2
Methodological Considerations

The selection of instruments for use in a particular study or the construction of new instruments to assess ethnicity, race, gender, sex, and related constructs requires a consideration of various methodological issues, in addition to the context in which these tools are to be developed and the ethical implications of the categories once they have been developed. Issues to be considered prior to deciding which of existing instruments to use or whether and how to construct a new instrument include the focus of the research question, the format to be used to collect the data, and how the population of interest is to be sampled. The selection of an instrument for use, or the development of a new instrument, also requires attention to the instrument's validity, reliability, and the possibility of misclassification associated with its use (McDowell and Newell, 1996). A basic understanding of these issues is important in order to better evaluate the literature that exists with regard to the constructs that are the focus of this text. This is not, however, a comprehensive discussion of these issues, which can be the focus of entire books themselves, and the reader is urged to consult the sources listed at the end of this chapter for additional guidance.

Framing the Research Question and the Research

How a research question is framed and the design of the study that will be undertaken for its investigation are critical issues to be resolved prior to identifying the instruments to be used or whether and how to develop a new instrument for the assessment of any of the constructs discussed in this text. Issues requiring consideration include the following.

(1) The time period of interest. Because one's self-identity may change over time with respect to ethnicity, race, sexual orientation, and related constructs, it is important to determine at what point in time these are to be assessed. For instance, does the research question demand an understanding of how an individual currently self-identifies? This might be relevant, for instance, in studies assessing current patient satisfaction with health care. Or, does the study focus on the impact of stigmatization on one's health status over time? In this case, it may be important

to assess the individual's identity over time and/or how an individual is perceived by others in terms of his or her race, ethnicity, sexual orientation, etc.

This is an important consideration even in instances in which the researcher has decided to rely on pre-formulated categories for the classification of the research participants. For instance, the manner in which the federal government has defined various ethnicities and races has changed over time. (See chapter 3.) A study that spans time periods that use different classification systems may find that the choices provided to respondents at the initiation of the study period may no longer be in use towards the end, and researchers may have to reconcile responses to the newer categories.

(2) The focus of the research question. The concepts of race, ethnicity, sex, gender, sexual orientation and related concepts are multidimensional. As an example, depending upon the focus of the research, a determination of ethnicity may require an assessment of the ethnicity of an individual's parents and grandparents in addition to a consideration of the origin of the individual research participant. A sexual history that focuses on the number and sex of one's sexual partners may be sufficient to answer a question focusing on the sex of one's sexual partners, but it may not be adequate to determine an individual's sexual orientation, which is a function of emotional attraction, physical attraction, sexual fantasies, self-definition, and opportunity.

These characteristics are also subject to identification not only by the individual who may be a participant in the research, but also by the observer as well. For instance, an individual may self-identify his or her race (suspending, for the moment, a discussion as to whether race exists), but an individual's race is also subject to the perception of the observer who, on the basis of criteria that he or she as somehow developed, will make a judgment regarding the individual's race. Consequently, it is important to consider in framing the research question whether participants' self-identity as to race, ethnicity, sexual orientation, gender, etc. is important or whether it is the perception of specified observers that is critical. For instance, how an individual identifies him- or herself with respect to race may not be as important as how the individual is viewed by others, if the focus of the study is an exploration of the effects of political marginalization.

Selecting the Sample

How the study sample is selected and the size of the sample are critical issues. A biased sample may lead to erroneous conclusions and an inadequate sample will not have sufficient statistical power to detect the hypothesized effect. This section very briefly reviews issues related to sampling. The issue is a complex one, and readers are referred to other texts for an in-depth exploration of the topic (Cochran, 1977; Kish, 1965; Levy and Lemeshow, 1980).

The sampling procedure is framed around the sampling unit. In many cases, this will be the individual, but it can be a family or household, or area of a community. The sample that will be constructed consists of the sampling units that have been

selected from among those that are eligible for inclusion in the study (Kelsey, Thomson, and Evans, 1986). For example, if an investigator wished to know the proportion of households of a particular ethnic group had health insurance, the sample would consist of a portion of those households where a designated member was of that ethnicity.

The sampling frame refers to the list of the population from which the sample will be drawn. In some cases, this is unknown and unknowable. For instance, a study focusing on the experience of homophobia by gay and lesbian individuals would have a difficult, if not impossible, task to construct a sampling frame, since it would not be possible to know of every individual who self-identified as gay or lesbian, since many may not wish to acknowledge their orientation publicly.

