SURVEY DESIGN RECOMMENDATIONS

William P. Fisher, Jr., Ph.D.
Public Health & Preventive Medicine
LSU Health Sciences Center - New Orleans

Item writers and data analysts should follow seventeen basic rules of thumb to create surveys that

1) are likely to provide data of a quality high enough to meet the requirements for measurement specified in a probabilistic conjoint measurement (PCM) model;
2) implement the results of the PCM tests of the quantitative hypothesis in survey and report layouts, making it possible to read interpretable quantities off the instrument at the point of use with no need for further computer analysis; and
3) are joined with other surveys measuring the same variable in a metrology network that ensures continued equating (Masters, 1985) with a single, reference standard metric

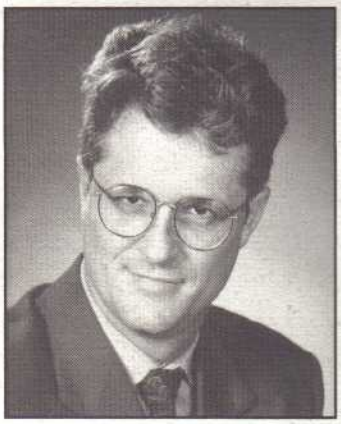
First, make sure all items are expressed in simple, straightforward language.

Second, restrict each item to one idea, meaning avoid conjunctions (and, but, or), synonyms, and dependent clauses. A conjunction indicates the presence of at least two ideas in the item. Having two or more ideas in an item is unacceptable because there is no way to tell from the data which single idea or combination of ideas the respondent was dealing with. If two synonymous words really mean the same thing, only one of them is needed. If the separate ideas are both valuable enough to include, they need to be expressed in separate items. Dependent (if, then) clauses require the respondent to think conditionally or contingently, adding an additional and usually unrecoverable layer of interpretation behind the responses that may muddy the data.

Third, avoid "Not Applicable" or " No Opinion" response categories. It is far better to instruct respondents to skip irrelevant items than it is to offer them the opportunity in every item to seem to provide data, but without having to make a decision.

Fourth, avoid odd numbers of response options. Middle categories tend to attract disproportionate numbers of responses. Again, it allows the respondent to appear to be providing data, but without making a decision concerning preferences. If someone really cannot decide which side of an issue they come down on, let them decide on their own to skip the question.

Fifth, do not assume that respondents will be unable to make more than one or two distinctions in their responses, and do not simply default to the usual four response options (Strongly Agree, Agree, Disagree, Strongly Disagree, or Never, Sometimes, Often, and Always, for instance). The LSU HSI PFS, (Fisher, Marier,



William P. Fisher, Jr.

William P. Fisher, Jr. was formerly Senior Research Scientist for Program Evaluation at Marianjoy Rehabilitation Hospital & Clinics in Wheaton, IL, serving on the Management Team, and on the Clinical Programs and Quality Assessment & Improvement Committees. After completing the University of Chicago's Social Sciences Divisional Master's degree in 1984, William was a Spencer Foundation Dissertation Fellow, earning a Ph.D. in Chicago's Department of Education in 1988, concentrating in Measurement, Evaluation, and Statistical Analysis (MESA). Dr. Fisher is still a MESA Research Associate, is on the Editorial Board of the Journal of Outcome Measurement, and on the Advisory Board of the Institute for Objective Measurement. He is professionally active in diverse organizations. Current tasks include designing and implementing an outcome measurement system for Louisiana's statewide public hospital system; consulting on the Social Security Administration's Disability Process Redesign Project; and drafting scale-free health status measurement standards for the ASTM E31 Committee on Medical Informatics.

Eubanks & Hunter, 1997; Fisher, Eubanks & Marier, 1997) for example, employs a six-point rating scale and is intended for use in the Louisiana statewide public hospital system, which provides most of the indigent care in the state. To date, about 75% of the respondents have less than a high school education and incomes of less than $15,000 per year, but they have shown little or no difficulty in providing consistent responses to the questions posed. Part of the research question raised in any measurement effort concerns determining the number of distinctions that the variable is actually capable of supporting, besides determining the number of distinctions actually required for the needed comparisons. Starting with six (adding in Very Strongly Agree/Disagree categories to the ends of the continuum) or even eight (adding Absolutely Agree/Disagree extremes) response options gives added flexibility in survey design. If one or more categories blends with another and isn't much used, the categories can be combined. Research that starts with fewer categories, though, cannot work the other direction and create new distinctions. More categories have the added benefit of boosting measurement reliability, since, given the same number of items, an increase in the number of functioning (used) categories increases the number of distinctions made among those measured.

Sixth, write questions that will provoke respondents to use all of the available rating options. This will maximize variation, important for obtaining high reliability.

