**Inter-Rater Reliability: Definitions, Obstacles and Remedies**

When utilizing an instrument, e.g., the Certificate of Eligibility, to determine qualification for services, it would be imperative to have that tool accurately reflect what it purports to measure. Researchers in the social sciences describing a parameter of this goal, put it this way: "Reliability refers to the ability of a measure to yield consistent results each time it is used" (Monette, 2002, 212); or ". . . a highly reliable measure is one that gives the user consistent results over time, places, and occasions" (Adams and Schvaneveldt, 1991, 86-87). Inherent in the concept of reliability is the fact that regardless of who gathers the data, the information will be identical. For instance, if a recruiter secures information from a family on May 1, 2007, and returns in a month to verify those facts, the eligibility data (e.g., qualifying arrival date, qualifying activity, etc.) will be similar. Or if a supervisor conducts a monitoring visit with the family, she/he will receive the same basic information garnered by the initial recruiter.

There are three types of reliability: stability reliability, representative reliability, and equivalence reliability (Neuman, 1997, 138-139). With stability reliability, the researcher wants to know if the measure is stable over time. Determination of this feature occurs via the test-retest method in which the same population takes the same measure at different time periods, hopefully yielding the same results. If the instrument provides similar answers across varying populations (e.g., different genders or ethnicities), then representative reliability exists. Finally, the use of multiple indicators offering similar results is equivalence reliability, with the indicators in this case being different recruiters employing the same measure.

A form of equivalence reliability is that of interrater or intercoder reliability which occurs when there is agreement among the coders or raters on the measure being employed (Neuman, 1997, 138-139). One way of testing for this would be to have a number of coders use the same measure, followed by a comparison of results. Measurement of interrater reliability takes the form of a reliability coefficient arrived at by a set of statistical techniques. The closer the value to +1.00 the greater the agreement among coders which is optimal (Adams and Schvaneveldt, 1991, 87). There is some variation in what researchers consider acceptable levels of reliability, ranging from 75% or more (Bailey 1987), with more seasoned raters in the 85% and above range, while others opt for the ideal of at least 85-90% (Monette, 2002, 213, 448).

Numerous variables affect reliability as Monette indicates, citing the work of several researchers (Holsti, 1969; Weber, 1990; Scott, 1990): "Reliability in content analysis depends on many factors, including the skills of the coders, the nature of the categories, and the degree of clarity or ambiguity in the documents" (Monette, 2002, 212). To that end, it is relevant to briefly explore some of the obstacles to reliability, followed by ways to address these challenges.

Problems in accurate data collection in the interactional process between interviewer and respondent can reside in either party. This does not suggest, however, that the dilemma is simply a matter of malevolence or intentionality. For instance, the mistakes given by a respondent may relate to forgetfulness, embarrassment, the influence of the presence of others, an uncomfortable setting, or some misunderstanding of their role (Neuman, 1997, 254-259). Citing Turner and Martin (1984), Neuman explained that not only were interviewees often unclear of the expectations for them but they also responded

to specific questions by applying them to their own situations, making them easier to understand without necessarily being accurate, at least from the viewpoint of the researcher (Neuman, 1997, 254). At the same time, any ambiguity and lack of clarity in the questionnaire itself can lead to incorrect information (Isaac and Michael, 1995, 146). Assuming that a questionnaire has clarity and that respondents will grasp the full meaning of these inquiries is not always correct:

> "Inaccurate reporting is not a response tendency or a predisposition to be untruthful. Individuals who are truthful on one occasion or in response to particular questions may not be truthful at other times or to other questions" (Wentworth, 1993, 130).

Gender, ethnicity, socioeconomic class and the social setting may all contribute to differing responses on differing occasions (Neuman, 1997, 25256-262).

Evasions and outright deception may also occur with culpability being in either the respondent or the interviewer (Neuman, 1997, 259). One researcher, describing the attempts by members of various groups to either inflate their status or avoid sensitive areas, claimed: "In all other research settings I've known about in any detail, lying was common, both among members and to researchers, especially about the things that were really important to the members" (Douglas, 1976, 73). As mentioned earlier, interviewer bias and intentional subversion can be quite problematic. In one case, a researcher described the process of "curbing" whereby interviewers simply completed responses without actually contacting anyone (Frey, 1989). The inept, unskilled or unethical interviewer can sabotage even the most well-defined instrument.

Strategies for addressing these issues focus upon the development and refinement of the instrument for research coupled with the training of those designated to implement the research. In the case in question, then, efforts would target the Certificate of

Eligibility and its various components along with the careful selection, training, and monitoring of those applying the questionnaire in the field.

Relative to the instrument determining eligibility, researchers maintain that the questions be constructed in language that will provide for unambiguous clarity between researcher and respondent so that there is little chance of misinterpretation (Isaac and Michael, 1995, 146). Clearly conceptualizing all the constructs to be measured and revising them when need be is critical (Neuman,1997, 140; Monette et al., 2002, 124). Moving to a higher level of measurement is another suggestion that will enhance reliability (Monette et al., 2002, 125; Neuman, 1997, 140). Collecting information that transcends a nominal measure (race, gender, religion) and moving to information that is an ordinal measure (socioeconomic class which can be categorized and ranked in a number of brackets) will contribute to greater reliability.