In instances in which it would be difficult to locate and recruit study participants due to the nature of the research, investigators often rely on a convenience sample comprised of volunteers. This strategy, however, can introduce bias into the selection process. Snowball sampling, in which already-recruited participants identify other individuals as potential participants, permits investigators to more easily locate and recruit "hidden" individuals, but may also introduce selection bias because the individuals recruited through respondents are more likely to be like the respondents.

Probability sampling is advisable when it is feasible, in order to reduce the possibility of selection bias. There are four basic designs for probability sampling: simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

Simple random sampling requires knowledge of the complete sampling frame in advance (Kelsey, Thompson, and Evans, 1986). This strategy means that each sampling unit in the population has an equal chance of being selected for participation. This method does not require advance knowledge of the population itself, but may be very inefficient.

Systematic sampling refers to the selection of sampling unit, such as individuals or households, at regularly spaced intervals within the sampling frame, such as every third household. This method has several advantages in that it does not require advance knowledge of the sampling frame, as it can be constructed as the process progresses and is generally relatively simple to implement.

Stratified sampling requires the division of the population into strata and a sample is selected from each such strata. This process is significantly more complex than the other strategies, but offers increased precision and may facilitate the inclusion of specific groups of persons.

Like stratified sampling, cluster sampling divides the sample into groups, such as clusters of homes. A sample is then taken of these clusters for inclusion in the study or, alternatively, a subsample of these sample clusters is utilized. As an example, an investigator wishing to study the prevalence of violence in public housing projects might divide all such projects into clusters by geographic area and then take a sample of these clusters. The households within these clusters could then be queried about the violence in the public housing projects.

Data Collection

Numerous strategies can be used to collect race/ethnicity/sexual orientation data including self-administered questionnaires and surveys, telephone interviews, and face-to-face interviews. Questions can be open-ended, or respondents can be presented with a pre-formulated listing of acceptable responses. Or, researchers may decide to rely on secondary databases and must necessarily, then, utilize the categories embedded in those databases. Depending upon the source of the database, respondents may have had to select their responses from a pre-formulated list, or the categorization of the individual may have been accomplished by an interviewer. The strategies that are selected have implications for the response that will be obtained. These various approaches are compared below.

Self-Completed Instruments versus Interviews

There are advantages and disadvantages associated with either strategy of having respondents complete instruments on their own, or conducting telephone or in-person interviews with respondents.

Self-completed instruments, whether with pencil and paper or through the use of a computer, have the potential to obtain more accurate responses from participants for several reasons. First, because the individual does not have to interact with anyone in giving the response, he or she may be more willing and comfortable to divulge particularly sensitive or embarrassing information in this manner. Computer-assisted self-interviewing (CASI), which permits respondents to type their answers on a computer keyboard in response to items on the computer screen, may be particularly helpful (Camburn and Cynamon, 1993; O'Reilly, Hubbard, Lessler, Biemer, and Turner, 1994). Second, the individual may feel less time pressure to complete the instrument because they are not facing or speaking with anyone directly. As a result, their answers may be more thoughtful.

However, there are several problems associated with self-completed instruments. Individuals may not be able to read or read well and may be embarrassed to disclose this. If this is the situation, they may be tempted to circle any response or write in any number just to complete the form. Unless the instruments are reviewed immediately by someone with the individual still there, it is also possible that individuals may have inadvertently missed items and the instrument remains incomplete. In some cases, depending upon the study design and/or the study population, it may be difficult to relocate or contact the respondent to obtain the missing data. Additionally, self-completed instruments are generally not appropriate for questions that are complex or open-ended and would require lengthier responses (Aday, 1996).

In-person or telephone interviews offer several advantages in that they allow investigators to complete an instrument with a participant, so the participant's ability to read may not be relevant and it is less likely that items contained in the instruments will be inadvertently missed. However, there may be an increased likelihood that answers will be inaccurate of the questions are felt to be embarrassing

or stigmatizing. With phone interviews, particularly if there has not already been a relationship established with the study, there is a greater chance that the prospective participant will simply hang-up or that they will screen calls and not answer (Aday, 1996).

