DISSERTATION

RAND

Assessing Patient Experiences with Healthcare in Multi-Cultural Settings

Leo Sergio Morales

RAND Graduate School

DISSERTATION

RAND

Assessing Patient Experiences with Healthcare in Multi-Cultural Settings

Leo Sergio Morales

RAND Graduate School

This document was prepared as a dissertation in September 2000 in partial fulfillment of the requirements of the doctoral degree in public policy analysis at the RAND Graduate School. The faculty committee that supervised and approved the dissertation consisted of Ronald Hays (Chair), José Escarce, and Dana Goldman.

The RAND Graduate School dissertation series reproduces dissertations that have been approved by the student's dissertation committee.

RAND is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND® is a registered trademark. RAND's publications do not necessarily reflect the opinions or policies of its research sponsors.

Published 2001 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, Virginia 22202-5050
201 North Craig Street, Suite 102, Pittsburgh, Pennsylvania 15213
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact RAND Distribution Services (Telephone: 310-451-7002; Fax: 310-451-6915; or E-mail: order@rand.org). Abstracts of all RAND documents may be viewed on the World Wide Web (http://www.rand.org). Publications are distributed to the trade by National Book Network.

Acknowledgements

Many outstanding collaborators have made the research presented in this dissertation possible. First and foremost, I would like to acknowledge my committee chair, Ron Hays, Ph.D., Professor, UCLA, for his intellectual leadership, career guidance, and unwavering support over the past several years. I would also like to acknowledge my committee members – José Escarce, M.D., Ph.D., and Dana Goldman, Ph.D. -for their helpful comments and my coauthors who include Robert Weech-Maldonado, Ph.D, Assistant Professor, Pennsylvania State University; William Cunningham, M.D., Associate Professor, UCLA; Marc Elliott, Ph.D., Associate Statistician, RAND; Steve Reise, Ph.D., Associate Professor, UCLA; Grant Marshall, Ph.D., Behavioral Scientist, RAND; Honghu Lui, Ph.D., Assistant Adjunct Professor, UCLA; Karen Spritzer, B.A., UCLA; Beverly Weidmer-Ocampo, M.A., RAND Survey Research Group; and Julie Brown, B.A., RAND Survey Research Group; for their creative and scholarly contributions.

Helpful reviews of the research in this dissertation were also been provided by Robert Brook, M.D., Sc.D., Vice President, RAND; Martin Shapiro, M.D., Ph.D., Professor, UCLA; and Neil Wenger, M.D., Associate Professor, UCLA. In addition, Dr. Shapiro was kind enough to cover for me on the hospital ward rounds while I completed work on my dissertation.

Support for the research presented in this dissertation is from several sources. I received support from a training grant from the Health Resources and Services Administration to UCLA, (Shapiro, PI); an unrestricted research grant from The Medical Quality Commission to RAND

(Hays, PI); a minority supplement grant from the Agency for Healthcare Research and Quality to RAND (Hays, PI; Morales, Co-PI); and an unrestricted research grant from the Robert Wood Johnson Foundation to UCLA (Morales, PI).

To my parents, Leo and Gladys, who gave me the vision and encouragement to pursue higher education. To my wife Laurie, who has loved me through the good times and the hard times, and who has given me my greatest joy in life, Carlos Rubén Morales.

- Leo Sergio Morales

ABSTRACT OF DISSERTATION

Assessing Patient Experiences with Healthcare in Multi-Cultural Settings

By

Leo Sergio Morales

Doctor of Philosophy in Policy Analysis

RAND Graduate School

2000

Ron D. Hays, Ph.D., Chairperson

This dissertation examines patient experiences with healthcare in multi-cultural settings. The first three chapters present a theoretical conceptualization of patient satisfaction, a general framework for producing culturally appropriate survey instruments, and an assessment of the readability level of the Consumer Assessments of Health Plans Study (CAHPS®) 2.0 surveys. The results of the readability assessment, which are based on readability formulas, show that both the Spanish and English versions of the CAHPS® 2.0 survey instruments require approximately a seventh grade reading ability.

The next three chapters present results from empirical studies examining racial/ethnic differences in reports and ratings of care. The first study is based on 7,093 patient surveys collected in English and in Spanish for the United Medical Group Association (UMGA) study and examines differences among Hispanics responding in Spanish, Hispanics responding in English, and non-Hispanics whites responding in

English in ratings of communication by doctors. After controlling for age and gender, Hispanics responding in Spanish rated communication by their doctors lower than Hispanics responding in English ($p<0.05$) and non-Hispanics whites responding in English ($p<0.05$).

The next two studies are based on the National CAHPS® Benchmarking Database (NCBD) 1.0. The NCBD 1.0 contains 28,354 completed adult and 9,870 child CAHPS® 1.0 surveys from 118 health plans across the United States. Reports about access to care, promptness of care, communication by doctors, the courtesy and helpfulness of doctor's office staff, health plan customer service, and global ratings of doctors (both personal doctors and specialists), healthcare and health plans were examined for racial/ethnic differences. After controlling for age, gender, education, and health status, significant differences were found among the racial/ethnic groups examined. Generally, Hispanic and Asian/Pacific Islander adults reported worse care compared to white adults, Black adults reported better care compared to white adults, and Native Alaskan/American Indian adults reported similar care compared to white adults. Comparable trends were found among respondents to the child surveys with the exception that Blacks reports worse care for their children than whites.

The next two chapters of the dissertation examine the psychometric properties of measures included in the UMGA and NCBD 1.0 databases. The first study uses item response theory procedures to test for differential item functioning among white and Hispanic survey respondents to the UMGA survey. The results indicate that despite some differences in item functioning, valid comparisons between whites and

Hispanics on indicators of satisfaction with care are possible. The second study yields similar results from a confirmatory factor analysis of the CAHPS® 1.0 measures among whites and Hispanics.

The final chapter of the dissertation summarizes the findings of the empirical investigations conducted, and derives policy implications from these studies. The dissertation concludes by noting that language barriers remain salient for minority patients who have gained access to the health care system. Greater efforts need to be directed at improving care for racial/ethnic minorities.

Table of Contents

1. Introduction: Assessing Racial and Ethnic Differences in Patient Evaluations of Care

Background

Improving the quality of health care and reducing racial/ethnic disparities in health are two principal objectives of current national health policy. Mounting evidence suggests that inequalities in the quality of care received by subgroups of the population contribute to the disparities in health we observe (Fiscella, Franks, Gold, & Clancy, 2000). To improve the quality of health care for racial and ethnic minorities and thereby reduce disparities in health, reliable and valid measures of health care are needed, particularly measures that are applicable across many different racial/ethnic and linguistic groups.

Largely a result of the consumer movement, patient evaluations of health care have emerged as one of the most important commonly collected indicators of quality of care. Patient satisfaction - one type of patient evaluation - is widely acknowledged by investigators and policy makers as an essential outcome of health care, distinct from the efficacy of care (Cleary & McNeil, 1988). Further, patient evaluations of health care have been linked to several important health-related behaviors including the initiation of malpractice litigation (Penchansky & Macnee, 1994; Vaccarino, 1977), disenrollment from health plans and providers (Allen & Rogers, 1997; Newcomer, Preston, & Harrington, 1996; Schlesinger, Druss, & Thomas, 1999), and adherence to medical regimens, including keeping follow-up appointments with health care providers (Hall, Milburn, Roter, & Daltroy, 1998).

Thus, patient satisfaction has been implicated as both an antecedent and consequence of good health (Marshall, Hays, & Mazel, 1996; Hall Judith A, Roter Debra L, & Milburn Michael A, 1999).

In response to the growing demand for a state of the art survey instruments to assess patient evaluations of health care, the Agency of Healthcare Research and Quality (AHRQ) funded a national effort called the Consumer Assessments of Health Plans Study (CAHPS®) (Crofton, Lubalin, & Darby, 1999). The goal of CAHPS® was to develop reliable and valid measures of patient evaluations of health care that are applicable to patients throughout the life cycle and across a variety of settings. The primary intended use of the CAHPS® surveys was to inform consumers about the experiences of other consumers with the health plans available to them. In response to these objectives, the CAHPS® research consortium (Harvard Medical School, RAND, Research Triangle Institute, and Westat) developed survey instruments applicable to adults and children, in managed care and fee-for-service settings. Because the CAHPS® consortium recognized the growing diversity of the US population, surveys were translated into Spanish (Weidmer, Brown, & Garcia, 1999). In addition to the surveys, scoring algorithms and reporting formats were developed.

Currently, many large providers and purchasers of care routinely assess health care using the CAHPS® survey instruments. Medicare and the Office of Personnel Management survey their beneficiaries yearly (Schnaier et al., 1999). Many state Medicaid programs including California, Texas and Washington State have adopted the CAHPS® surveys as part of their routine quality assurance and quality improvement

strategy (Brown, Nederend, Hays, Short, & Farley, 1999). The National Council on Quality Assurance, the largest accrediting body of Health Maintenance Organizations, requires health plans to administer the CAHPS® surveys for accreditation (National Committee for Quality Assurance, 2000).

Patient evaluations of health care can yield important insights about how well different subgroups within populations are being served by the health care system. Patient evaluations have been successfully used to assess the quality of medical care services among Hispanic, Asian, American Indian, and African American patients (Morales, Cunningham, Brown, Liu, & Hays, 1999; Murray-Garcia, Selby, Schmittdiel, Grumbach, & Quesenberry, 2000; Meredith & Siu, 1995; Taira et al., 1997; Morales, Elliott, Weech-Maldonado, Spritzer, & Hays).

A strength of patient evaluations is that they can implicitly and explicitly assess the cultural and linguistic appropriateness of health services. They can implicitly assess the cultural and linguistic appropriateness of care because they capture experiences with care from the patient's perspective, thus they incorporate the cultural lens through which patients experience health care. Patient evaluations can explicitly measure cultural and linguistic appropriateness of care with the inclusion of survey questions asking about domains of quality of care related to cultural and linguistic appropriateness (i.e., interpreter services, non-English patient materials). Thus, patient evaluations may be one of the best tools available to policy makers for assessing and monitoring racial/ethnic disparities in quality of care.

Collecting reliable and valid consumer data in culturally and socioeconomically diverse populations is complex. However, without it, accurate assessments and monitoring of racial/ethnic disparities in care is not possible. Researchers concerned about the quality of survey data have raised methodological concerns about the use of consumer surveys in culturally and linguistically diverse patient populations. In addition to concerns about providing adequate translations into multiple languages (Herdman, Fox-Rushby, & Badia, 1997; Badia, Garcia-Losa, & Dal-Re, 1997; Bullinger et al., 1998), there is concern about cultural differences in the interpretation of questions (Angel & Thoits, 1987),(Liang, Van Tran, Krause, & Markides, 1989; Weissman, Sholomskas, Pottenger, Prusoff, & Locke, 1977) response styles (Hayes & Baker, 1998), and the literacy requirements to comprehend survey questions (Brown et al., 1999). Although methods for empirically testing surveys for measurement equivalence or survey scales across groups are available (Smith Larissa L & Reise Steven P, 1998; Widaman Keith F & Reise Steven P, 1997; Reise Steven P, Widaman Keith F, & Pugh Robin H, 1993), these methods have been rarely applied to patient evaluations of health services.

In this thesis, two large data sources are analyzed for racial/ethnic differences in patient evaluations of health care. The data sources are the National CAHPS® Benchmarking Database (NCBD) 1.0 and the United Medical Group Association (UMGA) study database (Hays, Brown, Spritzer, Dixon, & Brook, 1998). The NCBD 1.0 is an aggregation of CAHPS 1.0 survey results from across the United States. The NCBD project is administered by QMAS with finding from the AHRQ and the Health Care Financing Administration (HCFA). Both adult and child

survey results from Medicaid and commercial settings are included in NCBD 1.0. The UMGA database contains survey results from a probability sample of patients receiving care from 63 physician groups located on the west coast of the United States. The UMGA database only contains results from adult survey respondents. A particular strength of this database for conducting cross-cultural research is that it includes over 150 surveys completed in Spanish.

This thesis also presents methods for producing reliable and valid survey instruments to assess patient evaluations of care in multicultural settings. A general framework for producing culturally appropriate survey instruments is developed and psychometric analyses are conducted to evaluate the equivalence of scales contained in the CAHPS® and UMGA surveys between racial/ethnic groups. The specific psychometric analyses conducted include an evaluation of the measurement equivalence of the of satisfaction ratings from the UMGA database between whites and Hispanics and an evaluation of the factor structure of the CAHPS® 1.0 measures among whites and Hispanics using the NCBD 1.0 database.

In following sections of this chapter, a detailed outline of this thesis is presented and theoretical foundation of patient satisfaction is briefly reviewed. Finally, the implications of patient satisfaction theory for the application of patient evaluation research in multicultural setting are discussed.

Thesis Outline

Chapter 2 presents a general framework for producing culturally appropriate survey instruments. Chapter 3 presents an analysis of the readability of the CAHPS® 2.0 survey instruments. Chapters 4 to 6 present three separate studies of racial/ethnic differences in patient evaluations of health care. Chapter 4 is an analysis of the UMGA data focusing on differences in patient evaluations of physician communication among English speaking non-Hispanic whites, English speaking Hispanics, and Spanish speaking Hispanics. Chapter 5 is an analysis of the NCBD 1.0 data focusing on differences in adult patient evaluations among Hispanics, Whites, Blacks, Asians, and American Indians. Chapter 6 is also an analysis of the NCBD 1.0 database, however, it focuses on differences in proxy evaluations of care delivered to Hispanic, white, black, Asian, and American Indian children. Chapter 7 and 8 examine the psychometric properties of patient evaluations of care included in the UMGA and CAHPS® databases. Chapter 7 uses item response theory procedures to test the metric equivalence of UMGA satisfaction rating measures among whites and Hispanics. Chapter 8 examines the factor structure of the CAHPS® 1.0 measures among whites and Hispanics. Chapter 9 summarizes the findings from the investigations conducted in the thesis, drawing out the policy implications of these results.

Conceptualizing Patient Satisfaction

Most research on patient evaluations has focused on patient satisfaction and correlates of patient satisfaction. Understanding the theoretical conceptualization of patient satisfaction is important for identifying methodological issues that might arise when patient

satisfaction measures and other patient evaluations are used in multicultural settings. Thus, we begin by reviewing sociological, psychological, and health services research theories pertaining to patient satisfaction.

Several recent reviews have summarized the literature on patient satisfaction. Sherbourne and Hays (1995), Pasco (1983) and van Campen (1995) reviewed patient satisfaction with primary care services, Lebow (1983) and El-Guebaly (1983) looked at satisfaction with mental health services, and Lochman (1983) described satisfaction with medical consultants (van Campen, Sixma, Friele, Kerssens, & Peters, 1995; Pascoe, 1983; Lebow, 1983; el-Guebaly, Toews, Leckie, & Harper, 1983; Lochman, 1983; Sherbourne, Hays, & Burton, 1995). Without exception, these reviews were critical of the existing research on patient satisfaction. Regarding patient satisfaction with ambulatory services, Pasco (1983) noted that there was very little theory or model development, little standardization of measurement instruments, low reliability of instruments, and uncertainty about the validity of instruments. Van Campen confirmed Pascoe's earlier findings, noting that the research conducted on patient satisfaction lacked sufficient theoretical foundations, and that most of the instruments lacked methodological rigor regarding the reliability and validity of subscales.

Exceptions, however, to the vast majority of atheoretical research on patient satisfaction exist. In her seminal research, Linder-Pelz (1982) used several types of social and psychological theories, included discrepancy theories, fulfillment theories, and equity theories, to formulate hypotheses about the determinants of

patient satisfaction (Linder-Pelz, 1982). These theories fall under the general rubric of the "disconfirmation paradigm"(Zegers, 1968), in which satisfaction is determined by the disparity between a standard (expectancies, values, or norms) and perceived occurrences.

In discrepancy theories, satisfaction is conceptualized as the difference between what actually occurs and what is expected, adjusted for what is expected. Mathematically, discrepancies theories can be formulated as follows:

$$Satisfaction = \frac{(E - O)}{E},$$

where *E* is what is expected and *O* is what actually occurs. In fulfillment theories, satisfaction is conceptualized as the simple difference between what is expected and what occurs, unadjusted for how much is desired in the first place. Mathematically, fulfillment theories can be formulated as follows:

$$Satisfaction = E - O,$$

where *E* and *O* are as defined above. Finally, in equity theories, satisfaction is a function of whether people perceive they are being treated fairly. Equity theories differ from fulfillment and discrepancy theories in that they stress the importance of *interpersonal* comparisons between how one is treated and how others are treated rather than *intrapersonal* comparisons between one's own expectations and perceptions of what occurs.

Sophisticated conceptual models of patient satisfaction that incorporate disconfirmation theories of satisfaction have been

constructed. Thompson and Suñol (Thompson & Suñol, 1995) recently proposed model of patient satisfaction based on marketing research conducted by Anderson (Anderson Rolph E, 1973) and Parasuraman (Parasuraman, Berry, & Zeithaml, 1991). The assimilation-contrast model of perceptions proposed by Anderson draws heavily from cognitive dissonance theory. In it, he proposes that when perceptions of attribute performance differ only slightly from expectations, there is a tendency for people to displace their perceptions towards their expectations; the assimilation effect. However, there is a point on either side of this range beyond which people can no longer effect displacement but begin to exaggerate the increasing difference between perceptions and expectations; the contrast effect.

Figure 1 (page 21)depicts Anderson's model. The horizontal axis represents actual or objective attribute performance, the vertical axis represents perceived attribute performance, and the diagonal axis represents expectations. When the difference between expectations and actual attribute performance are small (between arrows), differences between perceptions and actual attribute performance are minimized. On the other hand, when expectations and actual between expectations and actual attribute performance are large (outside arrows), differences between perceptions and actual attribute performance are exaggerated.

Parasuraman's model (see Figure 2, page 22) posits a zone of tolerance as a range between adequate and desired levels of service expectations. The zone of tolerance in this model corresponds to the assimilation effects proposed in the previous model. Parasuraman's model takes the additional step of distinguishing between process and outcome expectations. This distinction seems to make sense in the health care

context, where patients might hold different expectations for process and outcomes. For example, the quality of hospital food might have a larger zone of tolerance and lower level of expected service performance than treatment efficacy.

Thompson and Suñol reject the notion of an "objective" measure of attribute performance. In their model (see Figure 3, page 23), attribute performance is judged only by service users on perceptual terms. Initial perceptions of attribute performance are represented by a downward sloping diagonal axis and post-assimilative/contrast perceptions are represented by an upward sloping diagonal axis. A zone of tolerance around predicted expectations is posited, bounded by a minimum predictable level and an achievable normative level, on the assumption that normative expectations will exceed predicted expectations. When initial perceptions exceed predicted expectations within the zone of tolerance, the model posits a smaller amount of satisfaction than predicted by initial perceptions alone due to an assimilation effect. However, when initial perceptions exceed predicted expectations outside the zone, the model posits more satisfaction than predicted by initial perceptions alone due to a contrast effect. Thompson and Suñol propose that the curves represented in their model differ across domains of patient evaluation.

Conceptualizing Expectations

As hinted at in the proceeding section, the ways that expectations are conceptualized vary among researchers. In a recent review of the literature on expectations, Thompson and Suñol identified four types of expectations: predictions, ideals, normative standards,

and unformed expectations (Thompson & Suñol, 1995). Researchers, who conceptualize consumer expectations as predictions, believe consumer expectations are predictions about what is likely to happen during an impending exchange or encounter. For example, Oliver stated, "It is generally agreed that expectations are consumer-defined probabilities of the occurrence of positive and negative events if the consumer engages in some event" (Oliver Richard L, 1981). In contrast, researchers who conceptualize consumer expectations as ideals, refer to expectations as the desires of consumers (i.e., what consumers want rather than what will be offered).

Expectations have also been conceptualized as normative standards. In this case, expectations are about what should or ought to happen during an impending transaction or exchange, rather than what is expected or desired. Normative expectations can be equated with what consumers have been told, or led to believe, or personally deduced that they ought to receive from health services. Normative expectations are related to a subjective evaluation of what is deserved in a situation, and to some extent is also a socially endorsed evaluation.

Finally, Thompson and Suñol (1995) defined a fourth type of expectations, unformed expectations. Unformed expectations occur when consumers are "unable or unwilling, for various reasons, to articulate their expectations, which may be because they do not have any, or find it too difficult to express, or do not wish to reify their feeling, due to fear, anxiety, conformity to social norms, etc. This may be a temporary phenomenon prior to the experience and the gaining of knowledge. It may include 'taken for granted' attributes of care" [p.

130]. The authors argue that unformed expectations may be quite common in health care settings, where previously healthy persons may encounter many new aspects of the health care system once they become sick. Thus they may encounter the health care system without preformed expectations.

Implications for Patient Evaluation Research in Multicultural Settings

Regardless of how satisfaction is modeled, a person's racial/ethnic background can have an important influence on his/her evaluations of healthcare. All patient satisfaction theories incorporate expectations as determinants of satisfaction, and research on the determinants of expectations suggests that sociodemographic factors including age, gender, and racial/ethnic background influence expectations (Kravitz, 1996).

Past experiences are thought to be important in shaping predictions about future experiences. Thus patient expectations about future contacts with the health care system, conceptualized as predictions, are likely influenced by past experiences with health care and the health care system. Because racial/ethnic minorities tend to be treated at different hospitals than whites (Blacks tend to receive care at teaching hospitals while whites tend to receive care at non-teaching hospitals) (Kahn et al., 1994), expectations about the quality of care may also differ. Thus Blacks and whites may judge current experiences with care differently as a result of different past experiences with care.

A person's culture shapes both their normative and ideal expectations. For example, the low regard of the Hmong people for western health care has been documented (Fadiman, 1998). An examination of the practices of their traditional healers reveals that they routinely spend up to four hours with each patient during a consultation, render diagnoses without blood tests or extensive personal histories, and that physical examinations of women never include vaginal pelvic exams. Although the Hmong are reported to view western medicine as sometimes helpful, it is easy to see how their expectations about health care would greatly deviate from their experiences, resulting in dissatisfaction. Clearly, comparing patient evaluations given by the Hmong and whites would have different implications than comparisons between Blacks and whites.

Research conducted in this thesis addresses the need for assessing racial/ethnic differences in quality of care, while acknowledging the methodological complexities of making racial/ethnic comparisons. Methods to minimize the potential for biased instruments are discussed, and empirically studies to examine survey scales for bias are conducted. At the same time, studies to assess quality of care differences among racial/ethnic groups are conducted. Although studies to test for differences in expectations across racial/ethnic groups are not conducted in this thesis, this topic is left to future research. However, understanding the determinants of patient evaluations of care is helpful in illuminating the complex issues that arise when studies that make racial/ethnic comparisons are undertaken.

References

Allen, H. M. Jr, & Rogers, W. H. (1997). The consumer health plan value survey: round two. Health Aff (Millwood), 16(4), 156-66.

Anderson Rolph E. (1973). Consumer dissatisfaction: the effect of disconfirmed expectancy on perceived product performance. Journal of Marketing Research, Vol. 10(1), 38-44.

Angel, R., & Thoits, P. (1987). The impact of culture on the cognitive structure of illness. Cult Med Psychiatry, 11(4), 465-94.

Badia, X., Garcia-Losa, M., & Dal-Re, R. (1997). Ten-language translation and harmonization of the International Prostate Symptom Score: developing a methodology for multinational clinical trials. Eur Urol, 31(2), 129-40.

Brown, J. A., Nederend, S. E., Hays, R. D., Short, P. F., & Farley, D. O. (1999). Special issues in assessing care of Medicaid recipients. Med Care, 37(3 Suppl), MS79-88.

Bullinger, M., Alonso, J., Apolone, G., Leplege, A., Sullivan, M., Wood-Dauphinee, S., Gandek, B., Wagner, A., Aaronson, N., Bech, P., Fukuhara, S., Kaasa, S., & Ware, J. E. Jr. (1998). Translating health status questionnaires and evaluating their quality: the IQOLA Project approach. International Quality of Life Assessment. J Clin Epidemiol, 51(11), 913-23.

Cleary, P. D., & McNeil, B. J. (1988). Patient satisfaction as an indicator of quality care. Inquiry, 25(1), 25-36.

