of data records, each of which has a large number of attributes, including everything from the question text and skip pattern to documentation notes and instructions to the interviewer. We will refer to this as the *database approach.* (The reader can profitably read the entry on DATA MANAGEMENT, which explains relational databases, as it is germane to this section as well.) The thesis here is that the database approach provides an overarching conceptual approach to integrating all the steps in conducting a CAPI interview and allows for significant operational economies over the alternative programming approach.

In the design stage, survey specialists enter data into the various fields of a screen that populates the rows of relational database tables representing the questionnaire. The data entered include question text, edit restrictions on allowable responses, the list of acceptable answers, and skip instructions. This approach to "authoring" the CAPI questionnaire requires minimal programming skill, and junior survey specialists can enter all but the most complex passages. A set of predesigned "queries" to the database generates diagnostic checks as the data are entered, with more sophisticated checks and diagnostics run in batch mode.

Each question in the instrument is represented by joining rows from the database tables. The interviewing software performs these joins, displays the question, stores the answer, and branches as per the branching instructions. The program that reads and executes the questionnaire data is not rewritten for each survey but remains stable and is reused for different surveys. (The same records can also be used on a PDA, Web server, or local server to enable a variety of platforms for data collection.) This simplifies training for both questionnaire authors and interviewers because all interactions with the system reuse a few standard screen displays.

The survey data are transmitted to the central office and loaded into a master relational database already containing the tables that define the questionnaire. At this point, the questionnaire data become the "metadata," or data that describe the survey data, greatly facilitating the generation of documentation reports that are run as queries to the relational database. Predesigned queries generate either a printable or HTML questionnaire or a codebook. Additional information relevant to each question can be stored in the database and used to augment survey documentation.

We can extract input data files for the next round of a longitudinal survey from the relational database. One can distribute the public use data over the Web to users by employing the relational database techniques used for commercial transactions; instead of putting books in their shopping cart, researchers can search the database for the variables they need, mark them for extraction, and have a remote server e-mail them their data file.

This approach is holistic and uses commercial database software to integrate all phases of a CAPI survey. A well-designed system will handle a broad range of surveys, allowing both central office and interviewing staff to reuse standard tools. This approach offers significant opportunities to control costs and keep operations coordinated and efficient (see DATA MANAGEMENT).

—Randall J. Olsen

REFERENCES

Costigan, P., & Thomson, K. (1992). Issues in the design of CAPI questionnaires for complex surveys. In A. Westlake, R. Banks, C. Payne, & T. Orchard (Eds.), *Survey and statistical computing* (pp. 47–156). London: North Holland.

Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. F., Martin, J., Nichols, W. L., et al. (Eds.). (1998). *Computer assisted survey information collection.* New York: John Wiley.

Forster, E., & McCleery, A. (1999). Computer assisted personal interviewing: A method of capturing sensitive information. *IASSIST Quarterly, 23*(2), 26–38.

Saris, W. E. (1991). *Computer-assisted interviewing.* Newbury Park, CA: Sage.

CONCEPT. *See* CONCEPTUALIZATION, OPERATIONALIZATION, AND MEASUREMENT

CONCEPTUALIZATION, OPERATIONALIZATION, AND MEASUREMENT

Research begins with a "problem" or topic. Thinking about the problem results in identifying concepts that capture the phenomenon being studied. Concepts, or CONSTRUCTS, are ideas that represent the phenomenon. Conceptualization is the process whereby these concepts are given theoretical meaning. The process typically involves defining the concepts abstractly in theoretical terms.

Describing social phenomena and testing hypotheses require that concept(s) be operationalized. Operationalization moves the researcher from the abstract level to the empirical level, where variables rather than concepts are the focus. It refers to the operations or procedures needed to measure the concept(s). Measurement is the process by which numerals (or some other labels) are attached to levels or characteristics of the variables. The actual research then involves empirically studying the variables to make statements (descriptive, relational, or causal) about the concepts.

CONCEPTUALIZATION

Although research may begin with only a few and sometimes only loosely defined concepts, the early stages usually involve defining concepts and specifying how these concepts are related. In exploratory research, a goal is often to better define a concept or identify additional important concepts and possible relationships.

Because many concepts in social science are represented by words used in everyday conversation, it is essential that the concepts be defined. For example, the concepts *norms, inequality, poverty, justice,* and *legitimacy* are a part of our everyday experiences and thus carry various meanings for different people. The definitions provided by the researchers are usually referred to as nominal definitions (i.e., concepts defined using other words). No claim is made that these definitions represent what these concepts "really" are. They are definitions whose purpose it is to communicate to others what the concept means when the word is used. A goal in social science is to standardize definitions of the key concepts. Arguably, the most fruitful research programs in social science—those that produce the most knowledge—are those in which the key concepts are agreed on and defined the same way by all.