Respondent Self-identification versus Observer Identification

A decision as to who should categorize the research respondent is going to depend to a great extent on the focus of the research question: is it the individual's self-identity that is at issue or the perceptions of others that are most relevant?

By allowing respondents to self-classify with respect to any of the variables of interest discussed in this text, the researcher will be better able to understand definitions and distinctions internal to the community of interest. For example, in a study conducted by Carballo-Diéguez and Dolezal (1994) in which they allowed respondents to self-identify with respect to sexual orientation, they found that among Latino men who have sex with men (MSM) who had had at least one male partner during the previous year, 20% self-identified as bisexual or *hombres modernos* (modern men), 10% self-identified as heterosexual, 65% self-identified as gay, and 4% self-identified as drag queens; 80% of those self-identifying as bisexual had had sex with a woman during the previous year, in comparison with 63% of the men self-identifying as heterosexual; and almost three-fifths of the men self-identifying as gay had had sexual relations with a woman; 8% had had sex with a woman during the previous year. Had they presented respondents with a preformulated list from which to select their sexual orientation, they would not have been able to understand the distinctions made within this community that are relevant to both risk behaviors and prevention interventions.

Allowing respondents to self-identify may then provide the researcher with additional flexibility in the development of the categories to be utilized in the study. A large number of categories resulting from respondent self-definition can be collapsed into fewer categories to increase statistical power.

Participant self-identification may, however, create difficulties for the researcher as well as providing these advantages. The terms selected by respondents to self-identify may not be comparable to categories then in use in the literature, making comparison across research studies difficult. Additionally, respondents may differ in the characteristics they choose as a basis for self-identification. For instance, in asking respondents to describe their ethnicity, some research participants may focus on their country of origin, some on their religion, some on the culture of their parents or grandparents, etc. The researcher may not be aware of the varying criteria used and may have difficulty reframing the responses.

Identification based on observer perception has the advantage of consistent application of pre-specified criteria. However, as described in chapter 3, there is often considerable variance between observer perception of identity and an individual's self-definition. The importance of any resulting difference necessarily depends on the focus of the research.

Pre-formulated Lists versus Open-Ended Questioning

The issue of whether to use pre-formulated lists or open-ended questioning is related to the issue of respondent versus observer identification of an individual. In general, researchers who rely on observer classification of participants with respect to race or ethnicity often work from a pre-formulated list, while the use of open-ended questioning is more frequently employed when relying on participant self-identification.

Pre-formulated lists offer numerous advantages to the researcher. Because there is a predetermined number of categories, analysis may be simpler and, particularly with smaller sample sizes, use of a small number of categories for any particular construct may enhance statistical power. However, reliance on pre-formulated lists presents difficulties if the levels of a variable are overlapping or if they do not consider all possible responses. (See discussion regarding the interpretation of categories, below.)

Regardless of whether one ultimately decides to utilize a pre-formulated list of categories from which to select a response or to have participants answer open-ended questions, the ordering of the questions may be critical. Various approaches are available to order questions. The ordering may be done

- Temporally, from earlier events to more recent events or from more recent events to events occurring in the more distant past
- According to complexity, from simpler topics or concepts to ones that involve increasing complexity
- According to themes, so that questions pertaining to the same theme are grouped together
- By level of abstraction, so that the most concrete items are grouped together and the most abstract are grouped together
- According to level of sensitivity, so that the items that require the greatest level of personal disclosure or focus on the most sensitive topics follow those that are the least sensitive (Schensul, Schensul, and LeCompte, 1999).

Using Scales

There are three primary forms of scales that are often utilized in health research: the Likert, Guttman, and Thurstone Equal-Appearing Interval Scales. Readers are referred to other sources for a more in-depth discussion regarding the construction of scales (Aday, 1996; Spector, 1992).

A Likert scale utilizes an ordinal response scale which allows the respondent to indicate his or her level of agreement of disagreement with a particular item. Five categories of agreement/disagreement are generally used: strongly agree, agree, uncertain or neutral, disagree, strongly disagree. A score is assigned to each such level and the scores are summed across all items to yield a summary score. The Adolescent Survey of Black Life (Table 7 in chapter 6) illustrates the use of a Likert scale.

In contrast, a Guttman scale is premised on the idea that there is a hierarchy in attitudes or perceptions and this hierarchy can be utilized to construct the scale. Positive responses to each item within a hierarchy are totaled to yield a total score (Aday, 1996).