Crofton, C., Lubalin, J. S., & Darby, C. (1999). Consumer Assessment of Health Plans Study (CAHPS). Foreword. Med Care, 37(3 Suppl), MS1-9.

el-Guebaly, N., Toews, J., Leckie, A., & Harper, D. (1983). On evaluating patient satisfaction: methodological issues. Can J Psychiatry, 28(1), 24-9.

Fadiman, A. (1998). The spirit catches you and you fall down : a Hmong child, her American doctors, and the collision of two cultures. New York, New York: Noonday Press.

Fiscella, K., Franks, P., Gold, M. R., & Clancy, C. M. (2000). Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. JAMA, 283(19), 2579-84.

Hall, J. A., Milburn, M. A., Roter, D. L., & Daltroy, L. H. (1998). Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol, 17(1), 70-5

Hall Judith A, Roter Debra L, & Milburn Michael A. (1999). Illness and satisfaction with medical care. Current Directions in Psychological Science, 8(3), 96-99.

Hayes, R. P., & Baker, D. W. (1998). Methodological problems in comparing English-speaking and Spanish- speaking patients' satisfaction with interpersonal aspects of care. Med Care, 36(2), 230-6.

Hays, R. D., Brown, J. A., Spritzer, K. L., Dixon, W. J., & Brook, R. H. (1998). Member ratings of health care provided by 48 physician groups. Arch Intern Med, 158(7), 785-90.

Herdman, M., Fox-Rushby, J., & Badia, X. (1997). 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res, 6(3), 237-47.

Kahn, K. L., Pearson, M. L., Harrison, E. R., Desmond, K. A., Rogers, W. H., Rubenstein, L. V., Brook, R. H., & Keeler, E. B. (1994). Health care for black and poor hospitalized Medicare patients. JAMA, 271(15), 1169-74.

Kravitz, R. L. (1996). Patients' expectations for medical care: an expanded formulation based on review of the literature. Medical Care Research and Review, 53(1), 3-27.

Lebow, J. L. (1983). Research assessing consumer satisfaction with mental health treatment: a review of findings. Eval Program Plann, 6(3-4), 211-36.

Liang, J., Van Tran, T., Krause, N., & Markides, K. S. (1989). Generational differences in the structure of the CES-D scale in Mexican Americans. J Gerontol, 44(3), S110-20.

Linder-Pelz, S. U. (1982). Social psychological determinants of patient satisfaction: a test of five hypothesis. Social Science and Medicine, 16(5), 583-9.

Lochman, J. E. (1983). Factors related to patients' satisfaction with their medical care. J Community Health, 9(2), 91-109.

Marshall, G. N., Hays, R. D., & Mazel, R. (1996). Health status and satisfaction with health care: results from the medical outcomes study. J Consult Clin Psychol, 64(2), 380-90.

Meredith, L. S., & Siu, A. L. (1995). Variation and quality of self-report health data. Asians and Pacific Islanders compared with other ethnic groups. Med Care, 33(11), 1120-31.

Morales, L. S., Cunningham, W. E., Brown, J. A., Liu, H., & Hays, R. D. (1999). Are Latinos less satisfied with communication by health care providers? J Gen Intern Med, 14(7), 409-17.

Morales, L. S., Elliott, M. N., Weech-Maldonado, R., Spritzer, K. L., & Hays, R. D. Differences in CAHPS Adult Survey Ratings and Reports by Race and Ethnicity: An Analysis of the National CAHPS Benchmarking Data 1.0. Health Services Research: Submitted.

Murray-Garcia, J. L., Selby, J. V., Schmittdiel, J., Grumbach, K., & Quesenberry, C. P. Jr. (2000). Racial and ethnic differences in a patient survey: patients' values, ratings, and reports regarding physician primary care performance in a large health maintenance organization. Med Care, 38(3), 300-10.

National Committee for Quality Assurance. (Home Page [Web Page]. URL http://www.ncqa.org/Pages/Main/index.htm [2000, May].

Newcomer, R., Preston, S., & Harrington, C. (1996). Health plan satisfaction and risk of disenrollment among social/HMO and fee-for-service recipients. Inquiry, 33(2), 144-54.

Oliver Richard L. (1981). Measurement and evaluation of satisfaction processes in retail settings. Journal of Retailing, 57(3), 25-48.

Parasuraman, A., Berry, L. L., & Zeithaml, V. A. (1991). Understanding customer expectations of service. Sloan Management Review, 39(Spring).

Pascoe, G. C. (1983). Patient satisfaction in primary health care: a literature review and analysis. Evaluation and Program Planning, 6(3-4), 185-210.

Penchansky, R., & Macnee, C. (1994). Initiation of medical malpractice suits: a conceptualization and test. Med Care, 32(8), 813-31. Notes: COMMENTS: Comment in: Med Care 1996 Mar;34(3):280-2

Reise Steven P, Widaman Keith F, & Pugh Robin H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological Bulletin, 114(3), 552-566.

Schlesinger, M., Druss, B., & Thomas, T. (1999). No exit? The effect of health status on dissatisfaction and disenrollment from health plans. Health Serv Res, 34(2), 547-76.

Schnaier, J. A., Sweeny, S. F., Williams, V. S., Kosiak, B., Lubalin, J. S., Hays, R. D., & Harris-Kojetin, L. D. (1999). Special issues addressed in the CAHPS survey of medicare managed care beneficiaries. Consumer Assessment of Health Plans Study. Med Care, 37(3 Suppl), MS69-78.

Sherbourne, C. D., Hays, R. D., & Burton, T. (1995). Population-based surveys of access and consumer satisfaction with health care. Consumer survey information in a reforming health care system: Conference Summary (pp. 37-56). Rockville, MD: Agency for Health Care Policy and Research.

Smith Larissa L, & Reise Steven P. (1998). Gender differences on negative affectivity: an irt study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. Journal of Personality & Social Psychology, 75(5), 1350-1362.

Taira, D. A., Safran, D. G., Seto, T. B., Rogers, W. H., Kosinski, M., Ware, J. E., Lieberman, N., & Tarlov, A. R. (1997). Asian-American patient ratings of physician primary care performance. J Gen Intern Med, 12(4), 237-42.

Thompson, A. G., & Suñol, R. (1995). Expectations as determinants of patient satisfaction: concepts, theory and evidence. International Journal for Quality in Health Care, 7(2), 127-41.

Vaccarino, J. M. (1977). Malpractice. The problem in perspective. JAMA, 238(8), 861-3.

van Campen, C., Sixma, H., Friele, R. D., Kerssens, J. J., & Peters, L. (1995). Quality of care and patient satisfaction: a review of measuring instruments. Medical Care Research and Review, 52(1), 109-33.

Notes: Reviews literature for satisfaction surveys based on 5 criteria: theoretical foundation, containing subscales representing

major aspects of QCPP, reliability and validity, feasibility in population study, and instrument applied in home care setting. Five instruments identified. One theoretically sound.

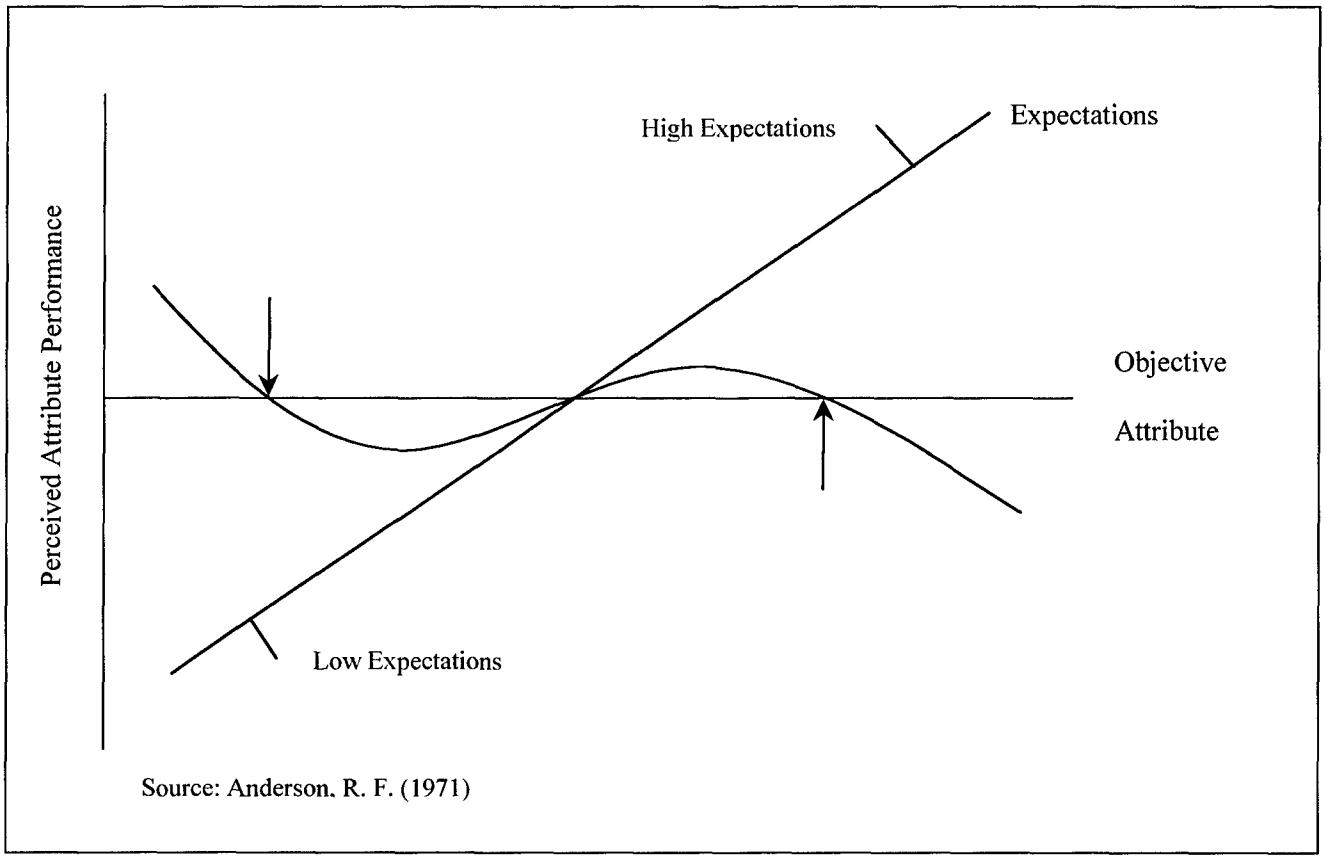
Weidmer, B., Brown, J., & Garcia, L. (1999). Translating the CAHPS 1.0 Survey Instruments into Spanish. Consumer Assessment of Health Plans Study. Med Care, 37(3 Suppl), MS89-96.

Weissman, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A., & Locke, B. Z. (1977). Assessing depressive symptoms in five psychiatric populations: a validation study. Am J Epidemiol, 106(3), 203-14.

Widaman Keith F, & Reise Steven P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. The Science of Prevention: Methodological Advances From Alcohol and Substance Abuse Research (pp. 281'324-xxxii, 458).

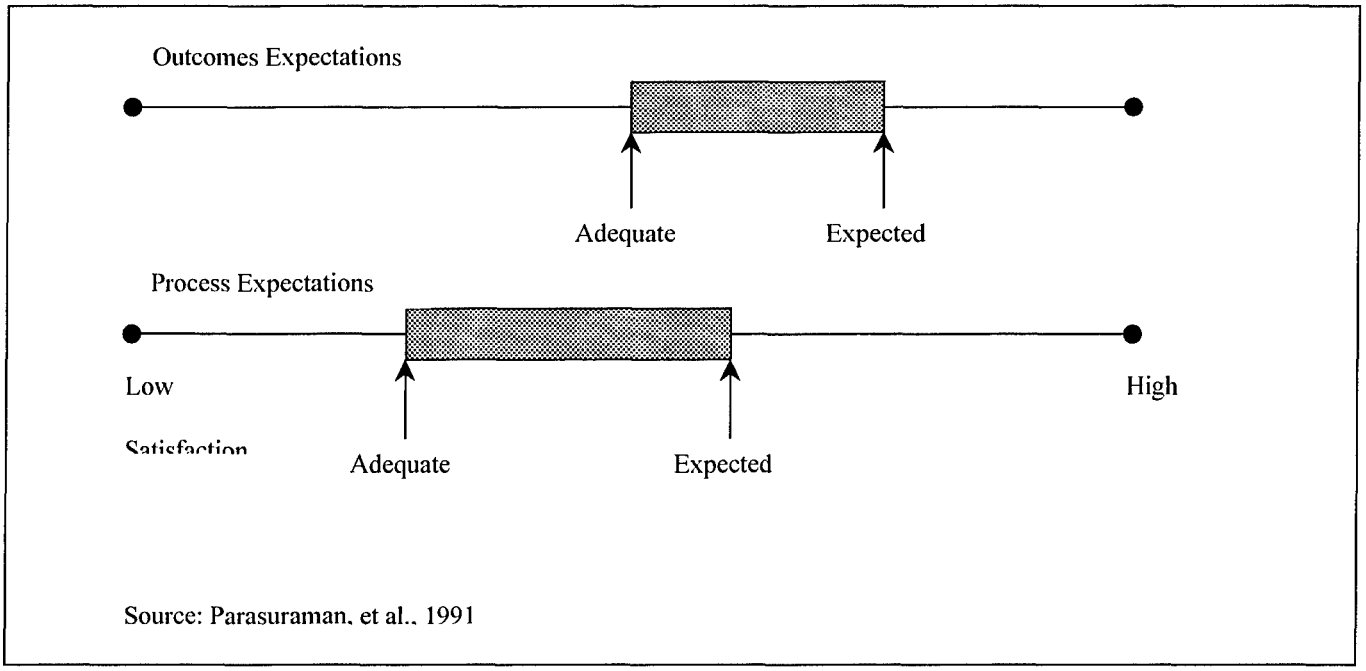
Zegers, R. A. (1968). Expectancy and the effects of confirmation and disconfirmation. J Pers Soc Psychol, 9(1), 67-71.

Figure 1. Assimilation-Contract Model of Perceptions.



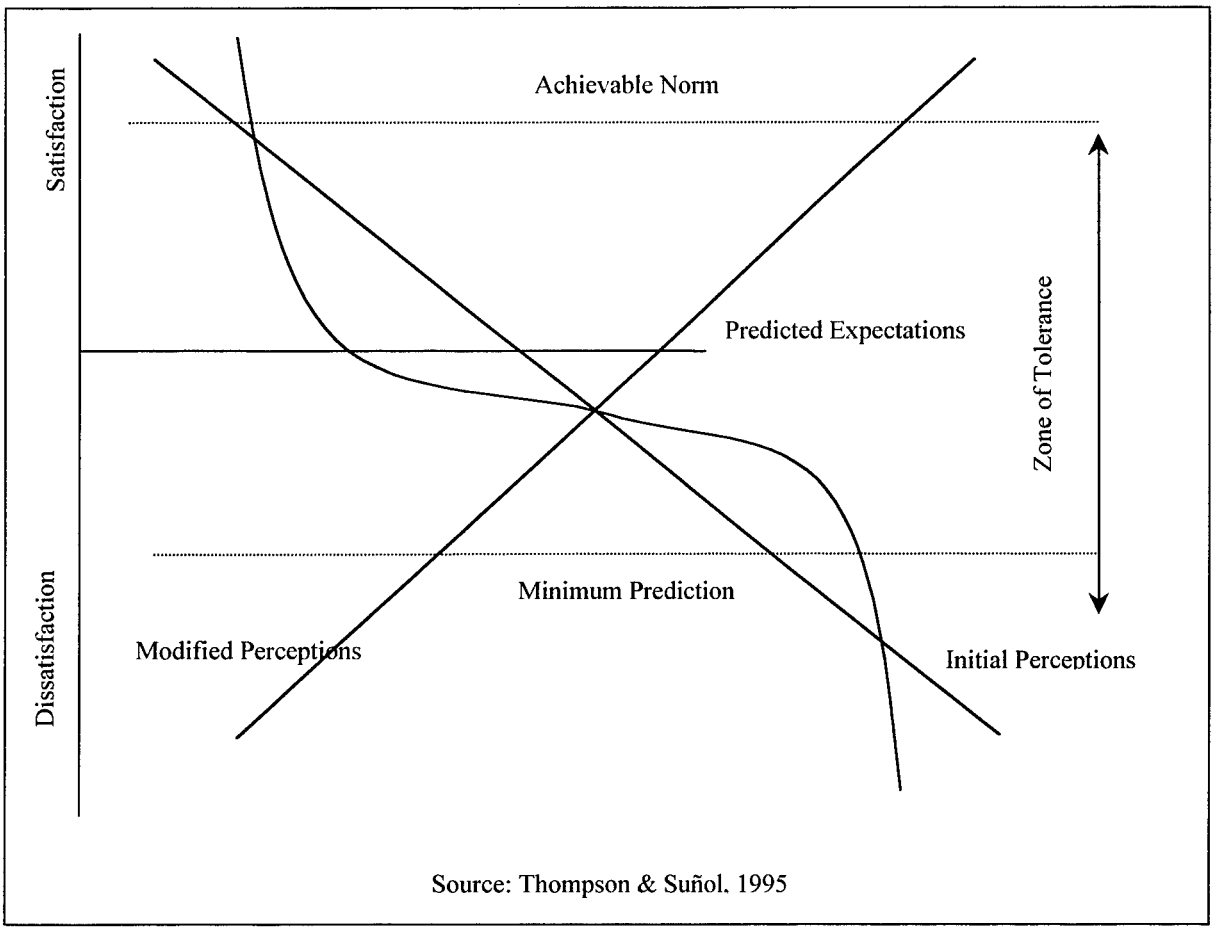
Source: Anderson, R. F. (1971)

Figure 2. Zone of Tolerence Model.



Source: Parasuraman, et al., 1991

Figure 3. Assimilation-Contrast Model of Patient Satisfaction



Source: Thompson & Suñol, 1995

2. CROSS-CULTURAL ADAPTATION OF SURVEY INSTRUMENTS: THE CAHPS® EXPERIENCE

Background

Collecting accurate health data on the growing number of ethnic minorities in the United States has increased in policy relevance in recent years. Today, most general-population sample surveys require translation into at least one language (usually Spanish), and often other languages as well. However, cross-cultural research is threatened by the failure to produce culturally and linguistically appropriate survey instruments for minority populations. Guillemin, Bombardier and Beaton consider that cross-cultural adaptation of instruments is a "prerequisite for the investigation of cross-cultural differences" (1993, p.1425). A survey conducted with an inadequate instrument may lead to erroneous conclusions that are difficult to detect during analyses. Conclusions drawn from such research may be mistakenly attributed to differences between the source and target populations. These risks, and the increasing importance of cross-cultural research, have led to a re-examination of the prevalent techniques for developing survey instruments that will be used in different languages and for assessing the cultural appropriateness of survey instruments that are utilized for this type of research.

In this paper we define culturally appropriate translated survey instruments as conceptually and technically equivalent to the source language, culturally competent, and linguistically appropriate for the target population. This paper provides recommendations for the cross-cultural adaptation of survey instruments and illustrates this with

examples of what is being done in the Consumer Assessment of Health Plans Study (CAHPS®).

The CAHPS® Surveys

CAHPS® is a 5-year initiative that aims to produce a set of standardized survey instruments that can be used to collect reliable information from health plan enrollees about the care they have received. CAHPS® items include both evaluations (ratings) and reports of specific experiences with health plans. CAHPS® surveys are constructed from two pools of items: "core" items that apply across the spectrum of health plan enrollees and supplemental items that are used in conjunction with the core to address issues pertinent to specific populations, such as Medicaid fee-for-service and Medicare managed care. The results of these surveys are then used to prepare reports that provide information to consumers who are trying to select a health plan.

CAHPS® recognizes the need to translate its instruments into several languages in order for its users to adequately collect data on its consumers. The CAHPS® survey instruments were translated into Spanish because it is the second most widely used language in the U.S. (Weidmer, Brown, and Garcia, 1999). As CAHPS® has expanded, several states and users have expressed the need to translate the CAHPS® instruments into other languages as well. The principal goal of the translation process of the CAHPS® surveys and protocols is to produce instruments that are culturally appropriate for the different groups in the selected languages. The main challenge is to produce such

instruments while maintaining equivalency with the English-language version.

Cultural Adaptation of Survey Instruments

Guillemin et al. have described the process of cross-cultural adaptation of surveys as "oriented towards measuring a similar phenomenon in different cultures; it is essentially the production of an equivalent instrument adapted to another culture" (1993, p. 1425). We define culturally appropriate translated survey instruments as conceptually and technically equivalent to the source language, culturally competent, and linguistically appropriate for the target population.

In translating, it is important to distinguish between technical and conceptual equivalence. Technical equivalence refers to equivalence in grammar and syntax, while conceptual equivalence refers to the absence of differences in meaning and content between two versions of an instrument. A technically equivalent instrument is a literal translation using the "equivalent denotative meaning" of the words in the original survey. However, different terms may have a different "connotative" or implied meaning in different cultures, requiring an assessment of conceptual equivalence in the translation of instruments (Marin and Marin, 1991).

Conceptual equivalence includes item and scalar equivalence of the source and translated surveys. Item equivalence signifies that each item has the same meaning for subjects in the target culture. Scalar equivalence is achieved when the construct is measured on the same metric in two cultures (Hui and Triandis, 1985). Health surveys

generally use categorical rating scales where response choices are ordered along a hypothesized response continuum (e.g., excellent to poor). It is important to determine if there is equivalence in the distances between the response choices in the two cultures (Keller et al., 1998).[1]

Cultural competence refers to the requirement that the translated instrument adequately reflect the cultural assumptions, norms, values, and expectations of the target population (Marin and Marin, 1991). Cross-cultural researchers differentiate between universal or common meaning across cultures ("etic") and group-specific ("emic") constructs or ideas. The source survey reflects the assumptions and values of the researcher's culture and in translating surveys, it is generally assumed that the constructs of the source survey are etic. Translated surveys should include both etic and emic items in order to reflect properly the reality being studied. This implies the development of new items that reflect the emic aspects of a concept in the target culture (Brislin, 1986).

Linguistic appropriateness refers to the language readability and comprehension of the translated instrument. The goal is to develop instruments using wording at a level easily understood by the majority of potential respondents. An instrument developed in the source language at an eighth grade reading level does not automatically preserve the same reading and comprehension level upon translation, and may actually increase considerably. The problem of equivalence in

[1]For a discussion of the Thurstone scaling exercise applied to the SF-36 see Keller et al. (1998).

reading level is further compounded if the target population is at a lower average reading level than the source language population.

In order to cross-culturally adapt survey instruments, we propose a framework (Figure 1) that comprises the following activities:

- Translation (steps 1 to 4)
- Qualitative analysis (step 5)
- Field test and analyses (step 6)

Based on the results of the field test, additional qualitative analysis may be necessary. The International Quality of Life Assessment (IQOLA) project group has used a similar protocol in translating the SF-36 Health Survey into different languages (Bullinger et al., 1998; Gandek and Ware, 1998).

Translation (Steps 1 to 4)

Most researchers today agree that it is no longer acceptable to use a direct-translation technique (or one-way-translation) for translating survey instruments. A review of the literature indicates that the most accepted approach to translation is one in which a variety of techniques are used to ensure the reliability and validity of the translated survey instrument (Brislin, 1986; Bullinger et al., 1998; Marin and Marin, 1991). The rationale behind this approach is that no single technique adequately demonstrates and improves the equivalence of an instrument, and that only a multi strategy approach that provides and evaluates different types of equivalence can produce an adequate translation. We recommend a process for translating surveys

that includes translation, back-translation, independent review, and review by committee.

Forward-translation

Professional translators (two or more) experienced in translating similar survey instruments, preferably native speakers of the target language, are retained to translate the survey instrument. The translators used for this task should have familiarity with the target population and with data collection procedures. Before starting the translation, the translators should be briefed on the objectives of the study, the demographic characteristics of the sample, the interviewing mode to be used, and the targeted reading level of the translation.

Back-translation

Once the instruments are translated they go through a process of back translation. In this process the translated instrument is given to two translators, native English speakers, who are instructed to translate the questionnaire back into English. It is important that this translator not have access to the original English language versions of the instrument and that he/she does not consult with the first translators.

Independent Review and Comparison

The third step in the translation process it to give the translated versions of the survey instruments to one or more bilingual reviewers. The reviewers are provided with the original English

versions and the back-translated versions and are instructed to compare the two, highlighting any discrepancies in meaning or equivalence.