Another recommendation is the incorporation of multiple indicators of a particular variable. For example, rather than having a box that says "agricultural related employment," an inclusion of numerous examples in that category would aid in the reliability process (Neuman, 1997, 140-141; Monette et al., 2002, 125). Doing a thorough evaluation of each item on multiple-item measures would also be useful (Monette et al., 2002 125). In re-evaluations of some Certificates of Eligibility, error rates were purported to be high in some geographical areas. It would be particularly important, however, to identify WHICH of the measures demonstrated higher error rates. In that way, adjustments to specific questions in the instrument could be made, contributing again to increased reliability. Finally, the employment of pretests, pilot studies and replication will aid in improving reliability. The development of drafts,

tested on a variety of individuals, followed by revisions of measures that are unclear contributes to the retrieval of quality data (Neuman, 1997, 141).

There are important considerations when utilizing an instrument with populations that are culturally diverse. Depending upon the population under study, language differences in addition to other cultural variants may play a significant part in the comprehension of questions and the collection of accurate responses. Those versed in studies of diverse groups offer practical suggestions: involve members of the study population in both the development of the instrument and in any pilot testing, seeking their feedback and critique on areas of confusion or obfuscation; in the event that the instrument is translated into another language, solicit population study members using the 'double translation' method (English into the target language with translation by another individual back into English, checking for mistakes); incorporate population members in any kind of pilot testing program, asking for input on why certain measures may not be resulting in accurate reporting or why the questions are being understood differently by the target population (Marin and Marin 1991; Tran and Williams, 1994; Monette et al., 2002, 122, 125).

Having developed, pilot tested, and revised the instrument encompassing the earlier recommendations, it is readily apparent that the training of those involved in the actual administration of the instrument and procurement of the data is essential. When numerous coders/raters/recruiters go into the field to interview families in the hope of getting accurate and factual information on their life circumstances with the goal of ascertaining appropriateness of migrant educational services for their children, effective training of these gatekeepers is paramount. When multiple persons employ the same

instrument, the potential for problems exists: "Survey researchers have known for a long time that different interviewers get different answers from respondents as a result of their own attitudes and demeanors" (Rubin and Babbie, 1997, 175). Speaking to the importance of the preparation for interviewers in a research endeavor, Wolfer claims:

> "Interviewer training entails familiarizing the interviewer with the details of the study, the organization of the survey, illustrations about how to effectively reinforce that the interview will be confidential, and how to probe should the need arise. Hence, interviewers need to be aware of the research topic and goals so that they are equipped to judge when respondents have given the level of information necessary for adequate analysis. This does not mean that interviewer is looking f or a specific response. . . –it just means that the interviewer is looking for a specific level of information. This will allow the interviewer to determine whether a response needs clarification for the researcher's purpose" (Wolfer, 2007, 309).

Also addressing the training issue, Neuman underscores the importance of the selection process and adequate compensation: "A professional-quality interview requires the careful selection of interviewers and good training. As with any employment situation, adequate pay and good supervision are important for consistent high-quality performance" (Neuman, 1997, 257). However, he then bemoans the fact that, unfortunately, this is often not the case. While the regimen for preparing interviewers will differ, there are some generic suggestions:

> "Researchers train professional interviewers in a one-to two-week training course, which usually includes lectures and reading, observation of expert interviewers, mock interviews in the office and in the field that are recorded and critiqued, many practice interviews, and role playing" (Neuman, 1997, 258).

Subsequent to the adequate training of interviewers and revision of their instruments, the process of ongoing evaluation continues:

> In face-to-face interviews, supervisors check to find out whether the interview actually took place. This means calling back or sending a confirmation postcard to a sample of respondents . . . and they may reinterview a small subsample,

> analyze answers, or observe interviews to see whether interviewers are accurately asking questions and recording answers" (Neuman, 1997, 257).

Experience demonstrates that the process of training and ongoing supervision in this manner can insure accurate outcomes in the eligibility determination process. In one instance, a training program in the Georgia AFDC and Food Stamp Program subsequent to the discovery of high error rates in eligibility determination had a positive outcome in worker job performance (Lindsey et al., 1995).

Similarly, others echo the importance of the ongoing process of supervision and monitoring process to ensure the quality and integrity of the process:

> "Proper supervision begins during interviewer training. The importance of contacting the right respondents and meticulously following established procedures should be stressed. Interviewers should be informed that their work will be carefully checked and that failure to follow procedures will not be tolerated. . . Completed interviews should be scrutinized for any evidence of falsification. Spot checks can be made to see if interviewers are where they are supposed to be at any given time, and respondents can be telephoned to see if they have, in fact, been interviewed" (Monette et al., 2002, 183-184).