Concepts vary in their degree of abstractness. As an example, human capital is a very abstract concept, whereas education (often used as an indicator of human capital) is less abstract. Education can vary in quality and quantity, however, and thus is more abstract than years of formal schooling. Social science theories that are more abstract are usually viewed as being the most useful for advancing knowledge. However, as concepts become more abstract, reaching agreement on appropriate measurement strategies becomes more difficult.

OPERATIONALIZATION

Operationalization of concepts involves moving from the abstract to the empirical level. Social science researchers do not use this term as much as in the past, primarily because of the negative connotation associated with its use in certain contexts. One such use has been in the study of human intelligence. Because consensus on the meaning of this concept has been hard to come by, some researchers simply argued that intelligence is what intelligence tests measure. Thus, the concept is defined by the operations used to measure it. This so-called "raw empiricism" has drawn considerable criticism, and as a consequence, few researchers define their concepts by how they are operationalized. Instead, nominal definitions are used as described above, and measurement of the concepts is viewed as a distinct and different activity. Researchers realize that measures do not perfectly capture concepts, although, as described below, the goal is to obtain measures that validly and reliably capture the concepts.

MEASUREMENT

Measurement refers to the process of assigning numerals (or other labels) to the levels or characteristics of variables. In moving from the abstract level of concepts to the empirical level of variables, the researcher must think of the concept as having characteristics that can be empirically observed or assessed. A number of concerns must be addressed. The first of these is LEVEL OF MEASUREMENT.

At the most basic level, the researcher must decide whether the underlying features of the concept allow for ordering cases (ordinal level) or allow only for categorizing cases (nominal level). Another distinction concerns whether the features of the concept are discrete or continuous, with fine gradations. Relying on these distinctions, most researchers are faced with variables that are nonordered and discrete (e.g., marital status and religious preference), ordered and discrete (e.g., social class), or ordered and continuous (e.g., income). Developments in statistics over the past several decades have dramatically increased the analysis tools associated with ordered discrete variables, but the majority of the research in social science still adheres to the dichotomy of categorical versus continuous variables, with the more "liberal" researchers assuming that ordinal-level data can be analyzed as if they were continuous. With the advances in statistical techniques, however, these decisions no longer need to be made; ordinal variables can be treated as ordinal variables in sophisticated multivariate research (see Long, 1997).

The first step in measurement, then, is to determine the level of measurement that is inherent to the concept. As a general rule, measurement should always represent the highest level of measurement possible for a concept. The reason for this is simple. A continuous variable that is measured like a category cannot be easily, if ever, converted to a continuous scale. A continuous scale, however, can always be transformed into an ordinal- or nominal-level measure. In addition, ordinal and continuous scales allow for the use of more powerful statistical tests, that is, tests that have a higher likelihood of rejecting the null hypothesis when it should be rejected.

Any measurement usually involves some type of measurement ERROR. One type of error, *measurement invalidity,* refers to the degree to which the measure incorrectly captures the concept (DeVellis, 1991). Invalidity is referred to as systematic error or bias. VALIDITY is usually thought of in terms of the concept being a target, with the measure being the arrow used to hit the target. The degree to which the measure hits the center of the target (the concept) is the measure's validity. Researchers want measures with high validity. The reason for this is simple, as shown by an example from a study of social mobility. If the researcher's goal is to determine the degree of social mobility in a society, and the measure consistently underestimates mobility, then the results with this measure are biased and misleading (i.e., they are invalid).

The other major type of measurement error is *unreliability.* Whereas invalidity refers to accuracy of the measure, unreliability refers to inconsistency in what the measure produces under repeated uses. Measures may, on average, give the correct score for a case on a variable, but if different scores are produced for that case when the measure is used again, then the measure is said to be unreliable. Fortunately, unreliability can be corrected for statistically; invalidity is not as readily corrected, however, and it is not as easily assessed as is RELIABILITY.

Validity is arguably the more important characteristic of a measure. A reliable measure that does not capture the concept is of little value, and results based on its use would be misleading. Careful conceptualization is critical in increasing measurement validity. If the concept is fuzzy and poorly defined, measurement validity is likely to suffer because the researcher is uncertain as to what should be measured. There are several ways of assessing measurement validity.

FACE VALIDITY, whether the measure "on the surface" captures the concept, is not standardized or quantifiable but is, nevertheless, a valuable first assessment that should always be made in research. For example, measuring a family's household income with the income of only the male head of household may give consistent results (be reliable), but its validity must be questioned, especially in a society with so many dual-earner households.