Review by Committee

Once the review process is completed, the forward-translators, the back-translators and the reviewer(s) hold a series of meetings to discuss problems found during the review process, to correct errors in grammar and syntax and to resolve problems of equivalence found among the versions. Decisions on wording and corrections are made by consensus. The rationale behind this is that a translator or back-translator can introduce his or her own bias or error into a translation. The review-by-committee approach is useful in neutralizing the cultural, social, and ethnic bias that can be introduced when using only one translator and one back-translator.

CAHPS® Translation

Rather than produce multiple, population-specific Spanish translations, CAHPS® sought to produce an instrument that would be understood by most respondents by using "broadcast Spanish," and that maintained a reading and comprehension level that would be accessible to most respondents. "Broadcast Spanish" refers to a type of Spanish that is understood by most Spanish speakers regardless of their country of origin or ethnic background (Marin and Marin, 1991).

A professional translator experienced in translating survey instruments similar to the CAHPS® instrument was retained. The translated instrument was then given to a bilingual reviewer experienced in designing and translating survey instruments for cross-

cultural research. The reviewer focused on identifying syntax and typographic errors, identifying questions or terms that sounded awkward and identifying terms that were conceptually problematic. Once this process was complete, the reviewer was provided with the English version and was asked to compare the two instruments, highlighting any discrepancies in meaning or equivalence.

In an effort to adhere as closely as possible to the English version, the translator produced an initial Spanish version of the survey instruments that was technically equivalent to the English version, but in many instances was not conceptually equivalent, and in some cases, not linguistically appropriate for the target population (by using terms that are seldom used in Spanish, anglicisms, or words that are too sophisticated for the target population). The translator had been instructed to aim for a translation that would be appropriate for a Spanish-speaking Medicaid population likely to have less than 6 years of formal education. However, this proved to be difficult to accomplish while maintaining equivalence to the English version.

A member of the RAND CAHPS® team met with the translator and the reviewer to go over discrepancies related to equivalence. The reviewer and the translator back-translated problem areas in the Spanish version to further distinguish the source of the problems before decisions were made about addressing them. A final review of the original English version, the translation, and the back-translation, was conducted by the committee—the translator, the reviewer, and CAHPS® team member—and alternative wording for problematic terms was implemented. Table 1 shows terms that were problematic because they were not conceptually equivalent, were too sophisticated for the target population, or were

too-infrequently used by most Spanish speakers. The alternative wording in the final version comes closer to the conceptual meaning in the English version and is easier for the respondents to understand.

Qualitative Analysis (Step 5)

Qualitative research consists of "research methods employed to find out what people do, know, think, and feel by observing, interviewing, and analyzing documents" (Shi, 1997, p. 398). These methods should be viewed as complementary to quantitative methods. Qualitative methods are particularly useful in assessing the cultural competence or content validity of the translated survey instrument[2]. It is important to evaluate whether the survey measures the group-specific domains of the phenomenon under study for the target population. Qualitative methods assist in identifying the "etic" (universal) and "emic" (culture-specific) constructs or behaviors of a group. This constitutes an evaluation of the "subjective" culture whereby consistencies or patterns in responses by members of a group are used to identify the group's cognitive structure (Marin and Marin, 1991). The assumption is that the group's norms, values, and expectancies influence the observed consistencies or similarities in responses of a given cultural group. Qualitative methods can also be used to assess the conceptual equivalence and linguistic appropriateness of the translated survey.

---

[2] Herdman, Fox-Rushby and Badia (1997) recommend that qualitative methods of instrument evaluation precede the translation of survey instrument.

We are using qualitative methods to investigate the appropriateness of the CAHPS® survey content for Spanish-speaking Latino patients enrolled in Medicaid. First, we want to determine whether the items and scales currently contained within CAHPS® address the key concerns and expectations of Latino patients with respect to their health care providers and health plans. Second, we want to verify that the translated survey items, initially developed in English, have similar meaning in Spanish. Finally, we want to determine the readability level of the Spanish language survey instruments and determine whether it is appropriate for the Spanish-speaking Medicaid population.

There are three types of qualitative research pertinent to cross-cultural research: focus groups, cognitive interviews, and readability assessments. In this section we discuss the use of focus groups and cognitive interviews. For a discussion on readability assessments and its application to the CAHPS® surveys see Morales et al. (1999) in this conference proceedings.

Focus Groups

Focus groups are a research tool that relies on group discussions to collect data on a given topic (Morgan, 1996). Participant interactions help to reveal experiences, values, beliefs, and feelings. In addition, group discussion helps uncover extent of consensus or diversity, and its sources. Focus groups have been used extensively in marketing research to obtain customer input on new products (Burns and Bush, 1995); however, their use in cross-cultural research has been more limited. The primary objective of the focus groups in cross-

cultural research is to assess whether the domains currently covered in the survey adequately address the needs and expectations of the target population, and to assess the need for developing new domains or expanding current domains. The focus group process usually starts with a literature review and analysis of health surveys that focus on the target population, to aid in the identification of issues and concepts particular to the cultural group.

Stewart and Shamdasani (1990) have identified eight steps in the design and conduct of focus groups:

- Formulation of the research question
- Identification of sampling frame
- Identification of moderator
- Generation and pre-testing of structured protocol
- Recruiting the sample
- Conducting the focus group
- Analysis and interpretation of data
- Writing the report

A group size of 8 to 12 respondents per focus group is recommended (Burns and Bush, 1995). Homogenous groups based on demographics or other relevant characteristics are also recommended. This is important to elicit conversation among participants. Focus groups in cross-cultural research generally involve culturally

homogenous groups. However, the researcher may consider additional relevant demographic characteristics in forming the groups. For example, elderly Latinos versus teenager Latinos.

The moderator is the most crucial factor to ensure the effectiveness of the focus group. The focus group moderator conducts the entire session and guides the flow of group discussion across specific topics. According to Burns and Bush, the moderator "must strive for a very delicate balance between stimulating, natural discussion among all of the group members while at the same time ensuring that the focus of the discussion does not stray too far from the topic" (1995, p. 200).

In analyzing the data, the qualitative statements of the participants are translated into categories or themes and an indication is given of the degree of consensus apparent in the focus groups. The results of the focus groups inform the development of new items for the survey and the modification of existing measures as needed.

CAHPS® Focus Group

A focus group was conducted on November 7, 1998 at one of the clinics of a local health plan. The participants were recruited from among the Latino patient population of the health plan's clinics in two Los Angeles County communities with high concentrations of Latinos. In order to be considered for participation in the focus group, patients had to be adults (18 and over) and primarily Spanish speaking.

A member of the RAND CAHPS® team moderated the focus group using a scripted discussion guide. The focus group was conducted entirely in

Spanish and lasted for approximately two hours. Twelve women, ranging in age from 24 to 73 years, attended the focus group. Eleven of the participants were from Mexico and one was from Nicaragua. All of the women had been in the U.S. for many years, ranging from 10 to 23 years.

The specific objectives of the focus group included:

- Determining Latino patients perceptions about health providers;
- Collecting information on communication issues between Latino patients and their providers;
- Gathering information on the use of interpreters by Latino patients;
- Seeking information on the role of the family in health seeking behavior and in making decisions about healthcare;
- Collecting information on Latino patients' satisfaction with their health care, and
- Determining the most important aspects related to health care for Latino respondents.

Briefly, the results of this focus group raised interesting points:

- Provider's communication is highly valued by Latinos: that a doctor spend enough time with them, that he/she ask them questions, and that he/she provide sufficient information about the patient's illness and medications.

Participants were less concerned with the doctor's Spanish speaking ability (although they do value it) and with the doctor's race or gender.

- Participants reported some dissatisfaction with the care that they received from their health plan. Their chief complaints related to issues of promptness of care. Specifically, patients complained of difficulty obtaining timely appointments and long delays in seeing the doctor once arrived at the clinic.

- Most of the participants reported problems in using interpreters. They complained about the quality of the translation. In addition, patients reported not discussing certain personal health problems because of being ashamed to speak in front of their interpreter.

- Some participants reported going to Mexico to receive health care and the rest reported that they too would seek health care in Mexico if they could afford it financially. Among the reasons for preferring the care received in Mexico were the promptness of care, continuity of care, and provider's communication and approach to care.

The findings from the focus group suggested that the substantive issues covered in version 2.0 of the CAHPS® Survey Instrument are culturally and substantively appropriate. Two of the findings from the focus group are not addressed as part of the survey and require further exploration. The first of these findings centers on the use and quality

of interpreters and how this affects provider-patient communication. Although the CAHPS® supplemental item set contains items that ask about the need and availability of interpreters, it does not cover the issue of interpreter quality and the effect of interpreters on communication between a provider and his/her patient. The second of these findings relates to patients who travel to Mexico to seek health care in spite of the fact that they can receive health care from their health plan. This information is being used to field test additional CAHPS® survey items addressing care in Mexico.

Cognitive Interviews

Cognitive-testing techniques are often used in the process of questionnaire development to investigate, assess, and refine a survey instrument (Berkanovic, 1980). Cognitive testing can detect and minimize some sources of measurement error by identifying question items or terms that are difficult to comprehend, questions that are misinterpreted by the respondents, and response options that are inappropriate for the question or that fail to capture a respondent's experience (Jobe and Mingay, 1991).

One of the most common forms of cognitive testing is the cognitive interview to examine the thought processes of the interviewee. There are two forms of cognitive interviews: the concurrent and retrospective approaches. With the concurrent technique, the respondent goes through a process of "thinking-aloud" or articulating the thought processes as he or she answers a survey item. In the retrospective or "debriefing" technique, the interviewer asks questions about the survey process after the respondent completes the

survey (Harris-Kojetin, Fowler, Brown, Schanaier, and Sweeney, 1999). Verbal probes or follow up questions may be used in either type of cognitive interview. One common probe is to ask the respondent to paraphrase the survey question. This helps to understand whether the respondent understands the question and gives it the intended interpretation. This may also suggest more appropriate wording for the survey item.

Prior to conducting the cognitive interviews, a structured protocol is developed to ensure that all participants receive similar prompts from the facilitators. The structured protocol is translated. Interviewers are bilingual in the target language and are trained in using cognitive interview techniques. Using notes taken during the cognitive interviews and audiotapes of each of the interviews, each interviewer writes up a summary for each interview in English. These summaries are then combined into one report outlining the results of the cognitive testing.

CAHPS® Cognitive Testing

The CAHPS team completed 150 cognitive interviews in different geographic locations (Harris-Kojetin et al., 1999). Seven cognitive interviews were completed in Spanish in California during June-July, 1996. A concurrent think-aloud technique with scripted probes was used in this case. The Spanish-language interviews were completed with adult women on Medicaid who were receiving AFDC benefits and were enrolled in either an HMO or a fee for service plan through Medicaid.

The primary objectives of the cognitive interviews were:

- To assess whether respondents understood the CAHPS® survey instruments.
- To determine the optimal response categories for ratings and reports of care.
- To identify the source of problems in comprehension: translation, reading level, survey content and cognitive task involved.

The results of each cognitive interview were summarized in reports and analyzed for points of convergence. In addition, the interviewers were debriefed and asked to provide general feedback on how well the instruments were working and to discuss content areas or issues that were problematic.

For the overall ratings, an adjectival scale (excellent, very good, good, fair, poor) was compared with a numeric scale (0-10). There was less translation difficulty with the numerical than the adjectival categories. It was particularly difficult to translate "fair" and "poor" into Spanish (Harris-Kojetin et al., 1999).

The cognitive tests were also used to explore whether key words and concepts worked equally well in Spanish and English. Specific wording and terms that were particularly problematic for Spanish-speaking respondents were modified based on the results of the cognitive testing and used to produce instruments that were ready for pretesting.

The interviewers reported that the survey instruments worked better with the respondents who seemed to be more educated or acculturated. Another issue identified by interviewers as problematic was that the instrument presumed that all prospective respondents were reasonably familiar with the terminology and landscape of the health care system in the United States. Familiarity with the system may be common for most Medicare and Medicaid recipients, but it also is related to length of time in the United States and to levels of acculturation, usually lower for non-English-speaking respondents.

Field Test and Analyses (Step 6)

A field test of the translated survey instrument is also recommended. Psychometric analysis can then be used to assess the reliability and validity of the translated survey instruments. Psychometric testing can also be used to test for measurement equivalence across cultural groups. Three types of analysis commonly used are:

Reliability estimates, such as Cronbach's (1951) alpha coefficients, to measure the internal consistency of the instrument. Cronbach's alpha is based on the number of items in the scale and the homogeneity of the items. The homogeneity of the items represents an average of the inter-item correlations in a scale and measures to what extent items share common variance.

Factor analysis to examine the internal structure of the instrument or construct validity of the scales. In addition, factor analysis can be used to test measurement invariance across groups (Reise, Widaman, and Pugh, 1993).

Item Response Theory (IRT) methods provide an ideal framework for assessing differential item functioning (DIF), defined as different probabilities of endorsing an item by respondents from two groups who are equal on a latent trait. When DIF is present, trait estimates may be too high or too low for those in one group relative to another (Thissen, Steinberg, and Wainer, 1993).

CAHPS® Field Test

A pretest of preliminary drafts of the CAHPS® 1.0 survey instruments was conducted as part of the Medicaid field-test data collection conducted by RAND in 1996 (Brown, Nederand, Hays, Short, and Farley, 1999). Only 23 respondents completed the interview in Spanish. All 23 Spanish-speaking respondents completed the interview by telephone. The total number of completes in Spanish was insufficient to conduct sensitivity analyses to determine whether the Spanish-language instruments were performing like the English-language instruments.

Conclusion

Adept translation of a survey instrument is an integral part of the instrument-development process, but it alone does not ensure that a culturally appropriate survey instrument will result. Cross-cultural adaptation of survey instruments requires that the translated instruments be conceptually and technically equivalent to the source language, culturally competent, and linguistically appropriate for the target population. Producing a survey instrument that is culturally appropriate for Latinos in the United States requires subjecting the

Spanish-language instruments to rigorous testing. That testing must include conducting focus groups and cognitive interviews that evaluate the cultural appropriateness of the survey content as well as the cognitive task required in the survey instrument, determining the reading level of survey instruments in Spanish, and field testing the survey instrument to ensure that the survey measures perform equally well in Spanish and English.

The results of the cognitive interviews and the focus groups may require modifying the English version of the survey instruments by adding domains to capture the experiences of Latino consumers, modifying the construction of items in English to make them more "translatable" into Spanish, modifying the Spanish version to accommodate ethnic and regional variations in Spanish language use, and simplifying the translation to make the reading level of the document appropriate for the target population.

In order to assess the cultural appropriateness of the CAHPS® 2.0 survey instruments among different Latino ethnic groups and to account for regional variations in care, focus groups and cognitive interviews will be conducted in San Diego, New York, and Miami. By conducting focus groups across these sites, we will incorporate Latinos of Mexican, Puerto Rican and Cuban origins in our focus groups. The qualitative component of CAHPS® is being done later than we would like. Ideally this phase would have taken place before finalizing the English-language instrument. Currently, we are also conducting a field study of the CAHPS® surveys among a Medicaid managed care population in the San Diego area. Our goal is to obtain 50% of completed surveys in Spanish.

References

Berkanovic, E. (1980). The effect of inadequate language translation on Hispanics' responses to health surveys. American Journal of Public Health, 70, 1273-1276.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), Field Methods in Cross-Cultural Research. Beverly Hills, CA: Sage Publications.

Brown, J. A., Nederand, M. A., Hays, R. D., Short, P. F., and Farley, D. O. (1999). Special issues in assessing care of Medicaid recipients. Medical Care, 37, MS79-MS88.

Bullinger, M., Alonso, J., Apalone, G., Leplege, A., Sullivan, M., Wood-Dauphinee, S., Gandek, B., Wagner, A., Aaronson, N., Bush, P., Fukuhara, S., Kaasa, S., & Ware J. E. (1998). Translating health status questionnaires and evaluating their quality: The IQOLA project approach. Journal of Clinical Epidemiology, 51, 913-923.

Burns, A. C. & Bush, R. F. (1995). Marketing Research. Englewood Cliffs, NJ: Prentice Hall.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Gandek, B. & Ware, J. E. (1998). Methods for validating and norming translations of health status questionnaires: The IQOLA project approach. Journal of Clinical Epidemiology, 51, 953-959.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. Journal of Clinical Epidemiology, 46, 1417-1432.

Harris-Kojetin, L. D., Fowler, F. J., Brown, J. A., Schnaier, J. A., & Sweeny, S. F. (1999). The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. Medical Care, 37, MS10-MS21.

Herdman, M., Fox-Rushby, J., & Badia, X. (1997). 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Quality of Life Research, 6, 237-247.

Hui, C. H. & Triandis, H. C. (1985). Measurement in cross-cultural psychology. Journal of Cross-Cultural Psychology, 16, 131-152.

Jobe, J. & Mingay, D. (1991). Cognition and survey measurement: History and overview. Applied Cognitive Psychology, 5, 175-192.

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

Marin, G. & Marin, B. V. (1991). Research with Hispanic Populations. Newbury Park, CA: Sage Publications.

Morales, L. S., Weidmer, B., & Hays R. D. (1999). Readability of CAHPS 2.0 child and adult surveys. Proceedings of the 7th Conference on Health Survey Research Methods.

Morgan, D. L. (1996). Focus groups. Annual Review of Sociology, 22, 129-152.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114, 552-566.

Shi, L. (1997). Health Services Research Methods. Albany, NY: Delmar Publishers Inc.

Stewart, D. W. & Shamdasani, P. M. (1990). Focus Groups. Newbury Park, CA: Sage Publications.

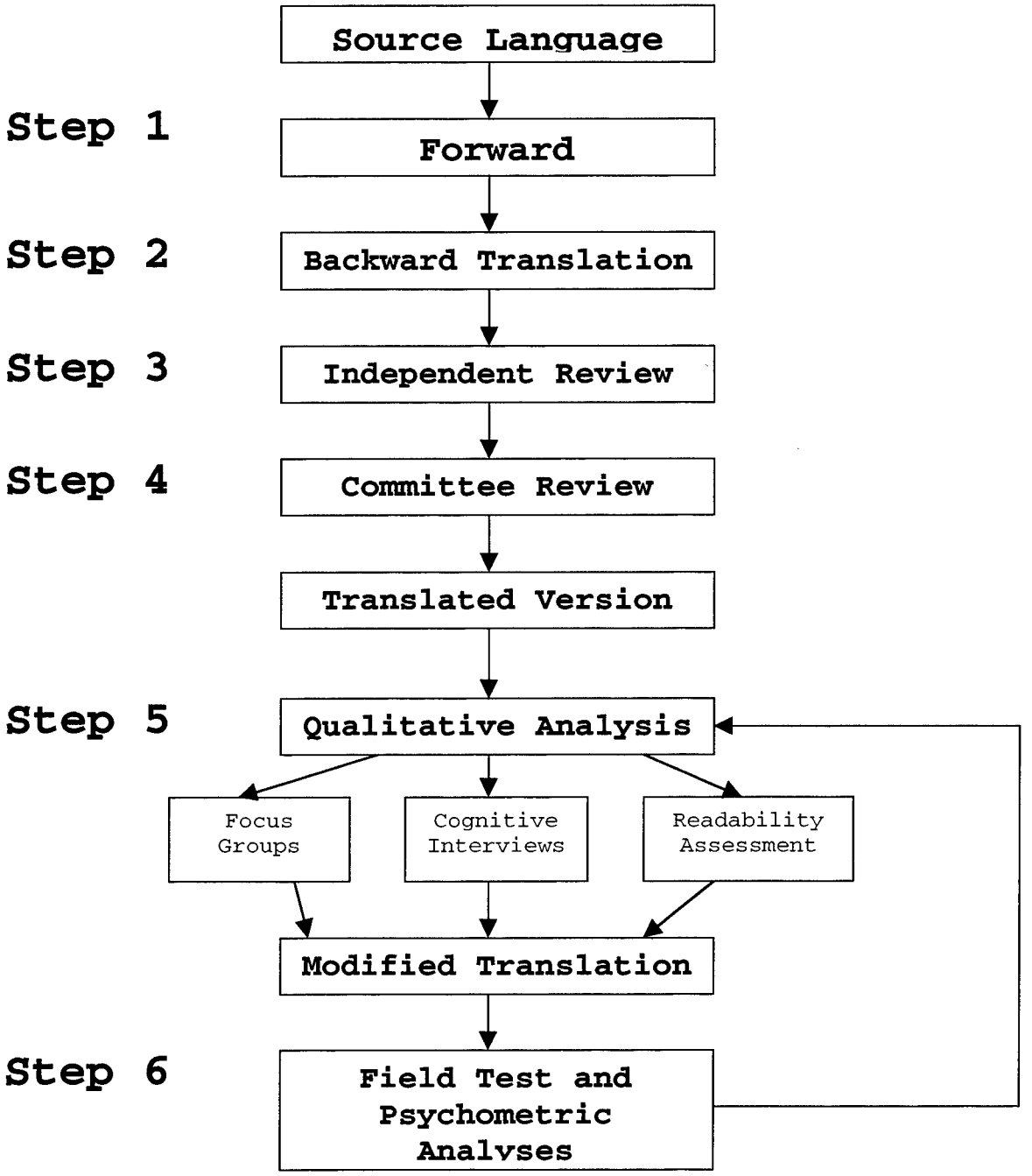
Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holand & H. Wainer (Eds.), Differential Item Functioning. Hillsdale, NJ: Erlbaum.

Weidmer, B., Brown, J., & Garcia, L. (1999). Translating the CAHPS 1.0 survey instruments into Spanish. Medical Care, 37, MS89-MS96.

Table 1. Terms That Presented Difficulty in Translation

Original English	Alternative wording used in the final Spanish Version	Back-translation
Health insurance plan	plan de seguro médico	medical insurance plan
Health provider	profesional de salud;	health professional
Rating/rate	calificación/califica	grade/grade
Usually	normalmente	normally
Preventive health steps	medidas de salud preventiva	preventive health measures
Listen carefully	escucharon atentamente	listen attentively
Health care	atención médica	medical attention
Prescription medicine	medicamentos recetados	prescribed medications
Male or female	niño o niña/hombre o mujer	boy or girl/man or woman
Background	ascendencia	ascendancy
Grade	año	year
School	estudios	studies
Highest	avanzado	advanced

Figure 1: Cultural Adaptation of Survey Instruments



3. READABILITY OF CAHPS® 2.0 CHILD AND ADULT CORE SURVEYS

Background

In recent years, the emergence of managed care has prompted interest in collecting survey information from health care consumers. Many public and private purchasers of care either already administer patient surveys to their beneficiaries or plan to in the near future. However, the growing diversity of the U.S. population poses major challenges for developing such survey instruments. First, the cultural and linguistic diversity of many beneficiary groups requires that surveys be appropriately translated into various languages and adapted for different groups. Second, because patient surveys are often self-administered, attention must be given to survey readability.

Research studies from many sources, including national literacy data, tell us that a large share of U.S. adults can only read at very basic levels. This problem is particularly striking among Medicaid beneficiaries. According to the 1993 National Adult literacy survey (Kirsch, Jungeblut, Jenkins & Kolstad, 1993), 75% of welfare recipients read at or below the eighth grade level and 50% read at or below the fifth grade level.

Moreover, low reading skills may be more concentrated among certain Medicaid beneficiary sub-groups more than others. For instance, immigrants and refugees from less-developed countries may be more likely than U.S.-born Medicaid beneficiaries to have low educational attainment and, as a result, low reading skills. Among recent Central American immigrants and refugees entering the U.S. from El Salvador and Guatemala, nearly 80% reported less than a high school

education (Lopez, 1996). Among foreign-born Hispanics living in the Los Angeles region, 10% report no schooling, 38% reported elementary school only, and 21% reported some high school (Cheng & Yang, 1996).

The mismatch between an intended respondent's reading ability and the survey instrument may have important implications for the validity of patient satisfaction research, particularly for self-administered surveys. Some of the consequences of this mismatch may include low response rates, especially in vulnerable populations, and unreliable responses because of poor item comprehension.

This study assesses the readability of the English and Spanish versions of the Consumer Assessments of Health Plans Study (CAHPS®) 2.0 adult and child core surveys. The linguistic and cultural adaptation of these surveys is discussed in a separate paper (Weech-Maldonado, Weidmer, Morales & Hays, 1999).

