableness of each statement.

A respondent's score is the median of his or her favorable responses.

Examples

?

Eisenberg, L. S., & Dirks, D. D. (1995). Reliability and sensitivity of paired comparisons and category rating in children. *Journal of Speech and Hearing Research, 38*, 1157-1167.

Children's subjective judgments of speech clarity using the methods of paired comparisons and category rating were evaluated in this investigation. Eighty children with normal hearing between the ages of 4 and 8 years judged the clarity of sentences that were systematically bandpass-filtered using conditions that increased intelligibility as estimated by the Articulation Index. Subjects were classified into four age groups (4, 5, 6, and 7-8 years olds), with 20 subjects per group. With use of materials and training methods suitable for children, judgments were obtained via the two psychophysical procedures (10 subjects per age group for each procedure). Results indicated that children 5 years of age and older were able to make reliable clarity judgments using either procedure; however, the method of paired comparisons was more sensitive than category rating in detecting differences between the bandpass-filtered conditions.

?

Eisenberg, Laurie S., Dirks, Donald D., & Gornbein, Jeffrey A. (1997). Subjective judgments of speech clarity measured by paired comparisons and category rating. *Ear and Hearing, 18*, 294-306.

The purpose of this study was to compare listeners' subjective judgments of speech clarity via paired comparisons and category rating using stimulus conditions that varied in the relative spacing between stimulus items, producing either a wide or narrow range of performance. Subjective judgments of speech clarity were measured in 12 normal-hearing (Experiment 1) and 8 hearing-impaired adults (Experiment 2). Sentences processed by six bandpass filters that increased monotonically in Articulation Index (AI) estimates constituted the stimuli to be judged. Using subsets of three filters from the group of six, subjective judgments were additionally obtained for stimulus conditions in which the performance ranges were wide (large differences in AI) and narrow (small differences in AI). Results showed that speech clarity judgments obtained by paired comparisons and category rating were highly related to the AI estimates for both normal-hearing and hearing-impaired subjects. When the performance range was wide, both methods provided similar judgments for the normal-hearing subjects. For the hearing-impaired subjects, paired comparisons were more sensitive than category rating. When the performance range was narrow, paired comparisons were more sensitive than category rating in differentiating between filters for both groups of subjects. This difference was less obvious for the normal-hearing subjects when paired comparison data were converted to a scale comparable to the category ratings. Large between-subject variability was evident for the hearing-impaired subjects on the psychophysical scaling procedures, most notably for category rating. The researchers

concluded that when judging the clarity among stimulus items for which performance varied over a wide range, category rating and paired comparisons provided comparable judgments for normal-hearing listeners. For conditions in which perceptual differences between stimulus items were restricted either by the choice of conditions or by the effects of sensorineural hearing loss, the method of paired comparisons was more sensitive.

5.18.7.1 Readings on Paired Comparisons

Bonebright, T. L. An investigation of data collection methods for auditory stimuli: Paired comparisons versus a computer sorting task. *Behavior Research Methods, Instruments, & Computers*28275–278.(1996).

Bossuyt, P.(1990). A comparison of probabilistic unfolding theories for paired comparisons data.New York: Springer Verlag.

David, H. A.(1963). The method of paired comparisons.Port Jervis, NY: Lubrecht and Cramer.

Eisenberg, L. S.Dirks, D. D. Reliability and sensitivity of paired comparisons and category rating in children. *Journal of Speech and Hearing Research*381157–1167.(1995).

Eisenberg, Laurie S.Dirks, Donald D.Gornbein, Jeffrey A. Subjective judgments of speech clarity measured by paired comparisons and category rating. *Ear and Hearing*18294–306.(1997).