Antidotes to some of the problems mentioned, including interviewer bias and unreliability, then, include the selection of quality personnel, rigorous training, and continual monitoring and evaluation of their work (Neuman, 1997, 257).

It is not the purpose of this review to plumb the depths of the selection, training, and evaluation process of interviewers. However, research provides a few directions for refining the interaction between the interviewer and respondent in the administration of a questionnaire format. "Double-coding," the procedure of having two individuals code the same material followed by an evaluation of the areas of agreement and disagreement can yield increased accuracy (Monette et al., 2002, 447). One can have even more individuals involved in this process, arriving at the reliability coefficient. If there is

significant discrepancy, then it is clear that problems may lie in the instrument itself, the training of the coders or possibly the simple ineptness of some coders. If some persons continually differ from the norm, they will need to be replaced (Monette et al., 2002, 212).

It would be remiss to omit the importance of computer-based technology possibilities relative to the efficient and accurate collection of data in the 21st century. As the tablet project of the Migrant Education Project indicates, there are increasingly available technologies that will allow for the accomplishment of goals such as reliable information from various respondents over wide populations. Training, evaluation, and monitoring can conceivably be dramatically enhanced by moving in this direction. Social scientists and researchers are leading the charge: "There are important linkages between social research and human service practice, and the imaginative application of computer technology is yet another way of forging this linkage" (Monette et al., 2002, 13). The same author points out that the computer can be utilized not only to store data but also to observe and study a variety of social phenomena (45).

In the interview process, portable computers are useful in many ways. For instance, drop-down menus on various measures seem particularly conducive to compiling more accurate data in the eligibility determination process: "Computers can also handle complex question branching, in which each question asked is dependent on earlier answers given by respondents" (126). In another situation, researchers employed a self-administered survey to adolescents, seeking sensitive information. They found that individuals preferred this modality to that of the paper and pencil model (Supple et al. 1999). Evidently, these tools can assist interviewers, supervisors, and all program

participants in more efficiently and accurately doing what they are commissioned to do, the comprehensive and factual determination of eligibility of families for migrant educational services.

## Reference


Adams, G.R. and J.D. Schvaneveldt. 1991. **Understanding Research Methods.** 2nd ed. New York: Longman.

Bailey, K. 1987. **Methods of Social Research.** 3rd. ed. New York: Free Press.

Douglas, J.D. 1976. **Investigative Social Research.** Beverly Hills, CA: Sage.

Frey, J.H. 1986. *An Experiment with a Confidentiality Reminder in a Telephone Survey.* **Public Opinion Quarterly.** Vol. 50 (Summer): 267-269.

Holsti, O.R. 1969. **Content Analysis for the Social Sciences and Humanities.** Reading, MA: Addison-Wesley.

Isaac, S. and W.B. Michael. 1995. **Handbook in Research and Evaluation.** 3rd ed. San Diego, CA: EdITS.

Lindsey, E.W., N.P. Kropf and S. Carse-McLocklin. 1995. *Training Public Assistance Workers in Policy and Interpersonal Helping Skills.* **Research on Social Work Practice.** Vol. 5, No. 1 (January), 20-35.

Marin, G. and B. VanOss Marin. 1991. **Research with Hispanic Populations.** Newbury Park, CA: Sage.

Monette, D.R., T.J. Sullivan and C.R. DeJong. 2002. **Applied Social Research.** Orlando, FLA: Harcourt Press.

Neuman, W.L. 1997. **Social Research Methods: Qualitative and Quantitative Approaches.** 3rd ed. Boston: Allyn and Bacon.

Rubin, A. and E. Babbie. 1997. **Research Methods for Social Work.** 3rd ed. Pacific Grove, CA: Brooks/Cole Publishing Co.

Scott, J.A. 1990. **A Matter of Record: Documentary Sources in Social Research.** Oxford, England: Polity Press.

Supple, A.J., W.S. Aquilino and D.L. Wright. 1999. *Collecting Sensitive Self-Report Data With Laptop Computers: Impact on the Response Tendencies of Adolescents in a Home Interview.* **Journal of Research on Adolescence.** Vol. 9, No. 4. 467-488.

Tran, T.V. and L.F. Williams. 1994. *Effect of Language of Interview on the Validity and Reliability of Psychological Well-Being Scales.* **Social Work Research.** Vol. 18 (March): 17-25.

Turner, C. and E. Martin, eds. 1984. **Surveying Subjective Phenomena.** Vol. 1. New York: Russell Sage Foundation.

Weber, R.P. 1990. **Basic Content Analysis.** 2nd ed. Beverly Hills, CA: Sage.

Wentworth, E.J. 1993. **Survey Responses: An Evaluation of Their Validity.** New York: Academic Press.

Wolfer, L. 2007. **Real Research: Conducting and Evaluating Research in the Social Sciences.** Boston: Pearson/Allyn and Bacon.