The CAHPS® Surveys

CAHPS® is a 5-year initiative that aims to produce a set of standardized survey instruments that can be used to collect reliable information from health plan enrollees about the care they have received and their experiences with their health plan. The results of the surveys are turned into reports that provide decision support to consumers selecting a health plan.

To date, several instruments have been developed as part of this study, each targeting a specific population served by health plans throughout the U.S. CAHPS® has also developed surveys for children, designed for a proxy respondent. Although variations exist between the different versions of these instruments depending on the target

population and the age of the respondent, a core set of survey questions is common to all versions of the survey. Five specific domains of care (getting needed care, getting care quickly, communication with providers, office staff courtesy and respect, health plan customer service) and global ratings (care overall, personal doctor or nurse, specialist care, health plan) are assessed in the CAHPS® 2.0 surveys.

The CAHPS® investigators recognized the need to translate its instruments into other languages. Indeed, the CAHPS® survey instruments were translated into Spanish (Weidmer, Brown & Garcia, 1999) because many participating health plans are located in states that have large numbers of Spanish speakers, including Texas, California, New Jersey and Florida. Hence, we evaluate the Spanish versions of the adult and child core surveys along with the English surveys.

Assessing Readability

Two major approaches are available for assessing the readability of documents - measurement and prediction. Measuring readability, by judgment or comprehension tests, involves using readers. Readability by judgment is usually obtained by asking literacy experts to determine the readability level of a document based on their experience or on use of an algorithm. Readability by comprehension test is obtained by administering a reading comprehension test based on the written material to readers of known ability. A test score criterion is chosen that defines comprehension of the material. When some proportion of readers of similar ability achieve that score, the reading ability of the test takers corresponds to the readability level of the document.

In the second approach, mathematical formulas predict the readability of a document. Unlike judgments or comprehension tests, readability formulas do not rely on readers to establish the readability level of written materials. Because no measurements are made, readability formulas are strictly prediction tools.

The selection of readability technique depends upon the factors of time, availability of subjects, level of resources available to conduct the assessment, and the degree of accuracy required in assessing the materials for the target groups (Klare, 1974). Predicting readability by formulas does not involve readers and is therefore much less expensive, but it only provides an approximate indication of the readability of a document. Measurements obtained by tests and judgments by experts require greater resources but provide more accurate assessment of readability. We chose to use the former approach for this study for two reasons. First, prior research had addressed the readability of the CAHPS® surveys through cognitive interviews (Harris-Kojetin, Fowler, Brown, Schnaier & Sweeny, 1999) and expert judgments (Brown, Nederend, Hays, Short & Farley, 1999). Second, available resources constrained us to using readability formulas.

To identify appropriate readability formulas for our study, we conducted a literature search. Our goal was to identify formulas appropriate for survey instruments in Spanish and English. Although we found references to numerous readability formulas, we did not identify any formulas appropriate for evaluating survey instruments in English or Spanish. The principal problem with applying readability formulas to survey instruments is that the formulas become unreliable when applied to passages of fewer than 100 words (Fry, 1990). Because the

CAHPS® surveys are composed of multiple closed-ended questions followed by a set of response options, passages of less than 100 words are common. Furthermore, the vast majority of formulas we identified were appropriate for English written materials but not Spanish.

Most readability formulas typically use two factors in their calculations: a sentence or syntactic factor and a word or semantic factor (Rush, 1985). Formulas using these two factors include the Fry Readability Graph (Fry, 1965), Dale-Chall (Dale & Chall, 1948), Fog (Gunning, 1968), Flesch (Flesch, 1948), and Flesch-Kincaid (Kincaid, Fishburne, Rodgers & Chissom, 1975). The SMOG (McLaughlin, 1969) is an exception because it has only a syntactic factor. The syntactic factor frequently estimates the grammatical complexity of the writing by using sentence length. The semantic factor purports to measure the degree of difficulty of the vocabulary in a piece of writing. Readability formulas usually estimate semantic load either with a measure of word length such as number of syllables or with a count of unusual words. Thus the assumption that word and sentence length are reasonable correlates of syntactic complexity and semantic load underlies reading formulas (Rush, 1985).

Readability formulas are typically validated against performance criterion passages of varying but known levels of difficulty. Two common sources of criterion passages are the McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1961) and the Gates-MacGinite reading tests (Gates & MacGinite, 1965). The validity of a particular readability formula is determined by how accurately it predicts the grade level of a criterion passage. In addition, the validity of more recent formulas is established in part through correlation with older formulas.

In addition to using a readability formula, some investigators have chosen to describe readability using a variety of counts of syntactic and semantic factors (Leadbetter, 1990). Fry recommends the use of word counts and sentence length to assess the readability of passages having fewer than 100 words (Fry, 1990). Because readability formulas were not originally intended for survey instruments, we have supplemented the readability formula results with counts of a variety of syntactic and semantic factors (see Table 1, page 73).

Adapting Survey Instruments for Readability Assessments with Formulas

Using readability formulas to assess the CAHPS® surveys required us to exclude the question response scales, leaving only the instructions, question preambles and the survey questions themselves. The question response scales were deleted from the text of the surveys because they do not have a sentence structure, which readability formulas assume.[1]

Fry Readability Graph

The Fry Readability Graph (FRG) is the principal readability assessment tool used in this study because it has been validated for Spanish and English language documents. Like most readability formulas, the FRG has syntactic and semantic factors - sentence length and syllables. To implement the FRG, one first randomly selects three sample passages of exactly 100 words - from the beginning, middle, and

[1] Other researchers have turned response options into sentences and included them in their readability analysis (Lewis, Merz, Hays, and Nicholas, 1995).

end of the source document. (Our source documents consisted of the CAHPS® surveys, stripped of all response options.) After the total number of sentences and syllables for each of the 100-word passages has been recorded, the average number of sentences and syllables is computed. The resulting figures are plotted on a graph and the resulting coordinate point is associated with an established grade level designation. An illustration of the FRG is shown in Figure 1. The FRG is appropriate for assessing materials from the first grade through the college level (Fry, 1969, 1977).

The FRG is one of the few readability assessment tools that is adapted for Spanish language documents (Gilliam, Peña & Moutain, 1980). Spanish language application of the FRG is similar to its English language application, with the exception of syllable counting. Because of differences in the structure of words in the two languages, the syllable counts for 100-word passages in Spanish tend to be much higher than for the same passage in English. To correct for this discrepancy, 67 is subtracted from the total syllable count for each 100-word passage in Spanish (Gilliam et al., 1980).

The comparability of the FRG applied to Spanish language documents (with the adaptation) and English language documents has been assessed. Using Spanish primary textbooks, the readability level of the FRG and the publisher's grade level were compared. In 10 of 12 cases, the FRG level grade level and publisher's grade level were the same (Gilliam et al., 1980). Unfortunately, a similar comparability study has not been conducted using the FRG for documents at higher reading levels.

FRASE Graph

The FRASE graph is a readability assessment tool specifically developed for Spanish written materials (Vari-Cartier, 1981). The FRASE graph addresses two limitations of the FRG by increasing the syllable count range beyond 182 per 100-words and altering the readability designations from grade levels to the reading difficulty designations used in English as a Second Language instruction (Beginning, Intermediate, Advanced Intermediate, and Advanced).

The FRASE graph is derived from the FRG, also basing its readability assessment on a syllable and sentence count. However, the FRASE graph uses five 100-word samples rather than three.

The FRASE graph has been extensively validated using subjective teacher judgments, Spaulding formula scores, cloze test scores, and informal multiple-choice test scores. Correlation coefficients between the FRASE graph readability designations and the alternative readability estimates ranged from 0.91 to 0.97, indicating that the FRASE graph is equivalent to other established methods for estimating readability (Vari-Cartier, 1981).

Fog Index

The Fog Index, which uses as few as 100 successive words to determine both sentence length and the number of words with three or more syllables, was developed by Gunning (1968). The counts are then substituted into a formula[2] and the reading difficulty is calculated according to formal grade level in school. For longer written works, the author recommends selecting several 100 word samples from various parts of the material averaging the results to determine the reading

[2] FOG Readability formula: Grade Level = 0.4*(average sentence length + percentage of words with 3 or more syllables).

level. This formula is appropriate for assessing materials from the fourth grade through the college level.

The Fog Index has not been adapted for Spanish language materials.

SMOG Grading Formula

The SMOG grading formula is based solely on syllables. It was developed by G. Harry McLaughlin as a fast and accurate test of readability (McLaughlin, 1969). The SMOG Grading Formula estimates the grade level of a document by counting the number of polysyllabic words (words with 3 or more syllables) in three chains of 10 consecutive sentences taken from the beginning, middle, and end of the document being assessed.

An advantage of the SMOG is that the standard error of the readability prediction has been estimated (SE=1.5 grades) based on validation studies using the McCall-Crabbs passages. A standard error of 1.5 grade levels means that the material being tested will be fully comprehended[3], by 68% of its readers who have reached a reading skill level within 1.5 grades of the SMOG score.

The SMOG grading formula has been adopted by the National Cancer Institute as the preferred method for assessing the readability of cancer communications after a comprehensive review of advantages and disadvantages, including how well alternative formulas predict readability (Romano, 1979).

---

[3] The reading ability, indicated by the grade placement score, needed to answer 100% of test questions on the Mcall-Crabbs passage for that grade level (Klare, 1974).

The SMOG grading formula has not been adapted for Spanish language materials.

Flesch Reading Ease Score

The Flesch Reading Ease Score is one of the most widely used readability assessment formulas. Rudolf Flesch published his first reading formula in 1945, based on the number of affixes, the average sentence length, and the number of personal references. He subsequently introduced the Reading Ease formula, which is based on number of syllables per 100 words and average number of words per sentence. When applied to a document, the Flesch Reading Ease formula results in a number ranging from 0 to 100. The lower the score, the more difficult the material is to read and comprehend. The Flesch Reading Ease Score has been validated against the McCall-Crabbs passages (Klare, 1974).

Studies have shown that scores of 90-100 characterize most comic books, scores of 60-90 characterize articles from the popular press (e.g., *Better Homes and Gardens, Newsweek*), and scores of 20-30 characterize reports from medical journals (e.g., *Journal of the American Medical Association, New England Journal of Medicine*) (Morrow, 1980).

In the computer adaptation of the Flesch Reading Ease formula, the syllable count is replaced by a vowel count, something computers can do more easily. Research by Coke & Rothkopf (1970) has showed that counting vowels provided very similar estimates as counting syllables.

The Flesch Reading Ease formula has not been adapted for Spanish language materials.

Findings

English Language Adult and Child CAHPS® Surveys

Table 2 (page 74) shows the readability formula and word and sentence difficulty results for the CAHPS® 2.0 adult and child core English language surveys. The average number of sentences and the average number of syllables are the main indicators of syntactic and semantic complexity used in all readability formulas except the SMOG. The average number of sentences per 100-word sample was 5.1 for the adult survey and 7.9 for the child survey. The average number of syllables per sentence per 100-word sample was 134.0 for the adult survey and 124.3 for the child surveys. In general, lower readability (less difficult) is assigned to written material that has shorter sentences and fewer syllables. The lower average number of sentences and higher average number of syllables in the adult survey may explain why the Flesch Reading Ease score for the adult survey is lower than that for the child survey (a lower score indicates more difficult text).

The FRG scores can be verified by plotting the average number of sentences and average number of syllables on Figure 1. The FRG results show that for both the adult and child surveys, a 7^{th} grade reading level is required for comprehension.

Applying the FOG Index to the adult and child surveys resulted in similar but not identical results. With the FOG Index, readability levels of 8^{th} grade for the adult survey and 6^{th} grade for the child survey were obtained. These results are consistent with the higher Flesch score obtained for the child survey than adult survey.

Recall that the SMOG Readability formula relies exclusively on counts of polysyllabic words found in three strings of 10 consecutive sentences selected randomly from the written material. Results from analyses using the SMOG are in agreement with results using the Fry graph that a 7th grade reading level is required for comprehension of both the adult and child surveys.

In Table 2, we also show the results of the readability formulas applied to a children's story (Kayner, 1999) and an article from national newspaper (*New York Times*, August 23, 1999: A1, A23). The FRG, Flesch Reading Ease score, FOG Index, and SMOG consistently rated the newspaper article at a higher level than either survey. The results of these analyses place the children's story at a reading level near that of both surveys.

Table 2 also shows the results of counts of syntactic and semantic components of the surveys and other materials. The sentence complexity counts (words per sentence, syllables per sentence, and characters per sentence) indicate that the adult survey had greater sentence complexity than the child survey and the *Cricket* reader, and less sentence complexity than the newspaper article. The counts of semantic factors
(1-syllable words, words with 2 or more syllables, number of characters per word, and number of syllables per word) are less easy to interpret. The newspaper had a greater average number of characters and syllables per word than either survey or the *Cricket* reader, indicating a greater use of longer words. The newspaper had a lower average number of 1-syllable words and a greater average of words with 2 or more syllables, also indicating a greater use of longer words.

Spanish Language Adult and Child CAHPS® Surveys

Table 3 (page 76) shows the readability formula and word and sentence difficulty results for the adult and child Spanish language surveys. The average number of sentences per 100-word sample was 6.8 for the adult survey and 4.4 for the child survey. The average number of syllables[4] per sentence per 100-word sample was 202.0 for the adult survey and 194.3 for the child surveys. Although the adult survey has more sentences and more syllables than the child survey, the results of the FRG indicate a 7th grade reading level for both.

The FRASE graph uses a similar method to the FRG to assess the readability of Spanish language materials. The FRASE graph results indicate that both the adult and child surveys require an *intermediate level* of reading skill to be fully comprehended. While the FRASE graph was intended to gauge the difficulty of materials used to teach Spanish as a second language, these results provide a useful indication of the readability level of the surveys. Furthermore, they provide a means of assessing the comparability of the child and adult survey readability levels.

Table 3 also shows the results of the readability formulas applied to an article from a Los Angeles Spanish language newspaper entitled "Resultado mixto en reduccion de clases" (*La Opinion*, June 24, 1999: A1) and a beginning reader, *Aventuras* (Freeman & Freeman, 1997). Both the FRG and FRASE graphs rate the readability of the surveys lower than the newspaper and higher than the reader.

---

[4] Unadjusted for the greater average number of syllables in Spanish language materials than in English language materials.

The syntactic counts (number of words, number of syllables, and number of characters) indicate that on average, the child survey sample had longer sentences than the adult survey. The semantic counts (number of characters and number of syllables per word) for both surveys were similar. The semantic counts of 1- and 2-syllable words were dropped from this analysis because the higher number of syllables in Spanish language materials makes them unreliable indicators of vocabulary complexity.

Discussion

The results of this study suggest that the CAHPS® 2.0 adult and child core surveys require a 7th grade reading level for adequate comprehension. The SMOG and Fry graphs both resulted in a 7th grade level readability assessment for the English language adult and child surveys. However, the FOG Index and the Flesch Reading Ease Score indicate that the adult survey may have a higher readability requirement than the child survey. This discrepancy may be due to greater sensitivity of the FOG Index and Flesch formulas to differences in number of sentence and/or number of syllables between the adult and child surveys than either the SMOG or FRG. While the FOG Index suggests that the magnitude of the difference between the adult and child surveys may be as great as 2 grade levels, it is difficult to determine the significance of the difference between Flesch scores of 71.3 and 89.6, since these scores are not tied to specific grade levels.

This study also shows that the English and Spanish versions of the CAHPS® surveys have comparable readability levels. Based on the Fry graph, both the English and Spanish adult and child versions for

the core CAHPS® surveys have7th grade readability levels. The similarity of the readability levels provides support for the success of the translation from English to Spanish.

Although the 7th grade reading level may be appropriate for commercially insured populations, it may be too high for Medicaid populations. According to the National Adult literacy survey (Kirsch et al., 1993), as many as 75% of welfare recipients read at or below the eighth grade level and 50% read at or below the fifth grade level. This suggests that the reading level required by the CAHPS® core surveys for full comprehension exceed the reading ability of more than 50% of welfare recipients. When one considers particular Medicaid beneficiary subgroups, the mismatch may be even greater.

A recent Public Policy Institute of California study reported that 42% of California Medicaid beneficiaries had less than a high school education (MaCurdy & O'Brien-Strain, 1997). Among recent immigrant Medicaid beneficiaries, 54% had less than a high school education; among Hispanic immigrant Medicaid beneficiaries who had arrived in the U.S. before 1985, 71% had less than a high school education. Since self-reported educational attainment tends to overstate literacy, the problem of low literacy and illiteracy among these groups is likely to be dramatic.

Poor comprehension of survey questions among those responding to patient surveys may also lead to unreliable results. For instance, adults with low literacy skills may not comprehend the term "health insurance plan." Indeed, cognitive interviews suggested that Medicaid beneficiaries frequently rated their overall care when asked to rate their health plan (Brown, 1996; Brown et al., 1999). Cognitive

interviews also found that Medicaid beneficiaries had trouble understanding the concept of a primary care provider or regular doctor and had trouble differentiating between a health plan and Medicaid (Brown, 1996; Brown et al., 1999).

Limitations of Readability Formulas

It is widely acknowledged that reading is an interactive process that occurs between the text and the reader. In fact, research shows that readers use experiences, knowledge, and information processing skills to comprehend text (Johnston, 1983).

Readability formulas, being strictly text-based, do not address the interactive nature of the reading process. Most reading formulas, including those used in this study, employ syntactic and semantic factors and do not directly address factors related to communicating meaning. For instance, readability formulas do not distinguish between written discourse and nonsensical combinations of words (Dreyer, 1984). Moreover, formulas cannot assess other critical factors such as the reader's interest, experience, knowledge or motivation, all of which may influence the reader's ability to comprehend the cognitive task asked by a survey (Duffy, 1985). Other factors related to readability and not assessed by a readability formula include typographical and temporal factors (e.g., time allotted to complete the reading task).

According to a recent paper on communicating with Medicaid beneficiaries, producing readable health materials requires thinking carefully about the audience to assess whether the intended respondents have the information with which to respond to the kinds of questions the survey asks (Hibbard, et al., 1997). It means organizing the material covered by the survey to make the survey easier to respond to,

and eliminating extra material that can overflow a page and overwhelm the survey respondent. It also means formatting a survey so that the instructions are simple to follow and using 12- to 14-point serif type, ample margins and headers to aid in organization. Finally, the overall content and design of the survey must be friendly, appealing and culturally appropriate to gain respondents' attention and increase their comprehension of important messages (Root & Stableford, 1999).

Many of the domains mentioned in the paragraph directly above were addressed during the development of the CAHPS® surveys. Cognitive interviews were used to identify items or terms that were difficult to comprehend, questions that were misinterpreted, and response options that that were inappropriate for the question or failed to capture the respondents' true experience (Harris-Kojetin et al., 1999). Literacy experts were consulted to improve readability of the survey (Brown et al., 1999). And careful translation procedures were followed to ensure the comparability of the English and Spanish versions of the surveys (Weidmer et al., 1999). These efforts provide additional evidence of the overall quality of the CAHPS® surveys.

This study is not intended to provide the definitive assessment of the readability of the CAHPS surveys. Rather, it aims to provide an additional rough gauge of their readability. Incidentally, a readability assessment by two literacy experts placed the readability level of the CAHPS surveys between the 6^{th} and 7^{th} grades (Julie Brown, personal communication, August 20, 1999).

Conclusions

Although the current readability level of the CAHPS® surveys may be appropriate for commercially insured populations, lower readability

is desirable for those who are publicly insured. As many as 50% of welfare recipients may fail to respond to the CAHPS® surveys because of a mismatch between the readability level of the surveys and the reading level of the intended respondents.

This situation may be exacerbated for certain subgroups of Medicaid beneficiaries, such as immigrants and refugees from less developed countries. According to research, non-English speaking patients and patients with low literacy skills face the greatest threat of receiving poor quality of care (Baker, Parker, Williams, Pitkin, Parikh, et al., 1996; Morales, Cunningham, Brown, Lui & Hays, 1999). Paradoxically, patients with low literacy skills also face the greatest barriers to responding to self-administered quality assessment tools such as the CAHPS® surveys.

Lowering the readability of the CAHPS® surveys, however, may be difficult. For reports about the CAHPS® surveys to help consumers make an informed choice about their health plan, the surveys need to collect information on a range of complex topics that require respondents to be familiar with concepts and vocabulary unique to health care. Shortening the survey and simplifying the vocabulary too much may cause the level of information gleaned from the CAHPS® surveys to fall, defeating the original purpose of CAHPS®.

Finding a balance between collecting important information and maintaining a reasonable level of survey readability will be an important consideration for researchers as future versions of the CAHPS® surveys are developed.

References

Baker,D., Parker, R., Williams, M., Pitkin, K, Parikh, N., Coates, W. & Imara, M. (1996). The health care experiences of patients with low literacy. *Archives of Family Medicine*, 5, 329-334.

Brown, J., Nederend, S., Hays, R., Short, P.& Farley, D. (1999). Special issues in assessing care of medicaid recipients. *Medical Care, 37*(3), MS79-MS88.

Brown, J. (1996). Report on cognitive interviews with Medicaid mothers for the Consumer Assessment of Health Plans Study. DRU-1471-AHCPR, Santa Monica, CA: RAND.

Cheng, L., & Yang, P. Q. (1996) The "Model Minority" deconstructed. In R. Waldinger & M. Bozorgmehr (eds.), *Ethnic Los Angeles* (pp.305-344). New York, NY: Russell Sage Foundation.

Coke, E., Rothkopf, E. (1970) Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology 54,* 208-210.

Dale, E. & Chall, J. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin, 28,* 37-54.

Dreyer, G. (1984). Readability and responsibility. *Journal of Reading, 27,* 334-339.

Duffy, T. (1985). Readability formulas: What's the use? In T. Duffy & R. Walker (eds.), *Designing Usable Texts* (pp. 113-143). Orlando, FL: Academic Press, Inc.

Flesch, R. (1948). A readability yardstick. *Journal of Applied Psychology, 32,* 221-233.

Fry, E. (1990). A readability formula for short passages. *Journal of Reading,* May, 594-597.

Fry, E. (1969).The Readability graph validated at primary levels. *The Reading Teacher, 22,* 534-538.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading, 21*(3), 242-252.

Gilliam, B., Pena, S. & Moutain, L. (1980). The Fry graph applied to Spanish readability. *The Reading Teacher,* January, 426-430.

Gates, A. & MacGinite, W. (1965). *Gates-MacGinite Reading Tests.* New York: Teachers College Press.

Gunning, R. (1968). The Fog Index after 20 years. *Journal of Business Communication, 6,* 3-13.

Hibbard, J., Slovic, P. & Jewett J. (1997). Informing consumer decisions in health care: Implications from decision-making research. *The Milbank Quarterly, 75*(3), 395-414.

Freeman, D. & Freeman, Y. (1997). *Aventuras.* Boston, MA: Houghton Mifflin Company.

Harris-Kojetin, L., Fowler, F., Brown, J., Schnaier, J. & Sweeny S. (1999). The use of cognitive interviews to develop and evaluate CAHPS®1.0 core survey items. *Medical Care, 37*(3), MS10-MS21.

Johnston, P. (1983). *Reading Comprehension Assessment: A Cognitive Basis.* Newark, DE: International Reading Association.

Kayner, G. (1999) Sun Flower. *Cricket, 26(12),* 4-7.

Kincaid, J., Fishburne, R., Rodgers, R. & Chissom, B. (1975). Derivation of new readability formulas for Navy enlisted personnel (Branch Report 8-75). Millington, TN: Chief of Naval Training.

Kirsch, I., Jungeblut, A., Jenkins, L. & Kolstad, A. (1993). *Adult Literacy in America.* Princeton, NJ: Educational Testing Service.

Klare, G. (1974). Assessing readability. *Reading Research Quarterly, 1,* 62-102.

Lewis, M., Merz, J., Hays, R., & Nicholas, R. (1995). Perceptions of intoxication and *impairment at arrest among adults convicted of driving under the influence of alcohol. Journal of Drug Issues, 25,* 141-160.

Leadbetter, C., Hall, S., Swanson, J. & Forrest, K. (1990). Readability of commercial versus generic health instructions for condoms. *Health Care for Women International, 11,* 295-304.

Lopez, D. (1996). Language and assimilation. In R. Waldinger & M. Bozorgmehr (eds.), *Ethnic Los Angeles* (pp. 139-163). New York: Russell Sage Foundation.