Seventh, write enough questions and have enough response categories to obtain an average error of measurement low enough to provide the needed measurement separation reliability, given sufficient variation. Reliability is a strict mathematical function of error and variation and ought to be more deliberately determined via survey design than it currently is (Linacre, 1993; Woodcock, 1992). For instance, if the survey is to be used to detect a very small treatment effect, measurement error will need to be very low relative to the variation, and discrimination will need to be focused at the point where the group differences are effected, if statistically significant and substantively meaningful results are to be obtained. On the other hand, a reliability of .70 will suffice to simply distinguish high from low measures. Given that there is as much error as variation when reliability is below .70, and it is thus not possible to distinguish two groups of measures in data this unreliable, there would seem to be no need for instruments in that range.

Eighth, before administering the survey, divide the items into three or four groups according to their expected scores. If any one group has significantly fewer items than the others, write more questions for it. If none of the questions are expected to garner very low or very high scores, reconsider the importance of step six above.

Ninth, order the items according to their expected scores and consider what it is about some questions that make them easy (or agreeable or important, etc.), and what it is about other questions that make them difficult (or disagreeable, unimportant, etc.). This exercise in theory development is important because it promotes understanding of the variable. After the first analysis of the data, compare the empirical item order with the theoretical item order. Do the respondents actually order the items in the expected way? If not, why not? If so, are there some individuals or groups who did not? Why?

Tenth, consider the intended population of respondents and speculate on the average score that might be expected from the survey. If the expected average score is near the minimum or the maximum possible, the instrument is off target. Targeting and reliability can be improved by adding items that provoke responses at the unused end of the rating scale. Measurement error is lowest in the middle of the measurement continuum, and increases as measures approach the extremes. Given a particular amount of variation in the measures, more error reduces reliability and less error increases it. Well-targeted instruments enhance measurement efficiency by providing lower error, increased reliability, and more statistically significant distinctions among the measures for the same number of questions asked and rating options offered.

Eleventh, as soon as data from 30-50 respondents are obtained, analyze the data and examine the rating scale structure and the model fit using a partial credit PCM model. Make sure the analysis was done correctly by checking responses in the Guttman scalogram against a couple of respondents' surveys, and by examining the item and person orders for the expected variable. Identify items with poorly populated response options and consider combining categories or changing the category labels. Study the calibration order of the steps and make sure that a higher category always represents more of the variable; consider combining categories or changing the category labels for items with jumbled step structures. Test out recodes in another analysis; check their functioning, and then examine the item order and fit statistics, starting with the fit means and standard deviations in BIGSTEPS Table 3. If some items appear to be addressing a different construct, ask if this separate variable is relevant to the measurement goals. If not, discard or modify the items. If so, use these items as a start at constructing another instrument. When the step structure and model fit are orderly, either continue gathering data on the existing survey and be prepared to make the same edits and changes later with more data, or modify the survey and gather new data in the new format.

Twelfth, when the full calibration sample is obtained, maximize measurement reliability and data consistency. First identify items with poor model fit. If an item is wildly inconsistent, with a mean square fit statistic markedly different from all others, examine the item itself for reasons why its responses should be so variable. Does it perhaps pertain to a different variable? Does the item ask two or more very different questions at once? It may also be relevant to find out which respondents are producing the inconsistencies, as their identities may

suggest reasons for their answers. If the item itself seems to be the source of the problem, it may be set aside for inclusion in another scale, or for revision and later re-incorporation. If the item is functioning in different ways for different groups of respondents, then the data for the two groups ought to be separated into different columns in the analysis, making the single item into two. Finally, if the item is malfunctioning for no apparent reason and for only a very few otherwise credible respondents, it may be necessary to omit only specific, especially inconsistent responses from the calibration. Then, after the highest reliability and maximim data consistency are achieved, another analysis should be done, one in which the inconsistent responses are replaced in the data. The two sets of measures should then be compared in plots to determine how much the inconsistencies actually affect the results.

Thirteenth, the instrument calibration should be compared with calibrations of other similar instruments used to measure other samples from the same population. Do similar items calibrate at similar positions on the measurement continuum? If not, why not? If so, how well do the pseudo-common items correlate and how near the identity line do they fall in a plot? If the rating scale step structures are different, are the step transition calibrations meaningfully spaced relative to each other?

Fourteenth, the calibration results should be fed back onto the instrument itself. When the variable is found to be quantitative and item positions on the metric are stable, that information should be used to reformat the survey into a self-scoring report. This kind of worksheet makes it possible to build the results of the instrument calibration experiment into the way information is organized on a piece of paper, providing quantitative results (measure, error, percentile, qualitative consistency evaluation, interpretive guidelines) at the point of use. No survey should be considered a finished product until this step is taken.

Fifteenth, data should be routinely sampled and recalibrated to check for changes in the respondent population that may be associated with changes in item difficulty.