Content validity is similar to face validity but uses stricter standards. For a measure to have content validity, it must capture all dimensions or features of the concept as it is defined. For example, a general job satisfaction measure should include pay satisfaction, job security satisfaction, satisfaction with promotion opportunities, and so on. As with face validity, however, content validity is seldom quantified.

CRITERION-RELATED VALIDITY is the degree to which a measure correlates with some other measure accepted as an accurate indicator of the concept. This can take two forms. *Predictive validity* uses a future criterion. For example, voting preference (measured prior to the election) is correlated with actual voting behavior. A high correlation indicates that voting preference is a valid measure of voting behavior. *Concurrent validity* is assessed by obtaining a correlation between the measure and another measure that has been shown to be a valid indicator of the concept. For example, a researcher in the sociology of work area wishes to measure employee participation by asking employees how much they participated in decision making. A more time-consuming strategy—videotaping work group meetings and recording contributions of each employee—would likely be viewed as a valid measure of participation. The correlation between the two would indicate the concurrent validity of the perceptual measure.

CONSTRUCT VALIDITY of a measure refers to one of two validity assessment strategies. First, it can refer to whether the variable, when assessed with this measure, behaves as it should. For example, if the theory (and/or past research) says it should be related positively to another variable Y, then that relationship should be found when the measure is used. The second use of construct validity refers to the degree to which multiple indicators of the concept are related to the underlying construct and not to some other construct. Often, this distinction is referred to with the terms *convergent validity* and DISCRIMINANT VALIDITY. FACTOR ANALYSIS can be used to assess this. For example, if a researcher has five indicators of cultural capital and four indicators of social capital, a factor analysis should produce two lowly correlated factors, one for each set of indicators.

It is important to stress that this second strategy uses factor analysis for confirmatory purposes, not for data dredging or concept hunting, as it is used sometimes in exploratory research. If the researchers have been careful in conceptualization, then the factor analysis serves to validate their measures. Factor analysis in this instance is not used to "discover" unanticipated concepts.

As stated above, reliability refers to the consistency of the results when the measure is used repeatedly. It is viewed as random measurement error, whereas validity is thought of as measurement bias. There are clear statistical consequences of unreliable measures, with larger variances and attenuated correlations at the top of the list. Reliability is usually assessed in one of two ways. TEST-RETEST RELIABILITY is assessed with the correlation of the measure with itself at two points in time. The difficulty associated with this type of assessment is that unreliability of the measure is confounded with actual change in the variable being measured. For this reason, a second and more often used strategy is to obtain multiple indicators of the concept. Here, internal consistency measures such as the CRONBACH'S ALPHA can be used. In both instances, a normed measure of association (correlation) allows for application of certain rules of thumb. For example, any measures with reliabilities under .60 should be regarded with considerable suspicion.

ADDITIONAL ISSUES

Measurement validity should not be confused with INTERNAL VALIDITY and EXTERNAL VALIDITY, which are research design issues. If researchers claim the results support their hypothesis that X causes Y, this claim is said to have internal validity if alternative explanations for the results have been ruled out. External validity refers to whether the study results can be generalized beyond the setting, sample, or time frame for the study. Both are very important concerns in building social science knowledge, but they are not the same as measurement validity.

Another measurement issue concerns whether single or multiple indicators should be used as measures. Some concepts, almost by definition, can be measured with single indicators. Some examples are income and education. Other more abstract concepts (e.g., political conservatism and marital satisfaction) are often measured with more than one indicator. The reasons for wanting multiple indicators are fairly straightforward. First, when a concept is more abstract, finding one measure that captures it is more difficult. Second, multiple-indicator measures are usually more reliable than single-indicator measures. Third, multiple-indicator measures, when used in STRUCTURAL EQUATION MODELING (SEM) (Bollen, 1989), allow for correcting for unreliability of the measures, produce additional information that allows for making useful model tests, and allow for estimating more complicated models such as reciprocal effects models. Although SEM achieves its strengths by treating multiple indicators separately, measures based on multiple indicators are often transformed into scales or indexes, which are the simple sums or weighted sums of the indicators.

A third measurement issue concerns the source of valid and reliable measures. As has been described, there exist a number of strategies for assessing how good a measure is (i.e., how valid and reliable it is). This assessment happens after the measurement has occurred, however. If the measure is not valid or reliable and other better measures cannot be easily obtained, discontinuation of the research must be seriously considered. For this reason, considerable attention must be given to identifying valid and reliable measures at the onset of the study. If the theory in a specialty area is well established and there already exists a strong research tradition, then valid and reliable measures likely already exist. For example, in the study of organizations, key concepts such as formalization and administrative intensity are well established, as are their measures. In such instances, a careful reading of this literature identifies the measures that should be used. It is always tempting for those new to a specialty area to want to "find a better measure," but unless previously used measures have low validity and reliability, this temptation should be avoided. Accumulation of knowledge is much more likely when measures of the key concepts are standardized across studies. Exploratory research, of course, is much less likely to use standardized measures. In such instances, concepts and measures are developed "along the way." Such research, however, should not be viewed as entirely unstructured and lacking standards of evaluation and interpretation. Maxwell (1996) convincingly shows that exploratory research can still be conducted within quasi-scientific guidelines.