MaCurdy, T. & O'Brien-Strain, M. (1997). *Who Will Be Affected by Welfare Reform in California?* San Francisco, CA: Public Policy Institute of California.

McCall, W. & Crabbs, L. (1961). *Standard Test Lessons in Reading.* New York: Bureau of Publications, Teachers College, Columbia University.

McLaughlin, G. (1969). SMOG grading - a new readability formula. *Journal of Reading, 12,* 636-646.

Morales, L., Cunningham, W., Brown, J., Lui, H.& Hays, R. (1999). Are Latinos less satisfied with communication by providers? *Journal of General Internal Medicine, 14,* 409-417.

Morrow, G. (1980). How readable are subject consent forms? *JAMA, 244*(1), 56-58.

Romano, R. (1979). *Readability in Cancer Communications: Methods, Examples and Resources for Improving the Readability of Cancer Messages and Materials*. Bethesda, MD: U.S. Department of Health, Education and Welfare, Public Health Service, National Institutes of Health, National Cancer Institute.

Root, J. & Stableford, S. (1999). Easy to read consumer communication: A missing link in medicaid managed care. *Journal of Health Politics, Policy and Law, 24*(1), 1-26.

Rush, R. (1985). Assessing readability: Formulas and alternatives. *Reading Teacher, 39,* 274-283.

Vari-Cartier, P. (1981). Development and validation of a new instrument to assess the readability of Spanish prose. *Modern Language Journal, 65*(Summer), 141-148.

Weech-Maldonado, R., Weidmer, B. O., Morales, L. S. & Hays, R. D. (In Press) Cross-cultural adaptation of survey instruments: The CAHPS® experience. In D. O'Rourke (ed.), Health Survey Research Methods: Seventh Conference Proceedings.

Weidmer, B., Brown, J., Garcia, L. (1999). Translating the CAHPS 1.0 Survey Instruments into Spanish. *Medical Care, 37*(3), MS89-MS96.

TABLE 1. Syntactic and Semantic Factor Counts Used in Readability Assessment.

	Syntactic (sentence) Factor	Semantic (word) Factor
Average number of sentences	√	
Average number of words per sentence	√	
Average number of syllables per sentence	√	
Number of characters per sentence	√	
Average number of syllables		√
Average number of 1-syllable words		√
Average number of 2-syllable words		√
Average number of characters per word		√
Average number of syllables per word		√

TABLE 2. Readability Levels of English Language CAHPS® 2.0 Surveys, English Language Newspaper and English Language Children's Book.

	CAHPS® 2.0 Adult Core	CAHPS® 2.0 Child Core	New York Times Article	Cricket Reader (Ages 9 and up)
Fry Readability Graph Score	7th Grade	7th Grade	12th Grade	5th Grade
Average number of sentences per 100-word sample	5.1	7.9	3.4	9.0
Average number of syllables per 100-word sample	134.0	124.3	153.3	133.0
Flesch Reading Ease Score	71.3	89.6	45.8	81.3
FOG Readability Score	8th Grade	6th Grade	12th Grade	5th Grade
SMOG Readability Score	7th Grade	7th Grade	12th Grade	7th Grade
Syntax Indexes				
Average number of words per sentence	19.8	15.2	30.7	11.4
Average number of syllables per sentence	26.5	18.9	46.4	15.1
Average number of characters per sentence	81.2	61.3	141.7	49.9
Semantic Indexes				
Average number of 1-syllable words per 100 words	76.3	83.7	65.3	75.7
Average number of 2- or more syllable words per 100 words	23.7	16.3	34.7	24.3
Average number of characters per word	4.1	4.1	4.7	4.4
Average number of syllables per word	1.3	1.2	1.5	1.3

Note. Fry Readability Graph and Flesch Reading Ease Score based on three 100-word passages taken from the beginning, middle and end of

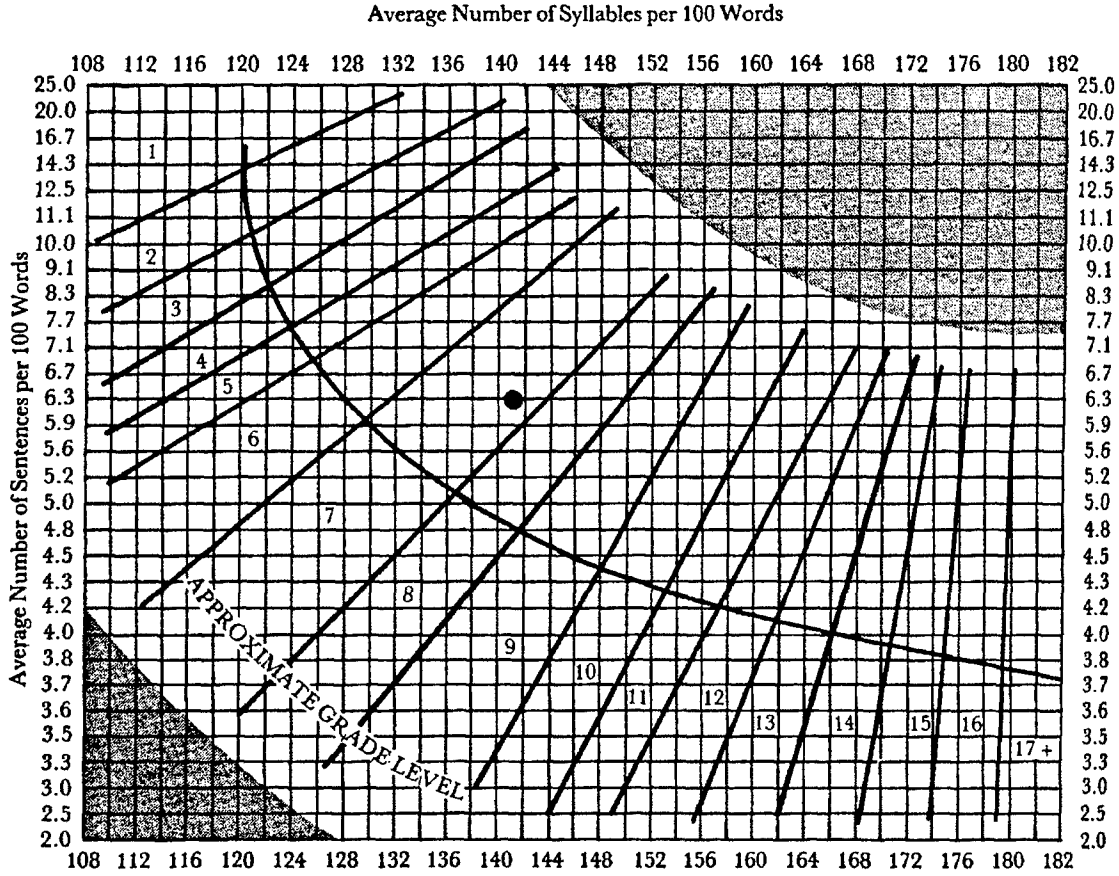
each document. The SMOG score is based on 3 continuous 10-sentence samples taken from the beginning, middle and end of each document.

TABLE 3. Readability Levels of Spanish Language CAHPS® 2.0 Surveys, Spanish Language Newspaper and Spanish Language Children's Book.

	CAHPS® 2.0 Adult Core	CAHPS® 2.0 Child Core	La Opinion	Aventuras
Fry Readability Graph	7th Grade	7th Grade	14th Grade	1st Grade
Average number of sentences per 100-word samples	6.8	4.4	2.8	16.7
Average number of syllables per 100-word samples	202.7	194.3	235.0	195.7
FRASE Graph	Intermediate	Intermediate	Advanced	Beginning
Syntax Indexes				
Average number of words per sentence	15.6	24.0	38.0	6.0
Average number of syllables per sentence	31.4	46.5	88.2	11.9
Average number of characters per sentence	74.0	110.5	191.6	26.0
Semantic Indexes				
Average number of characters per word	4.8	4.6	5.0	4.3
Average number of syllables per word	2.0	1.9	2.4	2.0

Note. Fry Readability Graph score based on three 100-word passages taken from the beginning, middle and end of each document. The FRASE assessment is based on 5 100-word samples taken from the document.

FIGURE 1
Graph for Estimating Readability—Extended
by Edward Fry, Rutgers University Reading Center, New Brunswick, NJ 08904



DIRECTIONS: Randomly select 3 one hundred word passages from a book or an article. Plot average number of syllables and average number of sentences per 100 words on graph to determine the grade level of the material. Choose more passages per book if great variability is observed and conclude that the book has uneven readability. Few books will fall in gray area but when they do grade level scores are invalid.

Count proper nouns, numerals and initializations as words. Count a syllable for each symbol. For example, "1945" is 1 word and 4 syllables and "IRA" is 1 word and 3 syllables.

4. ARE LATINOS LESS SATISFIED WITH COMMUNICATION BY HEALTH CARE PROVIDERS? A STUDY OF 48 MEDICAL GROUPS

ABSTRACT

Purpose: To examine associations of patient ratings of communication by health care providers with patient language (English vs. Spanish) and ethnicity (Latino vs. white).

Methods: A random sample of patients receiving medical care from a physician group association concentrated on the West Coast was studied. A total of 7,093 English and Spanish language questionnaires were returned for an overall response rate of 59%. Five questions asking patients to rate communication by their health care providers were examined in this study. All five questions were administered with a 7-point response scale.

Results: We estimated the associations of satisfaction ratings with language (English vs. Spanish) and ethnicity (white vs. Latino) using ordinal logistic models, controlling for age and gender [this only refers to the 1st model?). Latinos responding in Spanish (Latino/Spanish) were significantly more dissatisfied compared with Latinos responding in English (Latino/English) and non-Latino whites responding in English (white) when asked about: (1) the medical staff listened to what they say (29% vs. 17% vs. 13% rated this "*Very Poor,*" "*Poor,*" or "*Fair*"; $p<0.01$); (2) answers to their questions (27% vs. 16% vs. 12%; $p<0.01$); (3) explanations about prescribed medications (22% vs. 19% vs. 14%; $p < 0.01$); (4) explanations about medical procedures and test results (36% vs. 21% vs. 17%; $p < 0.01$); and (5) reassurance

and support from their doctors and the office staff (37% vs. 23% vs. 18%; $p<0.01$).

Conclusion: This study documents that Latino/Spanish respondents are significantly more dissatisfied with provider communication than Latino/English and white respondents. These results suggest Spanish speaking Latinos may be at increased risk for lower quality of care and poor health outcomes. Efforts to improve the quality of communication with Spanish speaking Latino patients in outpatient health care settings are needed.

INTRODUCTION

Although many studies have documented access to care barriers faced by Latinos (Andersen, Lewis, Giachello, Aday, & Chiu, 1981; Ginzberg, 1991; Schur, White, & Berk, 1995; Schur & Albers 1996; Valdez, Giachello, Rodriguez-Trias, Gomez, & de la Rocha, 1993), relatively few studies have examined satisfaction with care in this population once they have access to the health care system. Assessing satisfaction with care among Latinos, like other ethnic minority patient populations, is important because Latino patients have unique cultural and linguistic needs that are frequently not well served by the current health care system which is oriented to serving patients belonging to the dominant culture (Molina, Zambrana & Aguirre-Molina, 1997; Woolley, Kane, Hughes, & Wright, 1978; California Cultural Competency Task Force, 1994; Lavizzo-Mourey & Mackenzie, 1996). Moreover, this relative scarcity of research on satisfaction with care among Latinos exists at a time when the Latino population is growing rapidly, particularly in states such as California where Latinos already account for nearly a third of the resident population (McDonnell 1997).

The research on satisfaction with care among Latinos that does exist, tends to run in two general veins; comparisons of satisfaction between Latino and non-Latino patients, and comparisons between Spanish and English speaking patients. The results of research comparing satisfaction with care among Latinos and non-Latinos are mixed. On the one hand, in one of the first large studies of health care use by Latinos, Andersen and colleagues (1981) found that Latinos were more dissatisfied with appointment waiting time, information provided by

their physician and time spent with their physician than the general population. On the other hand, a more recent meta-analysis of patient sociodemographic characteristics and satisfaction concluded that there was no overall relationship between ethnicity and satisfaction with care while greater age and less education were positively associated with satisfaction (Hall & Dornan, 1990). Similarly, a recent study of satisfaction with care among clinic outpatients failed to find an association between race (including Latino) and patient satisfaction with provider communication or courtesy of the office staff (Harpole, Orav, Hickey, Posther, & Brennan, 1996).

The results of research comparing satisfaction with care among Spanish and English speaking patients are clearer; Spanish speaking patients tend to be more dissatisfied with care than English speaking patients. In a study of interpreter use in emergency rooms, Baker showed that monolingual Spanish speaking patients were more dissatisfied with communication than English speaking patients even with the use of interpreters (Baker, Parker, Williams, Coates & Pitkin, 1996). Hu and Covell (1986) found that outpatients whose primary language was English were more satisfied with their care in general than were patients whose primary language was Spanish and Harpole and colleagues (1996) found that Spanish speaking patients were less satisfied with office staff courtesy, but not communication with providers or timeliness of care. Other patient characteristics found to be associated with greater dissatisfaction with care include being unmarried, poorer health status, and younger age (Schur & Albers, 1995; Hall & Dornan 1990; DiMatteo & Hays, 1980; Linn & Greenfield, 1996; Cleary & McNeil, 1988; Aharony & Strasser, 1993).

In this study we investigate the association of patient ratings of communication by providers with ethnicity (Latino versus white) and language (Spanish versus English). In order to isolate the effects of language and ethnicity on satisfaction with communication we have included three comparison groups: non-Latino whites responding in English (whites); Latinos responding in Spanish (Latino/Spanish); and Latinos responding in English (Latino/English). At the outset of this study, we hypothesized that Latino/Spanish respondents would express the most dissatisfaction with communication because they were most likely to face language and cultural barriers to communication; followed by Latino/English respondents who may face cultural but not language communication barriers; followed by whites who are least likely to face either language or cultural barriers to communication.

METHODS

This analysis was based on survey data obtained from randomly selected patients receiving medical care from an independent association of physician groups located primarily in the western United States. The survey was designed to ask individuals about their health status, satisfaction with care, and use of health services during the past 12 months. At the time of the study, approximately two-thirds of the association's member medical groups were located in California. Of the 48 medical groups in the association participating in the study, 32 groups were located in Southern California, 10 groups were located in Northern California, and 21 groups were located in other states (Washington, Oregon, Texas, Arizona and New Jersey).

Patients at least 18 years of age and with a minimum of one provider visit during the 365 days prior to the study were considered

eligible for the survey. Each selected patient was mailed both Spanish and English language versions of the 12-page opscan questionnaire and cover letter along with a $2 cash payment and a return envelope. One week later, each individual was mailed a reminder/thank you postcard. Two weeks later, non-respondents were mailed a second packet of materials and a reminder telephone call was attempted. Each non-respondent was called back a maximum of six times. A total of 18,480 surveys were mailed out and 7,093 returned for an overall response rate of 59%, adjusting for undeliverable surveys, ineligible respondents, and deceased. Response rates across medical groups ranged from 46% to 73%, and were not significantly associated with ratings of health care (Hays, Brown, Spritzer, Dixon, & Brook, 1998).

A detailed description of the survey, including it's psychometric properties, is reported elsewhere (Hays, Brown, Spritzer, Dixon, & Brook, 1998). Briefly, the Spanish language version of the survey was created through a process of independent forward (English to Spanish) and back (Spanish to English) translation followed by reconciliation. The questionnaire included 153 items assessing the following: (1) intention to switch to another physician group; (2) intention to switch to another health plan; (3) ratings of care including ratings of communication health care providers; (4) reports about care; (5) utilization of care; (6) health status; and (7) a chronic condition inventory. The survey took approximately 27 minutes to complete. Overall, the health care rating questions showed excellent construct validity as measured by product-moment correlations between ratings of care and intentions to switch physician groups, continuity of care and

reports about care. The field period began October of 1994 and ended in June of 1995.

Dependent Variables

To assess satisfaction with provider communication, respondents were asked to rate five facets of provider communication: (1) medical staff listening to what you have to say (el personal médico presentando atención a lo que usted dice); (2) answers to your questions (las respuestas a sus preguntas); (3) explanations about prescribed medications (las explicacions sobre las medicinas que le recetan); (4) explanations about medical tests and procedures (las explicaciones de los procedimientos medicos y los resultados de los análisis); and (5) reassurance and support from your doctor and support staff (la tranquilidad y apoyo que le ofrecen los médicos y el personal). Each question was administered using a 7-point response scale (*Very Poor, Poor, Fair, Good, Very Good, Excellent, and The Best*) (*Muy Malo, Malo, Más o Menos, Bueno, Muy Bueno, Excelente/Buenisimo, Lo Mejor*) along with the option *Does Not Apply to Me* (*No Se Refiere a Mí*).

Independent Variables

Based on a review of the literature, three types of potential confounding variables were considered: demographic, socioeconomic including health insurance status, and health status. The following demographic variables were included in this analysis: gender (male; female) and age (60 or less; over 60). The following socioeconomic variables were included in this analysis: education (less than high school; high school; and more than high school), household income ($20,000 or less annual household income; more than $20,000), household

size (two or less persons; more than two persons), and insurance status (private, Medicare, Medicaid, other, or uninsured). Health status measures included in this analysis were a physical health composite score, a mental health composite score, and a checklist of comorbid conditions. The physical and mental health composite scores were derived from the RAND-36 Health Survey (Hays, Sherbourne & Mazel, 1993). The checklist of comorbid conditions inquired about presence of twenty-four different medical conditions, including prostate conditions for men and abnormal vaginal bleeding for women (see Appendix A).

Because respondents were allowed to identify more than one source of insurance coverage, we derived a single hierarchical variable that reflects a rank ordering of reported coverage. Persons were classified as having private insurance if they reported HMO, IPA, PPO or fee-for-service insurance. Persons who did not report private insurance but did report Medicaid coverage were classified as covered by Medicaid insurance (e.g., this included persons reporting Medicaid and Medicare coverage). Persons who did not report private or Medicaid coverage but did report Medicare coverage were classified as covered by Medicare. Persons who had none of these types of insurance coverage but did report "Other" insurance were classified as having other insurance. Finally, those who did not report coverage from any source, were classified uninsured.

A "Spanish Language Response Variable" (SLVR) was also used in this analysis. This variable controlled for potential differences in response patterns between Spanish and English language respondents attributable to linguistic and cultural differences in use of the response scale. Research has shown a potential problem with Spanish

translated Likert-type response scales (Angel & Guarnaccia, 1989; Hays & Baker, 1998; Shetterly, Baxter, Mason, & Hamman, 1996). The SLRV survey item asked about satisfaction with parking (How Do You Rate Arrangements for Parking?) using the same 7-point response scale used for the dependent variables. Assuming similar parking opportunities for Spanish and English language respondents, adding the SLRV to multivariate models of satisfaction with communication should statistically control for differences in ratings between Spanish and English language respondents attributable to linguistic and cultural differences in using the response scale alone.

Analysis Plan

Survey respondents included in this analysis were Latino/Spanish respondents, Latino/English respondents, and white respondents. Other respondents, including African Americans or Blacks, Asians or Pacific Islanders, Native Americans or American Indians, and those reporting their race/ethnicity as "Other," were dropped from the analysis. Of the total number of survey respondents (n=7,093), 88% were retained for this analysis (n=6,211).

Differences in demographic, socioeconomic and health status characteristics among Latino/Spanish respondents, Latino/English respondents and white respondents were examined using bivariate statistics. For categorical and continuous variables, chi-square and analysis of variance (ANOVA) were used, respectively.

Analyses of the five communication ratings questions were carried out in two steps. First, a communication summary score was constructed by averaging together the five provider communication ratings

questions. Then the score was normalized to a mean of 50 and standard deviation of 10 (T-score). T-scores were used rather than raw scores in order to ease interpretation (e.g., a score of 40 is one standard deviation below the overall sample mean). Associations between this score and each independent variable was examined using ANOVA and OLS regression. For these analyses, the satisfaction score was assumed to have interval scale properties.

Second, each satisfaction with communication question was independently modeled using multivariate ordinal logistic regression. Since subjects belonged to 1 of 48 medical groups, standard errors were adjusted (using a Huber correction) for potential intra-cluster variability. In total, we estimated three models for each satisfaction with communication question. In the first regression (model 1) we controlled for age and gender. In the second regression (model 2) we controlled for age, gender and the SLRV. In the third regression (model 3) we controlled for age, gender, income, household size, education, insurance status, health status and the SLRV.

The total number of response categories were reduced from seven (*Very Poor, Poor, Fair, Good, Very Good, Excellent, and The Best*) to five (*Very Poor/Poor, Fair, Good, Very Good, and Excellent/The Best*) in order to adequately satisfy the parallel slope assumption of the ordinal logistic model. Satisfaction of this assumption was tested using the chi-square score test in the SAS Logistic Procedure (SAS Institute, Inc., 1989). All other statistical analysis presented in this study were conducted using STATA, Version 5 (StataCorp, 1997). In accordance with the recommendations of DuMoucel and Duncan (1983), sampling weights are not used in the regression models.

RESULTS

Those returning the questionnaire had a mean age of 51 years (median, 49 years) compared with the mean age of the sampling frame that was 46 years (median, 43 years). Sixty-five percent of the responders were women, whereas only 58% in the sampling frame were women. The last medical visit for the study participants was, on average, 119 days (median, 88 days) before the beginning of the study. For those in the sampling frame, the average was 130 days (median 112 days). Four percent of the respondents and 3% of the sampling frame had hypertension as the last diagnosis recorded (according to the *International Classification of Diseases, Ninth Revision*, code) (Hays, Brown, Spritzer, Dixon, & Brook, 1998; World Health Organization, 1977).

Sample Characteristics

Latino/Spanish respondents compared with Latino/English respondents and whites reported lower educational attainment (<HS, 59% vs. 21% vs. 8%); lower annual income ($20,000 or less, 69% vs. 24%, 21%); larger family size (two or more persons, 87% vs. 68% vs. 43%); younger age (years, 40.2 vs. 42.2 vs. 51.9); fewer mean number of comorbid conditions (2 vs. 3 vs. 3); and were more likely married (married, 90% vs. 74% vs. 74%) (Table 1, page 102). The proportion of female respondents was smaller among Latino/Spanish respondents (56%) than among Latino/English respondents (65%) and whites (65%). Private health insurance was most commonly reported by whites (88%),followed by Latino/English respondents (84%) and Latino/Spanish respondents (64%). Having no insurance was most commonly reported by Latino/Spanish respondents (7%), followed by Latino/English respondents (2%) and

whites (1%). There was no meaningful difference between Latino/Spanish respondents, Latino/English respondents, and whites with respect to physical health (physical health index, 50 vs. 51 vs. 50) or mental health (mental health index, 50 vs. 49 vs. 50). However, Latino/Spanish respondents did report a lower average number of health conditions compared with Latino/English respondents and whites (1.9 vs. 2.5 vs. 3.1).

Satisfaction with Communication

Overall, Latinos reported greater average dissatisfaction with communication than whites (Table 2, page 103). Latino/Spanish respondents rated provider communication 5.4 points lower than whites (greater than one half standard deviation below the overall mean), while Latino/English respondents rated provider communication 1.7 points lower than whites. A difference of 2.5 points separated the average satisfaction ratings of older (60 + years) and younger (<60 years) patients. Other differences in average satisfaction ratings by respondent characteristics included: (1) a 0.4 point difference between males and females; (2) a 0.4 point difference between married versus not married; (3) a 0.2 point difference between education groups; (4) a 0.2 point difference between income groups; and (5) a 4.2 point difference between Medicare and uninsured respondents. In an OLS regression controlling for age, gender, physical and mental health, education, income, SLRV, insurance status and language/ethnicity, we found significant positive associations between the communication summary score and age ($p<0.01$), physical health ($p<0.01$), mental health ($p<0.01$), other insurance ($p<0.01$), and Latino/Spanish respondents ($p<0.01$).