The construction of a Thurstone Equal-Appearing Interval Scale is based on the ratings of items by a selected group of judges as to the extent to which they reflect a negative or positive attitude toward the issue in question. The judges are asked to place these items along an 11-point scale, ranging from most unfavorable to most favorable. The overall degree of favorableness of a particular items is determined by the median value of all of the judges with respect to that particular item. The items that have the least agreement among the judges re eliminated and the remaining ones are incorporated into a questionnaire (Aday, 1996).

Several factors should be considered in deciding whether to construct a new scale for a particular study or to use an existing one. The use of an existing scale may be preferable if it has been shown to have a high degree of reliability and validity and has been tested in the same or similar population as the study population to be assessed. The investigator should also consider whether it is available in the language used by the study population.

Validity

Four types of validity will be discussed here: content validity, criterion validity, construct validity, and factorial validity. *Content validity* refers to the comprehensiveness of the questions asked and whether they adequately reflect the intended goals. For instance, in designing an instrument to assess gender role, the investigator must ensure that all of the questions are relevant to the concept of gender role and that all salient aspects of gender role are covered by the questions. One way to assess the content validity of a proposed instrument is to ask other professionals familiar with the content area to review the items. Focus groups can be conducted with individuals who are representative of the groups with which the instrument is to be used, in order to get their feedback and suggestions to improve both the content of the instrument and the wording of its items. It may be difficult, however, to establish definitively that all of the items included in the instrument reflect all relevant items (Seiler, 1973).

In contrast, the term *criterion validity* refers to the extent to which an instrument correlates with another "gold standard" instrument designed to assess the same factor(s). The term criterion validity can be used with respect to particular items of an instrument or the instrument as a whole. Unfortunately, no instrument exists that is considered the gold standard for the assessment of many of the themes discussed in this text. Indeed, given the diversity that exists within and between communities, it is difficult to conceive of such a gold standard.

The *criterion validity* of an instrument can be assessed by calculating its sensitivity and specificity. *Sensitivity* refers to the proportion of individuals with a particular characteristic who are correctly classified as having that characteristic.

The question then becomes how to determine what the correct classification is. Since there is no instrument that is considered the gold standard, the classification of items by the newly constructed instrument cannot be assessed against the gold standard. One way to accomplish this comparison and assess sensitivity, however, might be to compare the results of the new instrument against individuals' self-classification. Conversely, the term *specificity* is the proportion of individuals without a specific characteristic who are correctly assessed by the instrument as being without that characteristic. The sensitivity and specificity of an instrument can be combined to give a single measure of accuracy. (For a discussion of how to calculate sensitivity, specificity, and accuracy, see McDowell and Newell, 1996; Morgenstern, 1996).

Construct validity requires that a conceptual definition of the construct be formulated, including the components of that concept. There is no perfect way to assess construct validity. Instead, construct validity may be suggested by the extent to which the results produced by the newly designed instrument correlate or do not correlate with other assessment instruments designed to measure the same constructs. For example, a high degree of correlation between the findings of a newly designed instrument to assess gender role and a pre-existing instrument used to assess the same construct would suggest that the new instrument displays *convergent validity*, said to be equivalent to assessing sensitivity (McDowell and Newell, 1996). If the new instrument does not correlate well with other instruments designed to assess different themes, it can be said that it displays *divergent validity*. However, it is unclear how high a level of correlation is required to say that there is adequate correlation (McDowell and Newell, 1996).

Factor analysis is often used to examine the conceptual structure of an instrument by assessing how well the items of the instrument fall into expected groupings. Using again the example of an instrument designed to measure gender role across various contexts, we might use factor analysis to determine whether the questions fall into two or more distinct groups such as masculinity and femininity. These groupings should be homogenous and unrelated to each other. (For a discussion of the difficulties associated with such a distinction, see chapter 3.)

The appropriate use of factor analysis requires that the variables be assessed using an interval-level scale (McDowell and Newell, 1996); reliance on interval-level scales may be somewhat rare in the context of assessing the constructs discussed in this text. Where this approach is used, though, it is important that there be at least five times the number of respondents in the sample than there are variables to be used in the analysis (McDowell and Newell, 1996). However, many journal articles may indicate that factor analysis was used to evaluate the content validity of instruments that utilize categorical responses, such as "never," "sometimes," "frequently," "always."