A final measurement issue concerns the use of available data, which now is the major source of data for social science research. Often, these data were collected for some purpose other than what the researcher has in mind. In such instances, PROXY VARIABLES are often relied on, and validity issues usually must be confronted. Reliability is usually less of a problem because proxy measures often seem to be sociodemographic variables. Fortunately, large-scale data collection organizations are now relying more on input from social scientists for measures when they collect their data. Having to locate and use proxy measures of key concepts, it is hoped, will become a thing of the past.

—Charles W. Mueller

REFERENCES

Bollen, K. (1989). *Structural equations with latent variables.* New York: John Wiley.

DeVellis, R. F. (1991). *Scale development: Theory and applications.* Newbury Park, CA: Sage.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

Maxwell, J. A. (1996). *Qualitative research design.* Thousand Oaks, CA: Sage.

CONDITIONAL LIKELIHOOD RATIO TEST

One procedure for testing a restriction on parameters estimated with CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION is the conditional likelihood ratio test. The intuition behind the conditional likelihood ratio statistic is analogous to that of the LIKELIHOOD RATIO STATISTIC. The conditional likelihood ratio statistic measures the reduction in the conditional log-likelihood function from imposing the restriction. If the restriction is valid, the reduction in the conditional log-likelihood function should be insignificant. The large sample distribution of the conditional likelihood ratio statistic is chi-squared with DEGREES OF FREEDOM equal to the number of restrictions imposed. As with the likelihood ratio statistic, the conditional likelihood ratio test has counterparts in conditional versions of the Wald and score (Lagrange multiplier) tests.

More formally, suppose that a parameter vector, $\boldsymbol{\beta}$, is estimated by maximizing the following conditional log-likelihood function, $\ln L = \sum f(y_2|\mathbf{X}, \boldsymbol{\beta}, \mathbf{Z}, \gamma)$, where γ is a parameter vector whose values are set by theoretical assumption or, more commonly, replaced with the estimate, $\hat{\gamma}$, which is obtained by maximizing the marginal log-likelihood function, $\ln L = \sum g(y_1|\mathbf{Z}, \gamma)$. Now consider a $(r \times 1)$ vector of restrictions on $\boldsymbol{\beta}$, defined as $\mathbf{c}(\boldsymbol{\beta}) - \mathbf{q} = \mathbf{0}$. The conditional likelihood ratio test is based on the difference between $\ln\ \hat{L}$, the conditional log-likelihood function at the unrestricted estimate of $\boldsymbol{\beta}$, and $\ln \hat{L}_R$, the conditional log-likelihood function at the restricted estimate of $\boldsymbol{\beta}$. Note that both are conditioned on γ, which is treated as known. Then, the conditional likelihood ratio statistic is $-2(\ln \hat{L}_R - \ln \hat{L}) \sim \chi^2_r$, which tests $\mathbf{c}(\boldsymbol{\beta}) - \mathbf{q} = \mathbf{0}$ as the null hypothesis.

One application of the conditional likelihood ratio test is the test of exogeneity in the context of simultaneous PROBIT models. Consider the specific example of a probit model in which one of the regressors, $y_1 = \mathbf{Z}\gamma + v$, is a continuous ENDOGENOUS variable with a NORMAL DISTRIBUTION. Conditional maximum likelihood estimation of this system of equations would involve obtaining a maximum likelihood estimate of γ by ORDINARY LEAST SQUARES and then using this estimate to condition the maximum likelihood estimation of the probit equation. The conditioning in this case is using $\hat{\gamma}$ to derive reduced-form residuals, $\hat{v}$, that would be included on the right-hand side of the probit equation to control for simultaneity bias. The conditional likelihood ratio test would compare the log-likelihood for this unrestricted model to the log-likelihood for the restricted model that imposes the restriction of exogeneity by excluding the reduced-form residuals. In this specific example, the conditional likelihood ratio statistic would have a chi-squared distribution with one degree of freedom. Note that under the null hypothesis of exogeneity, this conditional likelihood ratio test is asymptotically equivalent to the likelihood ratio test.

–Harvey D. Palmer

REFERENCES

Alvarez, R. M., & Glasgow, G. (2000). Two-stage estimation of nonrecursive choice models. *Political Analysis, 8,* 147–165.

Rivers, D., & Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics, 39,* 347–366.