Ordinal logistic models of individual satisfaction with communication questions also showed significant differences in ratings between Latino and whites respondents (Table 3, page 105). Table 3 only displays adjusted proportions using model 1 (adjusting for age and gender) because all models produced nearly identical results. To the question "How Would You Rate Medical Staff Listening to What You Have to Say?" 28.8% percent of Latino/Spanish respondents answered *Very Poor/Poor* or *Fair* compared with 17.2% of Latino/English respondents and 13.4% of whites. To the question "How Would You Rate Answers to Your Questions?" 26.6% of Latino/Spanish respondents answered *Very Poor/Poor or Fair* compared with 16.0% of Latino/English respondents and 12.4% of whites. To the question "How Would You Rate Explanations About Prescribed Medications?" 30.5% of Latino/Spanish respondents answered *Very Poor/Poor* or *Fair* compared with 18.6% of Latino/English respondents and 14.0% of whites. To the question "How Would You Rate Explanations About Medical Tests and Procedures" 36.0% of Latino/Spanish respondents answered *Very Poor/Poor* or *Fair* compared with 21.2% of Latino/English respondents and 17.3% of whites. Finally, to the question "How Would You Rate Reassurance and Support Form Your Doctor and the Office Staff," 28.8% of Latino/Spanish respondents answered *Very Poor/Poor* or *Fair* compared with 17.3% of Latino/English respondents and 13.4% of whites.

Controlling for covariates had minimal effects on the distribution of patient rating scores (Table 4, page 107). Table 4 shows the unadjusted and adjusted (models 1-3) distribution of responses to the question "How Would You Rate Medical Staff Listening to What You Say." Reading across table 4 shows the distribution of responses by model

within ethnic/language group. For example, the proportion of Latino/Spanish respondents answering *Very Poor/Poor* or *Fair* ranged from 27.7% (unadjusted responses) to 31.8% (model 3). Among whites, the proportion of respondents answering *Very Poor/Poor* or *Fair* ranged from 13.4% (unadjusted responses) to 13.7% (model 3). This demonstrates that the effect of alternative model specifications on the distribution of rating scores was minimal. Table 4 only presents this analysis for the question "How Would You Rate Medical Staff Listening to What You Say." The identical analyses of the four other communication ratings questions yielded similar results and thus are not shown in this paper.

SUMMARY AND DISCUSSION

This study evaluated satisfaction with provider communication among a sample of Latino and non-Latino patients responding to a patient satisfaction survey in Spanish and English. We show that Latino/Spanish respondents are significantly more dissatisfied with provider communication than Latino/English and white respondents. These disparities were not accounted for in multivariate regression models controlling for confounding variables such as age, gender, education or insurance status. We also show that Latino/English respondents are somewhat more dissatisfied with provider communication than whites, though this finding did not reach statistical significance.

Comparisons of satisfaction ratings by a number of demographic characteristics have been reported in the literature (Harpole, Orav, Hickey, Posther & Brennan, 1996; Sisk, Gorman, Reisinger, Glied, DuMouchel, & Hynes, 1996). These include age, gender and insurance status. In contrast to these same comparisons made in our study sample, the disparities in provider communication ratings by

ethnicity/interview language are substantial. For example, the disparity between Latino/Spanish and white respondents is 5.4 points (Table 2) compared with 2.5 points by age, 0.4 points by gender, 2.5 points by insurance status, and 0.2 points by annual income.

We also found a small difference in satisfaction ratings between Latino/English and white respondents. The disparity between Latino/English and white respondents was 1.7 points, which was greater than disparities we detected by gender (0.4 points), marital status (0.4 points), and education (0.2 points). The difference in provider communication ratings between Latino/English and white respondents may reflect more subtle and less easily measured, but no less salient, barriers to patient-physician communication. For example, greater differences in social class between physicians and their Latino/English patients than between physicians and their white patients may account for this finding.

If the disparities in satisfaction ratings between Latino and white patients reflect actual differences in quality of provider communication, then Latino patients, particularly Spanish-speaking Latino patients, are at increased risk for poor quality of care and poor treatment outcomes. Research shows that Latino patients are at risk for low quality of care compared with non-Latino whites (Todd, Samaroo, & Hoffman, 1993) and poorer treatment outcomes when there is not language concordance between the patient and provider (Perez-Stable, Napoles-Springer, & Miramontes, 1997). Unsatisfactory communication between Spanish speaking patients and their providers may result in lower quality of care and poorer treatment outcomes in a variety of ways. Poor communication between a physician and patient,

as indicated by dissatisfaction with provider listening and answering of questions, may result in excessive ordering of medical tests as a provider attempts to establish a diagnosis in the absence of an adequate patient history. Spanish-speaking patients receiving unsatisfactory explanations about taking their prescribed medications may inadvertently take them inappropriately, resulting in less than optimal outcomes including medication toxicities regardless of whether or not the prescriptions were technically appropriate. Greater dissatisfaction with care among Latino patient may also result in increased in plan disenrollment, doctor shopping and inappropriate follow-up (Manson, 1988; Rosen, Sanford, & Scott, 1991).

Because optimal treatment outcomes depend upon satisfactory communication between patients and physicians about medical test results, medications and treatment options, special attention should be given to improving communication with Spanish speaking patients. Various strategies for improving communication with Spanish speaking patients have been described in the literature (Baker, Parker, Williams, Coates, & Pitkin, 1996; Hornberger, Gibson, Wood, Dequeldre, Corso et al, 1994, Woloshin, Bicknell, Schwartz, Gany, & Welch, 1995). Among these, increased access to and use of professional interpreters is frequently mentioned. Professional interpreters can significantly improve satisfaction with care among Spanish speaking patients (Baker, Parker, Williams, Coates, & Pitkin, 1996). Moreover, use of professional interpreters improves Spanish speaking patient's understanding of their disease (Baker, Parker, Williams, Coates, & Pitkin, 1996). Bilingual doctors who have adequate fluency in Spanish can also improve Spanish speaking patients understanding of their

disease and satisfaction with care (Baker, Parker, Williams, Coates, & Pitkin, 1996). Bilingual doctors have also been shown to improve outcomes among Spanish speaking patients with hypertension and diabetes (Perez-Stable, Napoles-Springer, & Miramontes, 1997). Other strategies to improve the quality of care for linguistic and ethnic minority patients include teaching medical Spanish to health care providers, educating health care providers about the health beliefs and practices of their patients (California Cultural Competency Task Force, 1994) and the development of clinical practice guidelines that ensure cultural competence (Lavizzo-Mourey & Mackenzie, 1996).

Findings from this study need to be interpreted with caution for several reasons. First, those who participated in the study were similar but not identical to those in the sampling frame (Hays, Brown, Spritzer, Dixon, & Brook, 1998). Moreover, since language preference and race/ethnicity were self-reported and not available through administrative records, we were unable to calculate the language or race/ethnic specific response rates. Had we been able to adjust for Spanish language non-response, however, it is likely we would have found even greater disparities in provider communication ratings between English and Spanish language respondents since Spanish speaking patients are probably faced with the greatest communication barriers (including lower literacy) are least likely to respond to the survey. Second, our satisfaction rating scale (*Very Poor, Poor, Fair, Good, Very Good, Excellent, and The Best*) might have been interpreted differently by Spanish and English language survey respondents. Other reports in the literature suggest that Spanish language respondents tend to score lower on some rating scales (e.g., *Poor to Excellent*)

than English language respondents (Angel & Guarnaccia, 1989; Hays & Baker, 1998); thus, the direction of a language response bias, if present, would inflate disparities between Spanish and English language respondents. Analytically, we account for a potential language response bias by including the "Spanish Language Response Variable" in our analysis. Since adding the SLRV does not significantly change the results of our study, the satisfaction disparities we have identified are unlikely to be entirely attributable to a differential interpretation of the rating scale. Finally, this survey was conducted in Western United States where Mexican-Americans are the predominant Spanish-speaking ethnic group. Thus, the results of this study may not generalize to other U.S. Spanish-speaking ethnic groups such as Puerto Ricans or Cubans.

Our results suggest that health plans and other large providers of medical care to Latino patients should monitor patient dissatisfaction with provider communication and examine it's association with treatment outcomes. Satisfaction with care tools may be used to monitor treatment outcomes within and among health plans and aid Latino patients in choosing among multiple providers of care. When appropriately constructed, administered and reported, these tools may help to focus provider attention on specific aspects of patient-provider communication such as explanations about medications, treatment side effects, giving consent, or advanced directives.

REFERENCES

Aharony L, Strasser S. Patient satisfaction: what we know about and what we still need to explore. Med Care Rev. 1993;50(1):49-79.

Andersen R, Lewis SZ, Giachello AL, Aday LA, Chiu G. Access to medical care among the Hispanic population of the southwestern United States. J Health Soc Behav. 1981;22(1):78-89.

Angel R, Guarnaccia PJ. Mind, body, and culture: Somatization among Hispanics. Soc Sci Med. 1989;28(12):1229-1238.

Baker DW, Parker RM, Williams MV, Coates WC, Pitkin K. Use and effectiveness of interpreters in an emergency department. JAMA. 1996;275(10):783-8.

California Cultural Competency Task Force. Recommendations for the Medi-Cal Managed Care Program. California Department of Health Services; 1994.

Cleary PD, McNeil BJ. Patient satisfaction as an indicator of quality care. Inquiry. 1988;25(1):25-36.

DiMatteo MR, Hays R. The significance of patients' perceptions of physician conduct: a study of patient satisfaction in a family practice center. J Community Health. 1980;6(1):18-34.

DuMouchel WH, Duncan GJ. Using sample survey weights in multiple regression analysis of stratified samples. JASA. 1983;78(383):535-543.

Ginzberg E. Access to health care for Hispanics. JAMA. 1991;265(2):238-41.

Hall JA, Dornan MC. Patient sociodemographic characteristics as predictors of satisfaction with medical care: a meta analysis. Soc. Sci. Med. 1990;30(7):811-818.

Harpole LH, Orav EJ, Hickey M, Posther KE, Brennan TA. Patient satisfaction in the ambulatory setting. Influence of data collection methods and sociodemographic factors. J Gen Intern Med. 1996;11(7):431-4.

Hayes RP, Baker DW. Methodological problems in comparing English-speaking and Spanish-speaking patients' satisfaction with interpersonal aspects of care. Med Care. 1998;36(2):230-6.

Hays RD, Brown JA, Spritzer KL, Dixon WJ, Brook RH. Member ratings of health care provided by 48 physician groups. Arch Intern Med. 1998;158(7):785-90.

Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. Health Econ. 1993;2(3):217-27.

Hornberger JC, Gibson CD, Jr., Wood W, Dequeldre C, Corso I, Palla B, & Bloch DA. Eliminating language barriers for non-English-speaking patients. Med Care. 1996;34(8):845-56.

Hu DJ, Covell RM. Health care usage by Hispanic outpatients as a function of primary language. West J Med. 1986;144(4):490-93.

Lavizzo-Mourey R, Mackenzie ER. Cultural competence: essential measurements of quality for managed care organizations. Ann Intern Med. 1996;124(10):919-21.

Linn LS, Greenfield S. Patient suffering and patient satisfaction among the chronically ill. Med Care. 1982;20(4):425-31.

Manson A. Language concordance as a determinant of patient compliance and emergency room use in patients with asthma. Med Care. 1988;26:1119-1128.

McDonnell PJ. Latinos face civil rights crisis, panel told. Los Angeles Times 1997 December 4, 1997; B1.

Molina CW, Zambrana RE, Aguirre-Molina M. The influence of culture, class, and environment on health care. In: Molina CW, Aguirre-Molina M, eds. Latino health in the U.S.: A growing challenge. Washington, D. C.: American Public Health Association; 1997:23-43.

Perez-Stable EJ, Napoles-Springer A, Miramontes JM. The effects of ethnicity and language on medical outcomes of patients with hypertension or diabetes. Med Care. 1997;35(12):1212-9.

Rosen K, Sanford S, Scott J. Emergency department care of the Spanish-speaking patient. Ann Emerg Med. 1991;20(466).

SAS Institute, Inc. SI. SAS/STAT User's Guide, Version 6. 4 ed. Vol. 2 Cary, NC; 1989.

Schur CL, Albers LA. Language, sociodemographics, and health care use of Hispanic adults. J Health Care Poor Underserved. 1996;7(2):140-58.

Schur CL, White LA, Berk ML. Health care use by Hispanic adults: the role of financial vs. non-financial determinants. AHSR FHSR Annu Meet Abstr Book. 1995;12:103.

Shetterly SM, Baxter J, Mason LD, Hamman RF. Self-rated health among Hispanic vs non-Hispanic white adults: the San Luis Valley Health and Aging Study. Am J Public Health. 1996;86(12):1798-801.

Sisk JE, Gorman SA, Reisinger AL, Glied SA, DuMouchel WH, Hynes MM. Evaluation of Medicaid managed care. Satisfaction, access, and use. JAMA. 1996;276(1):50-5.

StataCorp. Stata Statistical Software: Release 5.0 College Station, TX: Stata Corporation; 1997.

Todd KH, Samaroo N, Hoffman JR. Ethnicity as a risk factor for inadequate emergency department analgesia. JAMA. 1993;269(12):1537-9.

Valdez RB, Giachello A, Rodriguez-Trias H, Gomez P, de la Rocha C. Improving access to health care in Latino communities. Public Health Rep. 1993;108(5):534-9.

Woloshin S, Bickell NA, Schwartz LM, Gany F, Welch HG. Language barriers in medicine in the United States. JAMA. 1995;273(9):724-8.

Woolley FR, Kane RL, Hughes CC, Wright DD. The effects of doctor-patient communication on satisfaction and outcome of care. Soc Sci Med. 1978;12(2A):123-8.

World Health Organization. International Classification of Diseases, Ninth Revision (ICD-9). 1977.

Table 1. Sample Characteristics by Ethnic Background and Interview Language

Variable	Latino/ Spanish (n=181)	Latino/ English (n=532)	White/ English (n=5,498)	p Value*
Age (mean)	40	42	52	<.01
Gender (%)				
Female	56	65	65	.04
Education (%)				
Less Than HS	59	21	8	<.01
HS	20	24	23	
More Than HS	21	55	69	
Annual Income (%)				
$20,000 or Less	69	24	21	<.01
Marital Status(%)				
Married	90	74	74	<.01
Household Size (%)				
2+ persons	87	68	43	<.01
Insurance Status (%)				
Private	64	84	88	<.01
Medicaid	3	2	1	
Medicare	8	4	6	
Other	18	9	5	
None	7	2	1	
Health Status (mean)				
Physical Health Index	50	51	50	.01
Mental Health Index	50	49	50	.20
Number of Comorbid Conditions	2	3	3	<.01

* Statistical significance was determined with chi-square (categorical variables) or ANOVA (continuous variables) depending on the variable.

Table 2. Average Satisfaction Scores by Respondent Characteristics

Characteristic	Summary Satisfaction Score*	p Value†
Age		
Less Than 60	49.1	<.01
60 or Older	51.6	
Gender		
Male	49.9	.16
Female	50.3	
Marital Status		
Married	50.3	.29
Other	49.9	
Education		
<HS	49.6	.58
HS	50.0	
>HS	50.1	
Income		
$20,000 or Less	50.2	.42
More Than $20,000	50.0	
Insurance Status		
Private	49.9	.02
Medicaid	49.3	
Medicare	51.6	
Other	50.5	
None	47.4	
Ethnicity/Interview Language		
Latino/Spanish	44.9	<.01
Latino/English	48.6	
Whites/English	50.3	

* Overall satisfaction scores based on equally weighted average of the five satisfaction with communication questions normalized to a mean of 50 and standard deviation of 10 (T-scores). Higher scores indicate greater patient satisfaction.

† Statistical significance was determined with ANOVA.

Table 3. Patient Ratings of Communication by Health Care Providers

How do you rate …	Adjusted Proportions*					p Value
	Very Poor/Poor	Fair	Good	Very Good	Excellent/ The Best	
Medical Staff Listening to What You Have to Say						
Latino/Spanish	9.1	19.7	28.3	22.5	20.4	<.01†
Latino/English	4.9	12.3	23.3	26.0	33.5	.12‡
White/English (reference group)	3.7	9.7	20.3	26.0	40.3	
Omnibus Test						<.01§
English-Spanish Latino Equivalence Test						<.01¶
Answers to Your Questions						
Latino/Spanish	6.7	19.9	30.2	23.2	20.0	<.01
Latino/English	3.6	12.4	24.9	26.8	32.3	.03
White/English (reference group)	2.7	9.7	21.6	27.0	39.0	
Omnibus Test						<.01
English-Spanish Latino Equivalence Test						<.01
Explanations About Prescribed Medications						
Latino/Spanish	10.3	20.2	29.6	20.2	19.7	<.01
Latino/English	5.6	13.0	25.3	24.0	32.1	.02
White/English (reference group)	4.1	9.9	21.8	24.3	39.9	
Omnibus Test						<.01
English-Spanish Latino Equivalence Test						<.01
Explanations About Medical Tests and Procedures						
Latino/Spanish	12.9	23.1	30.3	18.3	15.4	<.01

Latino/English	6.6	14.6	27.2	23.9	27.6	.13
White/English (reference group)	5.2	12.1	24.9	24.8	33.0	
Omnibus Test						<.01
English-Spanish Latino Equivalence Test						<.01
Reassurance and Support From Your Doctor and Support Staff						
Latino/Spanish	12.0	25.0	30.3	17.6	15.1	<.01
Latino/English	6.5	16.5	28.0	22.8	26.2	.05
White/English (reference group)	4.8	13.0	25.1	24.1	33.2	
Omnibus Test						<.01
English-Spanish Latino Equivalence Test						<.01

* Results from ordinal logistic model controlling for age, and gender (model 1). Standard errors adjusted for medical group membership.

† P value for Spanish/Latino coefficient (reference white/English).

‡ P value for English/Latino coefficient (reference white/English).

§ P value for adjusted Wald test of Spanish/Latino coefficient=0 and English/Latino coefficient=0 (omnibus test).

¶ P value for adjusted Wald test of Spanish/Latino coefficient = English/Latino coefficient.

Table 4. Unadjusted and Adjusted Ratings for Question, "Medical Staff Listening to What You Have to Say."

Language/Ethnic Response Group	Unadjusted Proportions*	Adjusted Proportions		
		Model 1†	Model 2‡	Model 3§
Latinos/Spanish				
Very Poor/Poor	8.7	9.1	9.1	11.0
Fair	19.0	19.7	18.9	20.8
Good	28.1	28.3	27.8	27.1
Very Good	22.9	22.5	23.4	23.0
Excellent	21.3	20.4	20.8	18.1
Latinos/English				
Very Poor/Poor	4.9	4.9	5.1	5.1
Fair	12.1	12.3	12.1	12.3
Good	23.2	23.3	22.5	21.5
Very Good	26.1	26.0	26.0	26.7
Excellent	33.8	33.5	34.3	34.4
Whites				
Very Poor/Poor	3.7	3.7	3.8	3.8
Fair	9.7	9.7	9.7	9.9
Good	20.3	20.3	20.0	19.1
Very Good	26.1	26.0	26.0	26.5
Excellent	40.2	40.3	40.5	40.7
Tests of Statistical Significance (p value)				
Latino/Spanish coefficient (white reference)	<.01	<.01	<.01	<.01
Latino/English coefficient (white reference)	<.01	.12	.88	.88
Omnibus Test (adjusted Wald test of Latino/Spanish coefficient=0 and	<.01	<.01	<.01	<.01

Latino/English coefficient=0)				
English-Spanish Latino Equivalence Test (adjusted Wald test of Latino/Spanish coefficient = Latino/English coefficient).	<.01	<.01	<.01	<.01

* Unadjusted ordinal logistic model. Standard errors adjusted for medical group membership.

† Ordinal logistic model controlling for age and gender (model 1).

‡ Ordinal logistic model controlling for age, gender and Spanish language response variable (SLRV) (model 2).

§ Ordinal logistic model controlling for age, gender, number of comorbid conditions, education, income, household size, insurance status and SLRV (model 3).

Appendix A. Checklist of Medical Conditions

Hypertension	Cancer
Myocardial Infarct	Migraines
Congestive Heart Failure	Cataracts
Stomach trouble	Deafness or trouble hearing
Limitation in use of leg or arm	Blurred Vision
Diabetes	Glaucoma
Angina	Macular Degeneration
Chronic Lung Disease	Liver Trouble
Chronic Allergies	Epilepsy
Seasonal Allergies	Sciatica or chronic back problems
Arthritis	Trouble seeing
Kidney problems	Thyroid problems
Dermatitis/other chronic skin rash	Males Only: Prostate Problems
	Females Only: Abnormal Vaginal Bleeding

5. Differences in CAHPS® Adult Survey Ratings and Reports by Race and Ethnicity: An Analysis of the National CAHPS® Benchmarking Data 1.0

Abstract

Objective: To examine racial/ethnic group differences in consumer reports and ratings of care using data from the National CAHPS® Benchmarking Database (NCBD) 1.0.

Data Sources: Adult data from the NCBD 1.0 is comprised of CAHPS® 1.0 survey data from 54 commercial and 31 Medicaid health plans from across the United States. A total of 28,354 adult respondents (age≥18 years) were included in this study. Respondents were categorized as belonging to one of the following racial/ethnic groups: Hispanic (n=1,657); non-Hispanic White (n=20,414); Black or African-American (n=2,942); Asian and Pacific Islander (n=976); American Indian and Alaskan Native (n=588); and Other racial/ethnic group or Multiracial (n=553). Persons who failed to indicate any racial/ethnic background were placed in a "Missing" category (n=1,224).

Study Design. Four single item global ratings (personal doctor; specialty care; overall rating of health plan; and overall rating of health care) and five multiple item report composites (access to needed care; provider communication; office staff helpfulness; promptness of care; and health plan customer service) from the CAHPS® 1.0 surveys were assessed.

Statistical Analyses. Multiple regression models were estimated to assess differences in global ratings and report composites between whites and members of other racial/ethnic groups, controlling for age, ender, perceived health status, educational attainment, and insurance type.

Principal Findings: Whites were more positive in reports about most aspects of care than members of other racial/ethnic groups. Inter-racial/ethnic group differences were diminished in global ratings of care.

Conclusions: Improvements in quality of care for racial/ethnic minority groups are needed. Comparisons of care in diverse populations based on global ratings of care should be interpreted cautiously.

Keywords: Reports and Ratings of Care, CAHPS®, Racial/Ethnic Differences, Patient Assessed Quality of Care.

Introduction

Dramatic changes are occurring in the racial/ethnic makeup of the United States. By 2050, the proportion of the US population who are white is projected to drop to below 50%, while the proportions of the US population who are Hispanic, Asian/Pacific Islander, and African-American are expected to exceed 25%, 13%, and 20%, respectively (Smith & Edmonston, 1997). In some states such as California, whites have already ceased to be the majority group (Johnson, 1999). For medical providers, these dramatic demographic changes pose the formidable challenge of providing effective and relevant healthcare to patients of many different racial/ethnic backgrounds.

Surveys that ask patients to assess their healthcare are important tools for determining how well doctors and other healthcare providers meet the needs of their patients. Properly constructed, these survey instruments can capture patients' experiences with care (reports) and evaluations of care (ratings) in a reliable and valid manner in multicultural settings. Furthermore, when analyzed according to the racial/ethnic background of the respondents, patient surveys can yield information about how well medical providers are meeting the needs of patients belonging to particular racial/ethnic subgroups. Although numerous studies have examined access to care for members of racial/ethnic subgroups, there have been relatively few studies of reports and ratings of care by members of racial/ethnic subgroups.

Most previous research on racial/ethnic differences in reports and ratings of care has focused on Hispanics and African-Americans. In a population-based study conducted in 1981, Anderson et al. (1981)

found more dissatisfaction with care among Hispanics compared to the US population. In more recent studies, Baker et al. (1996) reported greater dissatisfaction with ER provider communication among monolingual Spanish speaking patients than English speaking patients, and Morales et al. (1999) found greater dissatisfaction with primary care provider communication among Hispanics than whites.

Patient satisfaction research involving African Americans has yielded mix results. Bashshur et al. (1967) noted greater satisfaction with care among Blacks than whites in a HMO population. Subsequent studies, however, found that Blacks are more dissatisfied with care than whites. Hulka et al. (1975) reported more dissatisfaction with care among Blacks than whites in a sample community-based sample. Taira et al. (1997) also reported more dissatisfaction among Blacks than whites in a sample of patients from a university based primary care practice. Finally, Meredith and Sui (1995) reported more dissatisfaction with care among Blacks than whites in the Medical Outcomes Study (MOS).