5.18.8 Semantic Differential

Nature: The semantic differential seeks to measure the meaning of an object to an individual. The subject is asked to rate a given concept (e.g., “African American,” “Republican,” “wife,” “me as I would like to be,” “me as I am”) on a series of 7-point, bipolar rating scales. Any concept can be rated, whether a political issue, a person, an institution, or a work of art. The 7-point scales include such bipolar scales as the following: (a) fair-unfair, clean-dirty, good-bad, valuable-worthless; (b) large-small, strong-weak, heavy-light; and (c) active-passive, fast-slow, hot-cold (as shown in Table 5.13). The rating is made according to the respondent's perception of the relatedness or association of the adjective to the word or concept. The three subgroups measure the following three dimensions of attitude: (a) the individual's evaluation of the object or concept being rated, corresponding to the favorable-unfavorable dimension of more traditional attitude scales; (b) the individual's perception of the potency or power of the object or concept; and (c) the individual's perception of the activity of the object or concept.

Utility: A 100-item test can be administered in about 10-15 minutes. A 400-item test takes about an hour. The semantic differential may be adapted to the study of numerous phenomena through choice of concepts and scales. It may be useful in constructing and analyzing sociometric scales.

Table 5.13 EXAMPLE OF A SEMANTIC DIFFERENTIAL SCALE

Fifteen concepts: Love, Child, My Doctor, Me, My Job, Mental Sickness, My Mother, Peace of Mind, Fraud, My Spouse, Self-Control, Hatred, My Father, Confusion, Sex. Each concept was rated on the following 10 scales:

valuable	___:	___:	___:	___:	___:	___:	___:	worthless
clean	___:	___:	___:	___:	___:	___:	___:	dirty
tasty	___:	___:	___:	___:	___:	___:	___:	distasteful
large	___:	___:	___:	___:	___:	___:	___:	small
strong	___:	___:	___:	___:	___:	___:	___:	weak
deep	___:	___:	___:	___:	___:	___:	___:	shallow
fast	___:	___:	___:	___:	___:	___:	___:	slow
active	___:	___:	___:	___:	___:	___:	___:	passive
hot	___:	___:	___:	___:	___:	___:	___:	cold
tense	___:	___:	___:	___:	___:	___:	___:	relaxed

SOURCE: This semantic differential scale was used in a study reported by Charles E. Osgood and Zella Luria (1954), "A Blind Analysis of a Case of Multiple Personality Using the Semantic Differential," *Journal of Abnormal and Social Psychology, 49*, 579-591. For detailed information, see James G. Snider and Charles E. Osgood (Eds.) (1969), *Semantic Differential Technique: A Sourcebook* (Hawthorne, NY: Aldine).

Construction:

The investigator prepares a list of concepts appropriate to the theory guiding the variable to be measured.

Pairs of polar adjectives are selected on a priori grounds.

Selection of adjectives is determined empirically by asking different groups (comparative or experimental-control design) to take prescribed orientations in responding to an adjective-rating task. For example, members of one group of respondents could be asked to rate as they believe a person would rate the concept if he or she held a positive attitude; other respondents could be asked to rate as they believe a person would rate the concept if he or she held a strong negative attitude. Respondents are given the standard instructions for using the semantic differential form. Data are analyzed, and adjective pairs are selected that distinguish clearly between the groups. New groups of respondents are selected who take prescribed orientations in rating the concepts. Data are analyzed.

Examples

?

Bishop, J. Joe. (1999, April). *Locating Czech democracy: A semantic differential analysis ofthe meaning of democracy among students and teachers in three types of secondary schools*. Paper

presented at the Midwest Sociological Society Meeting, Minneapolis, MN.

The study explored the meanings of democracy held by teachers and students in each of the three types of secondary schools in an emerging democracy (the Czech Republic) by locating the meaning in multidimensional semantic space. Data were collected during 2 months of fieldwork conducted in the Czech Republic during the fall of 1997. Students and teachers representative of three different types of schools in one large city, two medium-sized cities, and one small town were asked to think about the type of government the Czech Republic had while they completed a semantic differential scale composed of 57 bipolar adjectives. Factor analysis was used to represent the adjective pairs as a smaller number of variable factors. Results indicated significant age differences on the evaluative factor; sex differences on the potency and stability factors; school-level differences on the evaluative, potency, and stability factors; and a social class/prestige difference on the stability factor. No significant difference was found on the pervasiveness factor.