Reliability

In addition to assessing the validity of an instrument, it is important to determine its reliability. *Reliability* refers to the consistency of a measurement across time,

respondents, and/or observers. Reliability is said to consist of two components: the true value of the measurement and a degree of error in the measurement that is obtained. Reliability is concerned with that portion of the measurement error that is random; the portion that is not random, or systematic, is referred to as bias. (Bias is discussed below in the context of misclassification.) Random error may occur for any number of reasons including interviewer fatigue, carelessness, and/or interviewee fatigue.

Inter-rater agreement or reliability refers to the extent to which different raters assess the respondent similarly. Inter-rater reliability for nominal or categorical data, such as categories of sexual orientation or ethnicity, can be reported using the Kappa coefficient. The kappa coefficient is obtained by constructing a table that indicates the proportion of agreement between the two raters. A weighted kappa formula is useful in discriminating between minor and major discrepancies between raters (Streiner and Norman, 1989).

Intra-rater reliability or test-retest reliability pertains to the assessment of a respondent by the same rater, and the extent to which a second assessment is consistent with the first, respectively. It has been recommended that the time interval between the assessments be brief in order to reduce the risk that an instrument will erroneously appear to be unreliable when it is actually detecting changes that have occurred between the assessments (McDowell and Newell, 1996). However, if the successive administrations of the instrument are spaced closely in time, the rater and/or the respondent may remember the answers to the previously administered assessment and this may influence the results of the subsequent assessment.

Various strategies have been formulated in an attempt to reduce these possibilities. The subsequent assessment may utilize an instrument that is parallel to the first, but that is not the same. The assessment of reliability in this situation would focus on the level of correlation between the two results. Alternatively, two equivalent but not identical versions of the same test can be merged into a single instrument to be utilized in a single session. Reliability is assessed by determining the comparability of the results if the measurement had been divided into two component versions (McDowell and Newell, 1996). This can be done by correlating odd- and even-numbered questions or by estimating correlations between all possible pairs of items. A greater level of correlation among the items will facilitate the correlation of two equivalent versions.

It should be noted that a higher level of internal consistency will produce greater test-retest reliability. Cronbach's alpha is often utilized to assess internal consistency. An unsatisfactory score for internal consistency can sometimes be improved by deleting items from the instrument that do not correlate highly with other items. However, the deletion of items that are critical to the construct(s) under evaluation may threaten the content validity of the instrument.

Misclassification

Misclassification, which is a form of information bias, occurs when the exposure or outcome status of a study participant is erroneously classified. Where the error depends on the value of other variables, the misclassification is said to be

differential; nondifferential misclassification occurs when the classification error is not dependent on the values of other variables (Rothman and Greenland, 1998).

Assume, for instance, that an investigator wishes to evaluate the relationship between race/ethnicity and risk of lung cancer. The nondifferential misclassification of race or ethnicity could potentially mask any association that might exist and, if severe enough, could even reverse the direction of the association.

Similarities between some selection biases and misclassification bias are apparent. For instance, assume that an investigator wishes to oversample study participants from a particular ethnic group. Assume further that a portion of those individuals are erroneously classified as members of a different ethnic group and are considered eligible or ineligible for study participation on the basis of this erroneous classification. This is an example of selection bias. The information that will be derived from their participation in the study may suffer from information bias due to the continuing misclassification of those individuals who have been enrolled into the study on the basis of this erroneous classification.

The constructs that are the focus of this text, such as sex and ethnicity, are often considered to be confounding variables where they are associated with both the disease and the exposure under investigation, they are associated with the exposure among the source population for cases, and they are not on the causal pathway between the exposure and the disease. The nondifferential misclassification of a confounding variable will reduce the extent to which the confounding may be controlled. As a consequence, bias may occur either towards or away from the null value, depending upon the direction of the confounding. The results can be especially misleading if there is a weak association between the exposure and the disease of interest and the confounding is strong (Rothman and Greenland, 1998).

Interpreting Categories

The construction of categories often requires interpretation. Yanow (2003) has identified six features: (1) category errors, (2) a defining point of view, (3) tacit knowledge, (4) marking, (5), occluded features and silences, and (6) situated, local knowledge and change. The construction of categories implies, first, that everything that can be encompassed within a category actually is and, second, that there is no overlap in these characteristics across the named categories (Yanow, 2003). Difficulties occur when items do not fall within any of the named categories or when their characteristics permit their classification into more than one of the existing categories. As an example, true hermaphrodites may be considered to be of both male and female sexes or of neither male nor female sex. (See chapter 4 for further discussion regarding categories of sex.)