Studies of satisfaction with care among Asians are few. Meredith and Sui (1995), using data from the MOS, reported greater dissatisfaction with care among Asians than whites, Blacks and Latinos. Taira et al. (1997) reported that Asians gave lower ratings of care than whites on multiple dimensions of care including communication, trust, interpersonal treatment and comprehensiveness of care provided. In a recent study of access and satisfaction with care conducted in physician group practices primarily on the west coast, Asians reported worse access to care and gave lower ratings of care than whites or

Latinos (Snyder et al., in press). We are unaware of any studies that have examined patient satisfaction among Asian/Pacific Islander sub-groups or patient satisfaction among American Indians/Alaskan Natives.

This study is based on data from the Consumer Assessment of Health Plans (CAHPS®) benchmarking database. The National CAHPS® 1.0 Benchmarking Database (NCBD 1.0) is the first nationwide aggregation of reports and ratings of care collected using the CAHPS® 1.0 survey instrument. In this study, we examine differences in CAHPS® reports and ratings of care among non-Hispanic Whites, Hispanics, Blacks, Asians/Pacific Islanders, American Indians/Alaskan Natives, and persons indicating membership in multiple race/ethnic groups or other race/ethnic groups. In addition, we classified separately those persons who did not indicate any race/ethnic category in the survey.

Based on our review of prior research, we hypothesized that Hispanics and Asian/Pacific Islanders would give worse reports and ratings of care than whites. Because the results of research comparing satisfaction with care between Blacks and whites are inconsistent, we hypothesized no difference in satisfaction with care between Blacks and whites. Due to the lack of published research on patient satisfaction among American Indians/Native Alaskans, or persons of multiple races/ethnicities, we did not formulate any specific hypotheses regarding these groups.

Methods

CAHPS® Survey Instruments

The CAHPS® surveys are currently the gold standard for assessing patient experiences with ambulatory care. CAHPS® has been adopted by Medicare (Schnaier et al., 1999), state Medicaid programs (Brown et al., 1999), the Office of Personnel Management, and the National Committee on Quality Assurance as part of its accreditation process (NCQA, 1998).

The CAHPS® survey instruments were developed by a consortium of investigators from RAND, Harvard Medical School, Research Triangle Institute, and Westat with funding from the Agency for Healthcare Research and Quality and the Healthcare Financing Administration. The principal goal was to produce survey instruments that could be used to collect reliable and valid information from health plan enrollees about the care they have received (Crofton et al., 1999). The data generated by the surveys is principally intended to provide information for consumers choosing among different health plans.

CAHPS® surveys have been developed for use in fee-for-service or managed care and in various types of settings including commercial, Medicare, and Medicaid. In addition, CAHPS® surveys are available for assessing adult and childcare and for administration via mail or telephone. English and Spanish versions of the CAHPS® surveys exist (Weidmer et al., 1999).

CAHPS® surveys ask a core set of questions that are applicable across settings (Hays et al., 1999). The core questions include reports and ratings. The reports items assess the frequency with which specific experiences took place (e.g., How often did doctors or other health professional listen carefully to you?). The ratings of care capture global perceptions of care (e.g., How would you rate all your health care?).

Data Source

The data for this study is the National CAHPS® Benchmarking Database 1.0 (NCBD 1.0). The NCBD 1.0 is comprised of CAHPS® 1.0 survey results collected by commercial and Medicaid health plans from across the United States. It includes the results of both adult and child CAHPS surveys, administered by telephone and mail, and in English and Spanish. Data limitations prevent us from identifying the mode or language that the survey was administered. Previous research, however, demonstrates the equivalence of CAHPS information collected via telephone and mail (Fowler et al., 1999).

The NCBD 1.0 contains survey results from 7 Medicaid sponsors comprising 31 health plans. The 31 Medicaid health plans consist of 29 health maintenance organizations and 2 primary care case management plans from the District of Columbia, Arkansas, Kansas, Minnesota, Oklahoma, and Washington. The mean response rate among the Medicaid health plans was 34% (Median = 37%) and ranged from 17% to 50%.

The NCBD 1.0 also contains survey results from 6 commercial sponsors comprising 54 health plans. The commercial health plans include 27 HMOs, 8 physician-provider organizations, 3 point-of-service, 1 fee-for-service, and 15 other unspecified types of health plans from the District of Columbia, Florida, Kansas, New Jersey, Oklahoma and Washington. The mean response rates among the commercial health plans was 63% (median = 54%) and ranged from 48% to 83%.

Together, the Medicaid (n = 8,813) and commercial databases (n = 19,541) contain 28,354 completed adult surveys. The mean response rate across the combined sample of commercial and Medicaid health plans is 52% (Median = 52%).

All participants contributed their data to NCBD 1.0 voluntarily. The purpose of this database is to facilitate comparisons among various users of CAHPS® surveys and to support research on consumer assessments of health care. The surveys were fielded in 1997 and 1998.

CAHPS® Measures

The dependent variables in this study are the four global rating questions (Personal MD, Specialists, Health Care, and Health Plan) and five multiple item reports (composites) (Access to Care, Promptness of Care, Doctor Communication, Office Staff, Health Plan Customer Service) derived from the adult core CAHPS® 1.0 survey (Table 1). The four rating questions are asked using a 0-10 response format, where 10 is the best possible rating. All of the questions included in the composites are asked using a *Never, Sometimes, Usually, Always* response

format except for the *Access to Needed Care* composite. This composite includes two questions asked using a *Yes, No* response format and two questions asked with a *Never, Sometimes, Usually, Always* response format. The composite summary scores were computed by first transforming linearly each individual item score to a 0-100 possible range, and then averaging individual item scores within each scale. To facilitate comparisons between rating and composite scores, the 0-10 rating scores were also transformed linearly to a 0-100 possible range.

The main independent variable in this study is race/ethnicity. The case-mix variables included in this study are age, gender, perceived health status, educational attainment, and insurance type. Patient gender is a binary variable (male, female). Seven indicator variables for age were constructed from the seven age categories (18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+) asked in the survey. Five indicator variables for perceived health status were constructed from the five categories (excellent [reference category], very good, good, fair, poor) asked in the survey. Six indicator variables for education were constructed from the six education categories (8^{th} grade or less; some HS, but did not complete; HS graduate or GED; 1-3 years of college; 4-year college graduate; more than 4-year college degree) asked in the survey. Insurance type is a binary variable indicating whether an observation belonged to a commercial or Medicaid health plan.

Respondents were assigned to racial/ethnic categories based on their answers to the following questions (both questions were asked of all respondents):

1. Are you of Hispanic or Spanish Family Background?

   Yes

   No

2. How Would You Describe Your Race?

   American Indian or Alaskan native

   Asian or Pacific Islander

   Black or African-American

   White

   Another Race or Multiracial (*write in*)_____

Survey respondents who answered "yes" to (1) were categorized as Hispanic, regardless of race (see Figure 1). Respondents who answered "no" to (1) were categorized according to their response to (2). Respondents that wrote in a response to (2) were placed in the category that most closely matching their stated race or ethnicity. For instance, if a respondent wrote in "Persian," they were placed in the Another Race/Multiracial category. Respondents who did not answer (1) or (2) or who answered "no" to (1) but did not indicate a race were also placed in the "Missing" category. Respondents who only answered (2) were assumed to be non-Hispanic.

Analysis Plan

Cross-tabulations of race/ethnicity with each of the other independent variables are provided for the sample. A chi-square

statistic was computed for each cross-tabulation to test the significance of the associations.

Unadjusted mean scores by racial/ethnic group were computed on each of the single item global ratings and the individual items comprising the multiple-item report composites. Intra-item racial/ethnic group mean differences were tested for statistical significance using one-way ANOVA models.

Cronbach's alpha was computed to assess internal consistency reliability of each multiple-item composite (Cronbach, 1951)

Multiple regression analyses were conducted to assess differences in global ratings and composites between whites and members of the other race-ethnic groups, controlling for age, gender, perceived health status, educational attainment, and insurance type. The CAHPS® developers recommend adjusting for age and health status when making comparisons between health plans (CAHPS® 2.0 Survey and Reporting Kit). We included gender (Like & Zyzanski, 1987; Weiss, 1988) and educational attainment (Fiscella, 1999; Ware et al., 1982; Fox & Storms, 1981) in our multivariate models because prior research has found significant associations between these variables and satisfaction with care. A separate ordinary least squares regression model was estimated for each global rating item and composite report. All regressions included the five case-mix adjustment variables.

Because the distribution of scores on both the global ratings and composites were negatively skewed (bunching at more positive end of the scale), each model was estimated using transformed as well as

untransformed dependent variables. The transformation we used - dividing the square of the variable by 100 - yielded approximately normal distributions. The regression results using the transformed and untransformed dependent variables were very similar. Thus, we only present results using the untransformed dependent variables.

To guard against finding statistically significant differences by chance alone, we examined the coefficients on all the race/ethnic indicator variables jointly (omnibus test) before testing the significance of individual coefficients on the race-ethnic indicator variables. Robust standard errors (correcting for intra-plan clustering) were estimated for all regression coefficients (Huber, 1964).

Because response rates varied across health plans, we derived non-response weights for this study. A weight proportional to the inverse of the response rate was computed for each plan (Brick and Kalton, 1996). Thus respondents belonging to a plan with a low response rate received a greater weight than respondents belonging to a plan with a higher response rate, and all respondents within the same plan received the same weight.

To interpret the magnitude of the differences in reports and ratings between the race/ethnic groups in this study, we estimated differences in CAHPS® reports and ratings in terms of intentions to change medical providers. Because intention to change providers was not measured on the same subjects as the CAHPS® measures, a two-stage procedure was necessary. Specifically, we regressed intentions on a

rating of access to care in the 1994-95 United Medical Group Study (UMGA) dataset (Hays et al., 1998) and used the regression coefficient to estimate the corresponding relationship between a CAHPS® report composite and intentions. We also regressed intentions on a rating of quality of care in the UMGA dataset and used the regression coefficient to estimate the corresponding relationship between the CAHPS global rating of care and intentions. These estimates provided a rough basis for interpreting the clinical meaningfulness of observed differences between race/ethnic groups.

All analysis was conducted using STATA, version 6.0 (StataCorp., 1999).

Results

Sample Description (Table 2)

There were significant differences ($p<0.01$), as assessed by chi-square tests, in the distributions of all case-mix variables across the race/ethnic groups. Specifically, the distributions of age, gender, perceived health status, educational attainment, and type of insurance differed across Hispanics, whites, Blacks, Asian/Pacific Islanders, American Indian/Alaskan Natives, and persons in the Other/Multiracial and Missing race/ethnicity categories.

Item Means and Standard Errors by Racial/Ethnic Group (Table 3)

Table 3 presents mean item scores by racial/ethnic group. This table provides the reader with the data to "drill down" below the level of composite scores. For example, one can compare mean scores on

individual items that constitute the *Access to Care* composite for Asian/Pacific Islanders - the Asian/Pacific Islander mean scores on *Ease of Finding* an MD and *Ease of Getting Approvals and Payments* were 81.34 (SE=1.46) and 49.49 (SE=2.06), which is significant at a $p<0.01$ level based on a two sample t-test with equal variances (Kanji, 1993). Similar pair wise comparisons can be made between other pairs of item mean scores. All items differed significantly by racial/ethnic group at the $p<0.01$ level.

Scale Internal Consistency Reliability

Internal consistency reliability estimates for the five multiple item composites ranged from moderate to high (0.62 to 0.89). Specifically, the alpha for *Access to Care* was 0.62; the alpha for *Promptness of Care* was 0.68; the alpha for *Doctor Communication* was 0.89; the alpha for *Office Staff* was 0.67; and the alpha for *Health Plan Customer Service* was 0.63.

Multivariate Results (Table 3)

With a few exceptions, whites gave better reports and higher ratings of care than non-whites, though there were fewer significant differences between whites and non-whites on ratings than reports. No group gave better reports about care than whites on the *Access to Care*, the *Promptness of Care*, or the *Health Plan Customer Service* composites, and only Blacks gave higher reports on the *Provider Communication* and the *Office Staff Helpfulness* composites than whites. Hispanics gave worse reports than whites on the *Access to Care*, *Promptness of Care* and *Health Plan Customer Service* composites. American Indian/Alaska Natives gave reports about care to similar whites.

The greatest number of significantly worse reports (compared to whites) was obtained from Asians/Pacific Islanders, persons in the multiple/other race category, and persons who did not report their race/ethnic information (race-missing). Asians/Pacific Islanders and persons in the race-missing category reported significantly worse care than whites on all composite reports. Persons in the multiple/other race category reported significantly worse care than whites in on all composites except *Health Plan Customer Service*.

There were fewer significant differences in global rating of care scores between whites and non-whites. Interestingly, there were no significant differences in global ratings between whites and Asians/Pacific Islanders. Persons in the race-missing category rated the *Health Care* and *Health Plan* global rating questions lower than whites. Hispanics rated the *Health Plan* rating question higher than whites and Blacks rated the *Health Plan* and *Health Care* rating questions higher than whites. American Indian/Alaskan Natives rated both *Personal Physician* and *Specialist Physician* rating questions lower than whites. Persons in the other race/multiracial category rated the *Personal Physician*, *Health Care*, and *Health Plan* rating questions lower than whites.

In summary, Hispanics gave worse reports about care than whites on three report composites, but rated care higher than whites on one global rating question. Blacks gave better reports about care than whites on two report composites, and rated care higher than whites on two global questions. Asians/Pacific Islanders gave worse reports about care than whites on all five report composites, but rated their care similarly to whites on all the global rating questions. American

Indians/Alaskan Natives gave reports about care similar to whites, but rated care lower than whites on two global rating questions. Persons in the other race/multiracial category reported worse care on four report composites and rated care lower than whites on three global ratings questions. Persons in the missing-race category reported worse care than whites on all report composites and rated care worse than whites on two rating questions.

Effect of Differences in Satisfaction on Intent to Change Medical Groups (Table 5) *SEE PAGE 149*

Based on differences in the access to needed care composite, we estimated that the odds ratio of intent to change providers was 1.11 for Hispanics, 1.27 for Asians/Pacific Islanders, 1.58 for persons in the multiracial/other race category, and 1.47 for persons in the missing race category relative to whites. Intent to change provider odds ratios were estimated and reported only for beta coefficients that were significant in the regression analyses.

Discussion

Differences in Reports and Ratings of Care

In this study we found significant differences in reports and ratings of care between whites and members of other racial/ethnic groups. In general, whites reported more positive experiences with care and rated their care more highly and non-whites.

Consistent with prior research and our initial hypotheses, Hispanics were less positive about their care than whites. Specifically, Hispanics reported worse promptness in receiving care and worse health plans customer service than whites. These results suggest that Hispanics may be at risk for worse outcomes care than whites due to delays in obtain needed care. In addition, poor health plan customer service can discourage proper utilization of the health care system by failing to communicate important information about accessing care. Appropriate customer service staff for Hispanic populations should include bilingual customer service agents, who are more difficult to recruit and can demand higher wages than mono-lingual English employees. Further, written materials need to be translated to Spanish and of the appropriate readability level for the target population's literacy level.

Contrary to our initial hypotheses, but consistent with prior research, Blacks rated their health care more highly and reported more positive experiences with care than whites. Specifically, Blacks reported significantly more positive experiences with physician communication and physician office staff, and rated their health care and health plan better than whites. These results indicate that on average, Black patients are more likely to report that doctors are taking as enough time to listen and explain things than white patients are. Furthermore, these results suggest that doctor's office staffs are providing courtesy or respect to Black patients.

Asian/Pacific Islanders had worse reports about care than whites across all domains of care probed by the CAHPS® surveys. These results

indicate that Asian/Pacific Islanders are on average, having more difficulty than whites finding a personal doctor, obtaining needed treatments, obtaining referrals to see specialists, and obtaining acute and routine care visits. They also indicate that Asian/Pacific Islanders are less likely than whites to have a doctor listen to them carefully, explain their treatments and diagnosis clearly, and spend enough time with them. Taken at face value, these results indicate that Asian/Pacific Islanders are experiencing significantly lower quality of care than whites and are at risk for poorer outcomes of care.

American Indian/Native Alaskans rated their personal doctors and specialists lower than whites, but did not differ significantly from whites on any of the reports about care. These generally positive results are surprising in light of the poor access to care and quality of care documented for American Indians and Native Alaskans in other studies (Cunningham et al., 1995). They suggest that the subgroup of American Indians/Alaskan Natives in our study sample is not representative of American Indians/Alaskan Natives in general. Excluded from our sample are American Indians living on reservations, who may be at greatest risk for low quality of care. Our study results suggest, however, that those American Indians and Alaskan Natives in our sample are relatively pleased with their care.

Compared to whites, persons who indicated multiple races or a race other than white, Black, American Indian/Alaskan Native, or Asian/Pacific Islander reported worse experiences with care and rated their healthcare lower. Little is know about persons who indicate multiple races. The sample description (Table 2) suggests that they

are predominantly middle-aged, educated beyond high school, and from the commercial sector. Although this group constituted only 2% of the NCBD 1.0 sample, more research is needed to better characterize this population in light of the 2000 census, which allows respondents to indicate multiple races (www.census.gov).

Persons in the missing race category reported worse experiences with care and rated their healthcare lower than whites. Little is known about persons who fail to give a racial/ethnic background. In the NCBD 1.0, persons not indicating a racial/ethnic group constituted 4% of the sample, tended to be middle-aged and educated beyond high school. Because this group of survey respondents had such negative perceptions about their care, it is possible that skipping questions is a form of protest against their health plan.

Intentions to Change Providers

One important consequence of lower reports and ratings of care is changing providers (Schlesinger et al., 1999; Allen and Rodgers, 1997; Newcomer et al., 1996). Based on differences in the access to care composite, we estimated that Hispanics (Odds Ratio [OR] = 1.11) Asian/Pacific Islanders (OR = 1.27), persons in the multiracial/other race category (OR = 1.58), and persons in the missing race category (OR = 1.47) were more likely to change providers than whites. Assuming that 14% whites intend to change providers, the odds ratio for Hispanics indicates that 15% of Hispanics intend to change providers representing a 9% difference in intentions to change providers between Hispanics and whites. Similar analyses indicate that 17% Asian/Pacific Islanders intend to change providers representing a 22% difference in

intentions between Asian/Pacific Islanders and whites; that 20% of persons in the multiracial/other race category intend to change providers representing a 46% difference between this group and whites; and that 19% of persons in the missing race category intend to change providers representing a 38% difference between persons in this group and whites.

Differences in the quality of care global rating item tell a similar story. Once again, assuming that 14% of white intend to change providers, we predict that 25% persons in the missing race category (OR = 1.54) and 26% of persons in the missing race category (OR = 1.57) intend to change providers. These differences respectively represent a 43% and 45% greater intent to change providers compared with whites, assuming that 14% of whites report the intent to change providers.

While not all persons expressing the intent to change providers actually follow through, a significant number do. Research shows that intent to change medical providers is an important signal that a health plan may not be providing adequate medical treatment (Schlesinger et al., 1999). Thus, the differences in the CAHPS® access to care report composite and global quality of care rating item that we observe by racial/ethnic group in this study, suggest that Asians/Pacific Islanders, persons in the multiracial/other category, and persons in the missing category, in particular, are at higher risk than whites to receive low quality care.

Inconsistencies between Report Composites and Global Rating Items

The inconsistency of the global ratings and composite scores in some instances are not entirely unexpected (Pasco, 1983). On the one hand, global ratings assess satisfaction by measuring beliefs about overall care, thus making few explicit assumptions about the specific domains of care relevant to respondents making these evaluations. On the other hand, composites assess satisfaction by measuring the frequency with which specific experiences take place, explicitly defining the domains of care used by respondents making these evaluations. Thus, if members of different racial/ethnic make their evaluations of care using non-identical domains of care or using similar domains of care but assigning different relative importance to them, then we would expect to find inconsistent relationships between global ratings and composite scores.

The most striking discrepancy between evaluations obtained using the global ratings and composites is among Asian/Pacific Islanders. Based on global ratings, Asian/Pacific Islanders and whites received similar care. Based on the report composites, however, Asians/Pacific Islanders are receiving significantly lower quality care than whites. Prior research has documented lower satisfaction with care among Asian/Pacific Islanders than whites. Although cultural reasons for more dissatisfaction among Asian/Pacific Islanders have been cited, our analysis documents both worse reports about care than whites and similar ratings of care between Asian/Pacific Islanders and whites. If a cultural bias in responding to survey assessment questions underlies the poor reports about care we observe among Asian/Pacific Islanders, why does it not also result in lower ratings? Without further research on item functioning among whites and Asian/Pacific Islanders, it is

impossible to determine whether the worse reports about care among Asian/Pacific Islanders than whites are attributable to bias or worse experiences with care.

Overall it appears that when patients are asked to provide global ratings, differences between whites and other racial/ethnic groups are diminished, and when patients are asked to provide specific reports about care experiences (composites), differences between whites and other race-ethnic groups are amplified.

Limitations

The overall mean response rate across both commercial and Medicaid health plans was 52% (median = 52%). To determine whether this response rate significantly biased our results, we would need to know whether the response rates varied by racial/ethnic group and whether reports of care and ratings of respondents differed from that of non-respondents for each racial/ethnic group. These data would allow us to assess the representativeness of our sample for each of the racial/ethnic groups in our analyses and the direction and magnitude of any non-response bias. Without these data we cannot know, for example, how representative Blacks in our study are of all Blacks in the surveyed health plans and whether the experiences of Blacks responding to the survey are similar to those not responding to the survey. Unfortunately, data limitations prevent us from estimating racial/ethnic group specific response rates. Nor can we estimate differences in reports about care and ratings between respondents and non-respondents by racial/ethnic group.

Another limitation of our data source is that it does not allow us to disaggregate Hispanics or Asian/Pacific Islanders by cultural/linguistic group, thus potentially obscuring important variation at the subgroup level. Information on each respondent's English proficiency, level of acculturation, national origin or immigration/refugee status may provide important insights into how race/ethnicity. For example, respondents with limited English proficiency (LEP) may have much lower levels of satisfaction with provider communication than respondents with high levels of English language proficiency (Morales et al., 1999; Baker et al., 1997), but because LEP respondents represent a small fraction of all respondents, these findings are obscured in the current analyses.

Previous studies have raised additional methodological concerns in surveying ethnically and linguistically diverse populations (Hayes et al., 1999; Marin et al., 1992; Ross and Mirowski, 1984; Warnecke et al., 1997; Pearl and Fairley, 1985; Ware, 1978; Bachman and O'Malley, 1984; Hui and Triandis, 1989). These concerns focus on differences in the measurement properties of survey instruments across racial/ethnic groups. For instance, investigators have reported that African Americans and Hispanics tend to choose the extremes of Likert response scales (extreme response tendency) more often than whites or to answer affirmatively to yes/no questions (acquiescent response tendency), regardless of question content or phrasing (Marin et al., 1992; Bachman and O'Malley, 1984; Hui and Triandis, 1989). The discrepancies we have observed between global and composite measures support these concerns.

Recent research has addressed, at least in part, the issue of measurement equivalence in diverse populations. Morales et al. (in

press) used item response theory methods to evaluate the equivalence of a composite measure of satisfaction with care administered to a sample of non-Hispanic white and Hispanic survey respondents. This research showed that in spite of small differences in item functioning in some items, valid comparisons between these groups using a multiple item scale is possible. Similar studies of item and scale functioning need to be performed on samples that include Asians/Pacific Islanders, Blacks, and American Indians/Alaskan Natives.

Conclusions and Policy Implications

Race and ethnicity continue to play an important role in American society. This study documents significant differences in reports and ratings of care between whites and non-whites. These racial and ethnic differences may put nonwhites at increased risk for low quality of care and poor outcomes of care. Until there is more complete public accountability of health plans and physician groups for the care they deliver to their patients, it is unlikely that the particular needs of their diverse and vulnerable patients will be addressed adequately.

Healthcare providers and purchasers of care need to collect information about the race and ethnicity of the individuals they serve. Without this critical information, representative sampling of non-whites will remain unverifiable. Available methods for assessing the equivalence of survey instruments across race and ethnic groups should be applied to widely used survey instruments such as CAHPS. Without these studies, doubt will continue to shroud the validity of racial/ethnic differences in reports and ratings of care. Quality

improvement strategies designed to improve the relevance to and effectiveness of healthcare in diverse patient populations are needed.

References

Allen, H. M., and W.H. Rogers. 1997. "The Consumer Health Plan Value Survey: Round Two." Health Affairs 16: 156-166.