Sixteenth, for maximum utility, the instrument should be equated with other instruments intended to measure the same variable, creating a reference standard metric.

Seventeenth, everyone interested in measuring the variable should set up a metrology system, a way of maintaining the reference standard metric via comparisons of results across users and brands of instruments. To ensure repeatability, metrology studies typically compare measures made from a single homogeneous sample circulated to all users. Given that this is an unrealistic strategy for most survey research, a workable alternative would be to occasionally employ two or more previously equated instruments in measuring a common sample. Comparisons of these results should help determine whether there are any needs for further user education, instrument modification, or changes to the sampling design.

References

Andrich, D. (1988). Rasch models for measurement. Sage University Paper Series on Quantitative Applications in the Social Sciences, vol. series no. 07-068. Beverly Hills, California: Sage Publications.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. Psychometrika, 42, 631-634.

Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1971). Keymath: Diagnostic Arithmetic Test. Circle Pines, MN: American Guidance Service.

Fisher, W. P., Jr. (1996, October). Rating scale measurement standards relevant to ASTM 1384 on the content and structure of the electronic health record. Unpublished paper. ASTM E31 Committee on the Electronic Health Record, Washington, DC.

Fisher, W. P., Jr. (1997a). Physical disability construct convergence across instruments: Towards a universal metric. Journal of Outcome Measurement, 1(2), 87-113.

Fisher, W. P., Jr. (1997b, June). What scale-free measurement means to health outcomes research. Physical Medicine & Rehabilitation State of the Art Reviews, 11(2), 357-373.

Fisher, W. P., Jr. (1998, May). Objectivity in psychosocial measurement: What, why, how. Second International Outcome Measurement Conference. University of Chicago.

Fisher, W. P., Jr. (1999). Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. Journal of the Louisiana State Medical Society (submitted).

Fisher, W. P., Jr., Eubanks, R. L., & Marier, R. L. (1997). Equating the MOS SF36 and the LSU HSI physical functioning scales. Journal of Outcome Measurement, 1(4), 329-362.

Fisher, W. P., Jr., Harvey, R. F., & Kilgore, K. M. (1995). New developments in functional assessment: Probabilistic models for gold standards. NeuroRehabilitation, 5(1), 3-25.

Fisher, W. P., Jr., Marier, R. L., Eubanks, R., & Hunter, S. M. (1997). The LSU Health Status Instruments (HSI). In J. McGee, N. Goldfield, J. Morton & K. Riley (Eds.), Collecting Information from Patients: A Resource Manual of Tested Questionnaires and Practical Advice (Supplement) (pp. 13:109-13:127). Gaithersburg, Maryland: Aspen Publications, Inc.

Fisher, W. P., Jr., & Wright, B. D. (1994). Introduction to probabilistic conjoint measurement theory and applications. International Journal of Educational Research, 21(6), 559-568.

Linacre, J. M. (1993). Rasch generalizability theory. Rasch Measurement Transactions, 7(1), 283-284.

Linacre, J. M. (1997). Instantaneous measurement and diagnosis. Physical Medicine and Rehabilitation State of the Art Reviews, 11(2), 315-324.

Mandel, J. (1977, March). The analysis of interlaboratory test data. ASTM Standardization News, 5, 17-20, 56.

Mandel, J. (1978, December). Interlaboratory testing. ASTM Standardization News, 6, 11-12.

Masters, G. N. (1985, March). Common-person equating with the Rasch model. Applied Psychological Measurement, 9(1), 73-82.

Masters, G. N., Adams, R. J., & Lokan, J. (1994). Mapping student achievement. International Journal of Educational Research, 21(6), 595-610.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. Social Studies of Science, 23, 129-173.

Pennella, C. R. (1997). Managing the metrology system. Milwaukee, WI: ASQ Quality Press.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. Applied Psychological Measurement, 3(2), 237-255.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Sparrow, S., Balla, D., & Cicchetti, D. (1984). Interview edition, survey form manual, Vineland Behavior Scales. Circle Pines, MN: American Guidance Services, Inc.

Suppes, P., Krantz, D., Luce, R., & Tversky, A. (1989). Foundations of measurement, Volume II: Geometric and probabilistic representations. New York: Academic Press.

Wise, M. N. (Ed.). (1995). The values of precision. Princeton, NJ: Princeton University Press.

Woodcock, R. W. (1973). Woodcock Reading Mastery Tests. Circle Pines, MN: American Guidance Service, Inc.

Woodcock, R. W. (1992). Woodcock test design nomograph. Rasch Measurement Transactions, 6(3), 243-244.

Woodcock, R. W. (1999). What can Rasch-based scores convey about a person's test performance? In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement: What every psychologist and educator should know. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), Measurement and personality assessment. North Holland: Elsevier Science Ltd.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement: What every educator and psychologist should know. Hillsdale, NJ: Lawrence Erlbaum Associates.