In discussing "a defining point of view," Yanow (2003) is referring to the manner in which categories are constructed, through the shared logic of a group of people about what characteristics are most salient to the construction of categories. The logic that underlies this shared category-making is often tacit, and may not appear logical to members of other groups ("tacit knowledge"). For instance, the United

States until recently has utilized a "one drop rule," whereby any indication of black ancestry indicated that an individual was racially black. This system might seem less than logical to individuals from societies in which skin color is viewed on a continuum, extending from very light to very dark. Tacit knowledge also operates where deviations from the norm ("marked cases") are assessed against those that are considered prototypical. What is considered deviant, or "marked," may not be obvious to those outside of the group constructing the category on the basis of tacit knowledge. The classification of homosexuality as deviant behavior is premised on a view of what is normal sexual behavior that is not universally shared.

A focus on specific features in the construction of categories may deflect attention from other, critical features, thereby occluding or obscuring them. A focus on skin color to explain health disparities, for instance, may deflect attention from issues that may be equally important or even more important to an understanding of existing disparities, such as differences in socioeconomic status or inability to access care due to lack of medical insurance.

Categories are not fixed; they are situated in local knowledge and, therefore, may change over time and reflect changes in local knowledge over time. Homosexuality was once categorized by the American Psychiatric Association as a mental illness, but is no longer included as such in its nosology. (See chapter 4.) Intersexuality continues to be viewed, in general, as an abnormal condition, but this understanding is being challenged in significant ways and it is possible that intersex conditions may, in the future, be considered as yet another reflection of biological diversity, rather than a condition requiring a cure. (See chapter 3). If this were to occur, the categorization of intersexuality as a medical condition in need of treatment would no longer be valid.

Summary

This chapter has explored the meaning of "category" and the various ethical implications and methodological complexities associated with the construction of categories. It is critical that researchers be cognizant of intragroup classifications in formulating categories to be used in their research and the risks that participants may face as a result of being associated with those categories. Researchers must also consider the validity and reliability of the measures used to delineate between various categories, the implications of misclassification, and the complexities involved in interpreting the resulting findings.

References

Aday, L. (1996). *Designing and Conducting Health Surveys*, 2nd ed. San Francisco, California: Jossey-Bass Publishers.

Carballo-Diéguez, A., Dolezal, C. (1994). Contrasting types of Puerto Rican men who have sex with men (MSM). *Journal of Psychology and Human Sexuality* 6(4): 41–67.

Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley.

Kelsey, J.L., Thomson, W.D., Evans, A.S. (1986). *Methods in Observational Epidemiology*. New York: Oxford University Press.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Levy, P.S., Lemeshow, S. (1980). *Sampling for Health Professionals*. Belmont, California: Lifetime Learning Publications.

McDowell, I., Newell, C. (1996). *Measuring Health: A Guide to Rating Scales and Questionnaires,* 2nd ed. New York: Oxford University Press.

Morgenstern, H. (1996). *Course Materials, Part II: Class Notes for Epidemiologic Methods, Epidemiology 201B*, University of California Los Angeles.

Rothman, K.J., Greenland, S. (1998). *Modern Epidemiology*, 2nd ed. Philadelphia, Pennsylvania: Lippincott-Raven Publishers.

Schensul, S.L., Schensul, J.J., LeCompte, M.D. (1999). *Essential Ethnographic Methods: Observations, Interviews, and Questionnaires.* Walnut Creek, California: AltaMira Press.

Seiler, L.H. (1973). The 22-item scale used in field studies of mental illness: A question of method, a question of substance, and a question of theory. *Journal of Health and Social Behavior* 14: 252–264.

Spector, P.E. (1992). *Summated Rating Scale Construction: An Introduction*. Thousand Oaks, California: Sage.

Streiner, D.L., Norman, G.R. (1989). *Health Measurement Scales: A Practical Guide to Their Development and Use*. New York: Oxford University Press.

Yanow, D. (2003). *Constructing "Race" and "Ethnicity" in America: Category-Making in Public Policy and Administration*. Armonk, New York: M.E. Sharpe.