?

Cogliser, Claudia C., & Schriesheim, Chester A. (1994). Development and application of a new approach to testing the bipolarity of semantic differential items. *Educational and Psychological Measurement, 54*, 594-605.

A method of testing semantic differential scales for bipolarity was developed using a new conception of bipolarity that does not require unidimensionality. Assessment of Fielder's Least Preferred Coworker instrument with 63 college student subjects using multidimensional scaling revealed its significant departures from bipolarity.

?

Hayashi, Naoki, Yamashina, Mitsuru, Ishige, Naoko, Taguchi, Hisako, Igarashi, Yoshito, Hiraga, Masashi, & Inoue, Yukiyo. (2000). Perceptions of schizophrenic patients and their therapists: Application of the semantic differential technique to evaluate the treatment relationship. *Comprehensive Psychiatry, 41*, 197-205.

This study is an attempt to evaluate the treatment relationship with schizophrenic patients by examining the patients' and their therapists' perceptions of themselves and each other, which are hypothesized to reflect features of the relationship. A sample of 158 schizophrenic patients and 11 psychiatrists who each maintained a supportive relationship with the patients as a therapist estimated their perceptions using the semantic differential (SD) technique with 17 adjective pairs. Eight composite scales with sufficient internal consistency were constructed from the estimations. The interrelationship among the perceptual elements, which was represented by correlation analysis of the composite scale scores, seemed consistent with the researchers' clinical experience. A factor-analytic study of the scales yielded three orthogonal factors that could be assumed to characterize the treatment relationship. The patient-therapist cooperation factor indicated the degree of trust between the two participants, supposedly the affective or relational aspect of the therapeutic alliance. The therapist passivity factor reflected the therapist's passive role-taking and the clinical stability of the patient. The patient strength factor was related

to the condition-related and characterological strength of the patient. It was demonstrated that the estimations performed by patients and therapists were valid and useful for evaluation of the treatment relationship in the current status.

5.18.8.1 Readings on Semantic Differential Scales

Bishop, J. Joe.(1999, April). Locating Czech democracy: A semantic differential analysis of the meaning of democracy among students and teachers in three types of secondary schools. Paper presented at the Midwest Sociological Society Meeting, Minneapolis, MN.

Cogliser, Claudia C.Schriesheim, Chester A. Development and application of a new approach to testing the bipolarity of semantic differential items. *Educational and Psychological Measurement*54594–605.(1994).

Hayashi, NaokiYamashina, Mitsurulshige, NaokoTaguchi, Hisakolgarashi, YoshitoHiraga, Masashilnoue, Yukiyo. Perceptions of schizophrenic patients and their therapists: Application of the semantic differential technique to evaluate the treatment relationship. *Comprehensive Psychiatry*41197–205.(2000).

Moss, Claude S.(2001). Dreams, images and fantasy: A semantic differential casebook.Ann Arbor, MI: Books on Demand.

Osgood, Charles E.Luria, Zella. A blind analysis of a case of multiple personality using the semantic differential. *Journal of Abnormal and Social Psychology*49579–591.(1954).

Osgood, Charles E., **Suci, George J.**, & **Tannenbaum, Percy H.(1957).** The measurement of meaning.Urbana: University of Illinois Press.

Schriesheim, Chester A., & others. The equal-interval nature of semantic differential scales: An empirical investigation using Fiedler's Least Preferred Coworker (LPC) scale and magnitude estimation and Case III scaling procedures. *Educational and Psychological Measurement*54253–262.(1994).

Edited by: **Snider, James G.**, & **Osgood, Charles E.** (Eds.). **(1969).** Semantic differential technique: A source-book.Hawthorne, NY: Aldine.

http://dx.doi.org/10.4135/9781412984386.n60