Anderson, R., S. Zelman-Lewis, A.L. Giachello, L.A. Aday, and G. Chu. 1981. "Access to Medical Care Among the Hispanic Population of the Southwestern United States." Journal of Health and Social Behavior 22, (March): 78-89.

Bachman, J.G., and P.M. O'Malley. 1984. "Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles." Public Opinion Quarterly 48: 491-509.

Baker, D.W., R.M. Parker, M.V. Williams, W.C. Coates, and K. Pitkin. 1996. "Use and Effectiveness of Interpreters in an Emergency Department." Journal of the American Medical Association 275: 783-8.

Bashshur, R.L., C.A. Metzner, and C. Worden. 1967. "Consumer Satisfaction With Group Practice, the CHA Case." American Journal of Public Health and the Nations Health 57, no.11 (Nov.):1991-9.

Brick, J.M., and G. Kalton. 1996. "Handling Missing Data in Survey Research." Statistical Methods in Medical Research 5: 215-238.

Brown, J.A., S.E. Nederend, R.D. Hays, P.F. Short, and D.O. Farley. 1999. "Special Issues in Assessing Care of Medicaid Recipients." Medical Care 37, no.3: MS79-MS88.

CAHPS® 2.0 Survey and Reporting Kit. Agency for Healthcare Quality and Research, U.S. Department of Health and Human Services, AHRQ Publication No: 99-0039, October 1999.

Chronbach, L.J. 1951. "Coefficient alpha and the internal structure of tests." Psychometrika 16, 297-334.

Crofton, C., J.S. Luliban, and C. Darby. 1999. "Foreword." Medical Care 37: MS1-MS9.

Cunningham, P.J., and L.J. Cornelius. 1995. "Access to Ambulatory Care for American Indians and Alaskan Natives; the Relative Importance of Personal and Community Resources." Social Science and Medicine 40, no.3: 393-407.

Fiscella, K., and P. Franks. 1999. "Influence of Patient Education on Profiles of Physician Practices." Annals of Internal Medicine 131, no.10 (November): 745-51.

Fowler, F.J., P.M. Gallagher, and S.E. Nederend. 1999. "Comparing Telephone and Mail Responses to the CAHPS® Survey Instruments." Medical Care 37, no.3: MS41-MS49

Fox, J.G., and D.M. Storms. 1981. "A Different Approach to Sociodemographic Predictors of Satisfaction to Health Care." Social Science and Medicine 15(A): 557-64.

Hays, R.D., J.A. Brown, K.L. Spritzer, W.J. Dixon, and R.H. Brook. 1998. "Member Ratings of Health Care Provided by 48 Physician Groups." Archives of Internal Medicine 158 (April): 785-90.

Hays, R.D., J.A., Shaul, V.S.L. Williams, J.S. Lubalin, L.D. Harris-Kojetin, S.F. Sweeny, and P.D. Cleary. 1999. "Psychometric Properties of the CAHPS 1.0 Survey Measures." Medical Care 37, no. 3: MS22-MS31.

Hayes, R.P., and D.W. Baker. 1998. "Methodological Problems in Comparing English-Speaking and Spanish-Speaking Patients'

Satisfaction with Interpersonal Aspects of Care." Medical Care 36: 230-6.

Huber, P.J. 1964. "Robust Estimation of a Location Parameter." Annals of Mathematical Statistics 35: 73-101.

Hui, H., and H.C. Triandis. 1989. "Effects of Culture and Response Format on Extreme Response Style." Journal of Cross-Cultural Psychology 20, no.3 (September): 296-309.

Hulka, B.S., L.L.Kupper, M.B.Daly, J.C.Cassel, F. Schoen. 1975. "Correlates of Satisfaction with Medical Care: A Community Perspective." Medical Care 13, no.8: 648-658.

Johnson, H.P. 1999. "How Many Californians? A Review of Population Projections for the State. California Counts, Population Trends and Profiles." Public Policy Institute of California 1, no.1 (October).

Kanji, G.K. 1993. *100 Statistical Tests*. Thousand Oaks, CA: Sage Publications.

Like, R., and S.J. Zyzanski. 1987. "Patient Satisfaction with the Clinical Encounter: Social Psychological Determinants." Social Science and Medicine 24, no.4: 351-57.

Marin, G., R.J. Gamba, and B.V. Marin. 1992. "Extreme Response Style and Acquiescence Among Hispanics—The Role of Acculturation and Education." Journal of Cross-Cultural Psychology 23, no.4 (December): 498-509.

Meredith, L.S., and A.L. Sui. 1995. "Variation and Quality of Self-Report Health Data. Asians and Pacific Islanders Compared with Other Ethnic Groups." Medical Care 33, no.11: 1120-31.

Morales, L.S., S.P. Reise, and R.D. Hays. "Evaluating the Equivalence of Health Care Ratings by Whites and Hispanics." In Press: Medical Care, May, 2000.

Morales, L.S., W.E. Cunningham, J.A. Brown, H. Liu, and R.D. Hays. 1999. "Are Latinos Less Satisfied with Communications by Health Care Providers?" Journal of General Internal Medicine 14: 409-17.

National Committee for Quality Assurance. 1998. Accreditation '99: Standards for the Accreditation of Managed Care Organizations, pages 11-17.

Newcomer, R., S. Preston, and C. Harrington. 1996. "Health Plan Satisfaction and Risk of Disenrollment Among Social/HMO and Fee-For-Service Recipients." Inquiry 33: 144-54.

Pasco, G.C. 1983. "Patient Satisfaction in Primary Health Care: a Literature Review and Analysis." Evaluation and Program Planning 6: 185-210.

Pearl, D.K., and D. Fairley. 1985. "Testing for the Potential for Nonresponse Bias in Sample Surveys." Public Opinion Quarterly 49: 553-60.

Ross, C.E, and J. Mirowski. 1984. "Socially-Desirable Response and Acquiescence in a Cross-Cultural Survey of Mental Health." Journal of Health and Social Behavior 25 (June): 189-97.

Schlesinger, M., B. Druss, and T. Thomas. 1999. "No Exit? The Effect of Health Status on Dissatisfaction and Disenrollment for Health Plans." Health Services Research 34: 547-76.

Schnaier, J.A., S.F. Sweeny, V.S.L. Williams, B. Kosiak, J.S. Lubalin, R.D. Hays, and L.D. Harris-Kojetin. 1999. "Special Issues Addressed in the CAHPS Survey of Medicare Managed Care Beneficiaries." Medical Care 37, no.3: MS69-MS78.

Smith, J.P., and B. Edmonston. 1997. *The New Americans: Economic, Demographic, and Fiscal Effects of Immigration*. Washington, DC: National Academy Press.

Snyder, R., W. Cunningham, T.T. Nakazono, and R.D. Hays. "Access to Medical Care Reported by Asians and Pacific Islanders in a West Coast Physician Group Association." In Press.

StataCorp. 1999. Stata Statistical Software: Release 6.0. College Station, TX: Stata Corporation.

Taira, D.A., D.G. Safran, T.B. Seto, W.H. Rogers, M. Kosinski, J.E. Ware, N. Lieberman, and A.R. Tarlov. 1997. "Asian-American Patient Ratings of Physician Primary Care Performance." Journal of General Internal Medicine 12 (April): 237-242.

Warnecke, R.B., T.P. Johnson, N. Chavez, S. Sudman, D.P. O'Rourke, L. Lacey, and J. Horm. 1997. "Improving Question Wording in Surveys of Culturally Diverse Populations." Annals of Epidemiology 7, no. 5 (July): 334-42.

Ware, J.E. 1978. "Effects of Acquiescent Response Set on Patient Satisfaction Ratings." Medical Care 16, no. 4 (April): 327-36.

Ware, J.E., A. Davies-Avery, and A.L. Stewart. 1982. "The Measurement and Meaning of Patient Satisfaction." Health and Medical Care Services Review 1, no. 1: 2-28.

Weidmer, B., J. Brown, and L. Garcia. 1999. "Translating the CAHPS 1.0 Survey Instrument into Spanish." Medical Care 37: MS89-MS97.

Weiss, G.L. 1988. "Patient Satisfaction with Primary Care: Evaluation of Sociodemographic and Predispositional Factors." Medical Care 26, no.4: 383-92.

Table 1. CAHPS® 1.0 questions from the 1.0 NCBD database.

Single Item Global Ratings	**Response Format**
Health Plan	
We want to know your rating of all your experience with your health insurance plan. How would your health plan?	0 - 10 Scale.
Health Care	
We want to know your rating of all your health care in the last 6 months from all doctors and other health professionals. How would you rate all your health care?	0 - 10 Scale.
Specialists	
We want to know your rating of the specialist you saw most often in the last 6 months. How would you rate the specialist?	0 - 10 Scale.
Personal MD	
We want to know your rating of your personal doctor or nurse. How would you rate your personal doctor or nurse?	0 - 10 Scale.
Multiple Item Composites	
Access to Care	
With the choices that your health plan gives you, was it easy to find a personal doctor or nurse for yourself?	Yes (1) No (0)
In the last 6 months, was it easy to get a referral when you needed one?	
In the last 6 months, how often did you the tests or treatment you thought were needed?	Never (1) Sometimes (2) Usually (3) Always (4)
In the last 6 months, how often did your health plan deal with approvals or payments without taking a lot of your time and energy?	
MD Communication	
In the last 6 months, how often did doctors or other health professional listen carefully to you?	Never (1) Sometimes (2) Usually (3) Always (4)
In the last 6 months, how often did doctors or other health professionals explain things in a way you could understand?	
In the last 6 months, how often did doctors or other health professionals show respect for what you had to say?	
In the last 6 months, how often did doctors or other health professionals spend enough time with you?	

MD Staff	
In the last 6 months, how often did office staff at a doctor's office or clinic treat you with courtesy and respect?	Never (1) Sometimes (2) Usually (3) Always (4)
In the last 6 months, how often did office staff at a doctor's office or clinic as helpful as you thought they should be?	
Promptness of Care	
In the last 6 months, how often did you get the medical help you needed when you phoned the doctor's office or clinic during the day Monday to Friday?	Never (1) Sometimes (2) Usually (3) Always (4)
In the last 6 months, when you tried to be seen for an illness or injury, how often did you see a doctor or other health professional as soon as you wanted?	
In the last 6 months, when you needed regular or routine care, how often did you get an appointment as soon as you wanted?	
In the last 6 months, how often did you wait in the doctor's office or clinic more than 30 minutes past your appointment time to see the person you went to see?	
Health Plan Customer Service	
In the last 6 months, how often did you get all the information or other help you needed when you called the health insurance plan's customer service?	Never (1) Sometimes (2) Usually (3) Always (4)
In the last 6 months, how often were people at the health insurance plan's customer service as helpful as you thought they should be?	
In the last 6 months, how often did you have more forms to fill out for your health insurance plan than you thought was reasonable?	

Table 2. Sample Characteristics.

	Hispanic	White	Black	Asian / PI	AI / NA	Other / Multi	Missing
N	1,657	20,414	2,942	976	588	553	1,224
Age (%)							
18-34	51	37	48	41	49	43	41
35-54	42	74	43	48	45	49	50
55+	7	12	9	11	7	9	9
Gender (%)							
Female	75	72	82	64	83	68	75
Male	25	28	17	36	17	33	25
Education (%)							
<HS	24	10	20	14	20	7	10
HS	30	30	38	23	36	27	22
>HS	45	61	42	64	44	66	69
Health Status (%)							
E	16	16	16	18	12	17	16
VG	29	36	27	33	23	32	34
G	34	34	32	37	37	35	32
F	17	12	20	11	21	14	14
P	4	3	5	2	8	2	4
Insurance Type (%)							
Commercial	48	62	30	64	30	66	52
Medicaid	53	38	70	36	70	35	48

Note. All P-values statistically significant at 0.05 level.

Table 3. Item means and standard errors by race and ethnicity.

	Hispanic (n=1,657)	White (n=20,414)	Black (n=2,942)	API (n=976)	AI/NA (n=588)	M/O (n=533)	Miss (n=1,224)
Global Ratings							
Personal MD	79.72 (0.72)	79.35 (0.17)	77.25 (0.61)	77.65 (0.82)	72.06 (1.53)	75.18 (1.15)	77.63 (0.73)
Specialists	77.63 (1.16)	77.01 (0.30)	70.36 (1.04)	77.03 (1.35)	63.17 (2.53)	70.92 (1.98)	73.93 (1.17)
Health Care	79.45 (0.70)	78.83 (0.18)	80.39 (0.52)	77.84 (0.85)	75.50 (1.25)	72.95 (1.25)	73.33 (0.83)
Health Plan	77.43 (0.57)	73.43 (0.16)	77.60 (0.44)	74.58 (0.69)	73.89 (1.03)	68.72 (1.07)	67.74 (0.72)
Access to Care							
Finding personal MD	80.98 (1.06)	80.58 (0.31)	82.24 (0.77)	81.34 (1.46)	71.76 (2.03)	71.47 (2.19)	70.71 (1.51)
Get referrals	66.51 (2.07)	70.17 (0.55)	68.66 (1.59)	74.65 (2.71)	65.20 (3.50)	65.80 (3.30)	60.63 (2.36)
Getting tests and treatments	71.70 (1.39)	77.17 (0.32)	69.76 (1.11)	69.59 (1.73)	71.17 (2.15)	65.66 (2.35)	69.41 (1.44)
Getting approvals and payments	54.47 (1.50)	66.71 (0.35)	50.72 (1.18)	49.49 (2.06)	58.53 (2.41)	58.92 (2.24)	62.03 (1.83)
MD Communication							
MD listens carefully	81.25 (0.80)	80.58 (0.21)	83.27 (0.60)	79.83 (1.04)	80.41 (1.29)	76.92 (1.45)	76.30 (0.96)
MD explained things well	81.64 (0.83)	83.76 (0.20)	84.65 (0.59)	79.58 (1.14)	83.11 (1.28)	81.71 (1.25)	80.91 (0.87)
MD respected your comments	82.80 (0.77)	81.83 (0.21)	84.61 (0.59)	78.23 (1.11)	81.93 (1.27)	77.22 (1.40)	75.89 (0.97)
MD spend	72.69	75.29	75.49	71.77	73.38	70.04	69.82

enough time	(0.91)	(0.23)	(0.70)	(1.23)	(1.45)	(1.54)	(1.03)

MD Staff							
MD staff courteous and respectful	85.66 (0.74)	87.81 (0.18)	88.09 (0.52)	81.96 (1.10)	84.73 (1.17)	84.61 (1.20)	84.39 (0.83)
MD Staff Helpful	78.78 (0.82)	79.09 (0.21)	80.79 (0.61)	75.70 (1.13)	77.05 (1.33)	75.31 (1.45)	73.97 (0.93)
Promptness of Care							
Getting phone help	74.91 (1.01)	78.87 (0.25)	77.40 (0.73)	68.71 (1.41)	73.68 (1.73)	73.22 (1.69)	74.41 (1.18)
Getting acute care	68.25 (1.17)	71.78 (0.30)	69.20 (0.91)	63.80 (1.57)	66.68 (1.87)	65.79 (2.02)	66.90 (1.39)
Getting routine care	68.78 (1.09)	71.93 (0.29)	71.70 (0.78)	66.25 (1.40)	71.12 (1.80)	68.81 (1.89)	66.09 (1.33)
Wait more than 30 minutes	66.16 (1.08)	70.17 (0.28)	68.25 (0.77)	68.44 (1.41)	67.31 (1.57)	65.46 (1.88)	54.66 (1.31)
Health Plan Customer Service							
Getting Information from Customer Service	65.54 (1.40)	67.48 (0.38)	71.91 (1.06)	67.27 (1.63)	69.28 (2.30)	65.82 (2.18)	61.49 (1.60)
Customer Service Helpful	67.45 (1.40)	69.56 (0.37)	73.22 (1.04)	69.04 (1.63)	73.14 (2.06)	68.50 (2.10)	62.01 (1.52)
Paperwork	85.35 (0.67)	87.28 (0.20)	87.50 (0.47)	84.64 (0.92)	86.83 (1.10)	87.68 (1.16)	57.93 (1.32)

Note. All intra-item racial/ethnic group mean differences significant at a $p<0.01$ level, as assessed by one-way ANOVA models for each item.

Table 4. Results of Ordinary Least Squares Multivariable Regressions.

	Reports About Care					Ratings of Care			
	Access	Prompt	Comm	Helpful	Service	MD Rate	Spec Rate	Care Rate	Plan Rate
Hispanic	-2.02*	-3.91*	0.32	-0.57	-2.41*	1.14	2.49	1.24	3.35*
	(0.94)	(1.32)	(0.85)	(0.85)	(1.14)	(1.00)	(2.01)	(1.04)	(1.04)
Black	-1.65	-1.03	2.65*	1.90*	-1.02	-0.86	-2.50	2.52*	3.24*
	(1.03)	(0.79)	(0.55)	(0.39)	(1.14)	(1.65)	(1.48)	(0.91)	(0.74)
Asian/Pacific	-4.56*	-8.27*	-5.21*	-8.18*	-3.23*	-1.49	-0.78	-1.23	1.36
Islander	(1.23)	(1.17)	(1.15)	(1.50)	(1.59)	(0.96)	(1.89)	(0.91)	(0.82)
American Indian	-6.14	-2.55	-0.73	-2.04	-0.16	-4.91*	-8.59*	-1.68	0.43
/Alaskan Native	(4.09)	(1.57)	(1.12)	(1.29)	(0.98)	(2.44)	(3.82)	(2.59)	(1.32)
Other Race/	-8.77*	-5.35*	-4.84*	-4.36*	-1.15	-3.98*	-3.61	-5.25*	-3.86*
Multiracial	(1.59)	(1.62)	(1.21)	(1.05)	(1.24)	(1.38)	(2.54)	(1.57)	(1.35)
Missing	-7.42*	-8.53*	-4.70*	-4.90*	-25.03*	-0.54	-1.04	-5.41*	-5.58*
	(1.61)	(1.95)	(1.12)	(1.06)	(8.04)	(1.53)	(1.85)	(1.36)	(0.98)
Sample Mean Score	75.13	72.29	80.16	83.06	82.17	78.77	75.81	78.64	73.89
Joint F-Test	12.12	13.93	12.58	13.30	2.93	3.40	3.22	9.17	19.47
Regression R^2	0.04	0.04	0.06	0.04	0.04	0.03	0.09	0.07	0.05

Note. Regression models included age, gender, education, health status, and an indicator for Medicaid. Robust standard errors (in parenthesis) are adjusted for intra-plan clustering of observations. White was the reference group in all regressions. Reports about care scores range from 0-100 (100=best) and ratings of care ranged from 0-10 (10=best).

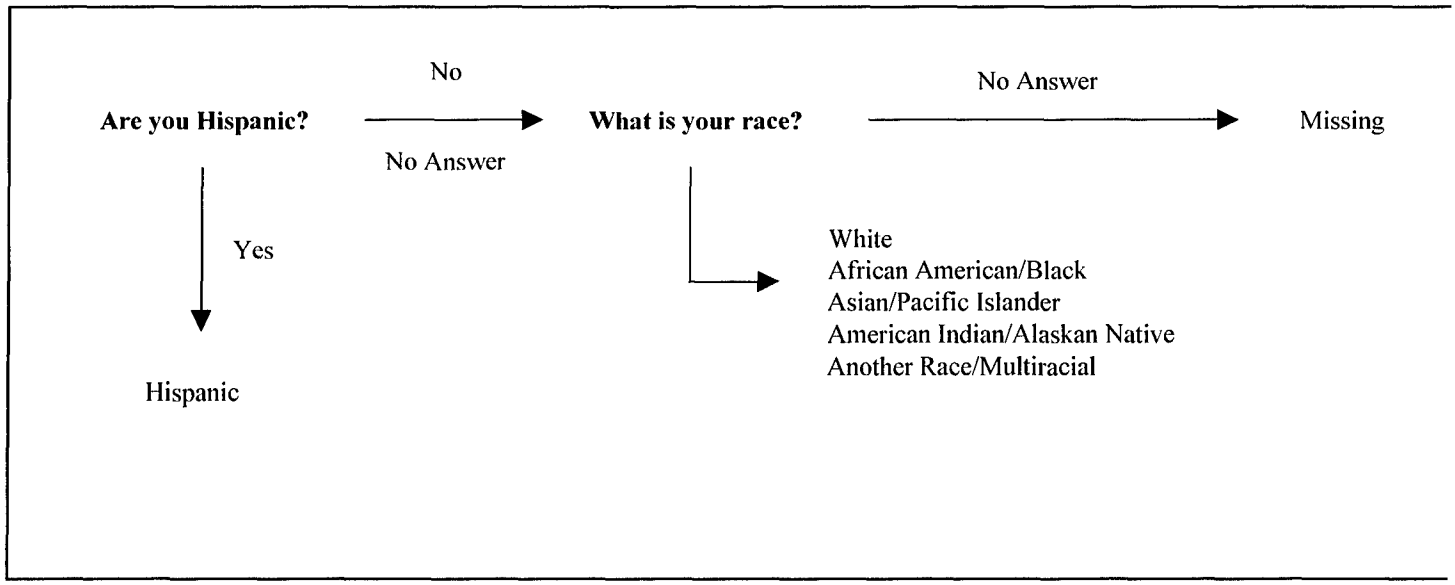
*Denotes statistical significance at 0.05 level.

Table 5. Effect of dissatisfaction on intent to change medical groups

	Access to Needed Care Composite		Global Rating of Health Care	
	$BETA_{CAHPS}$	Odds Ratio	$BETA_{CAHPS}$	Odds Ratio
Hispanic	-2.02	1.11	NS	-
Black	NS	-	2.52	0.81
Asian / Pacific Islander	-4.56	1.27	NS	-
American Indian / Alaskan Native	NS	-	NS	-
Other Race / Multiracial	-8.85	1.58	-5.25	1.54
Missing	-7.42	1.47	-5.41	1.57

Note. Whites are the reference group for the odds ratio. NS indicates "Not Significant."

Figure 1. Racial/ethnicity assignment algorithm.



6. RACIAL AND ETHNIC DIFFERENCES IN PARENTS' ASSESSMENTS OF PEDIATRIC CARE IN MEDICAID MANAGED CARE

Abstract

Objective. This study examines whether parents' reports and ratings of pediatric health care vary by race/ethnicity and language in Medicaid managed care.

Data Sources. The data analyzed are from the National CAHPS® Benchmarking Database (NCBD 1.0) and consist of 9,870 children Medicaid cases receiving care from 33 health plans in Arkansas, Kansas, Minnesota, Oklahoma, Vermont, and Washington from 1997 to 1998.

Study Design. Data are analyzed using regression methods. The dependent variables are CAHPS® 1.0 ratings (personal doctor, specialist, health care, health plan) and reports of care (getting care needed, timeliness of care, provider communication, staff helpfulness, plan service). The independent variables are race/ethnicity (White, Black, American Indian, Asian, and Hispanic), Hispanic language (English or Spanish), and Asian language (English or Other), controlling for gender, age, education, and health status.

Data Collection. The data was collected by telephone and mail, and surveys were administered in Spanish and English. The average response rate for all plans was 42.1%.

Principal Findings. Racial/ethnic minorities had worse reports of care than whites. Among Hispanics and Asians, language barriers had a larger negative impact on reports of care than race/ethnicity. However,

lower reports of care for racial/ethnic groups did not translate necessarily into lower ratings of care.

Conclusions. Health plans need to pay increased attention to racial/ethnic differences in assessments of care. This will facilitate the development of culturally and linguistically appropriate health care services for racial/ethnic minorities.

Key Words. Race/ethnicity, consumer assessments of health plans (CAHPS), reports and ratings of care, patient satisfaction.

Background and Significance

Consumer assessments of health care are increasingly being used as an indicator of the quality of care provided by health plans and providers. In 1999, the results of the Consumer Assessments of Health Plans (CAHPS) were made available to 90 million of Americans enrolled in Medicare and private health plans (Cleary, 1999). These evaluations provide important information about how well health plans and clinicians meet the needs of the people they serve (Crofton, Lubalin, & Darby, 1999). Patient evaluations of care have been associated with utilization (Zastowny, Roghmann, & Cafferata, 1989) and compliance with medical regimens (Hall & Dornan, 1990; David and Rhee, 1998). In addition, dissatisfaction with care has been linked with doctor shopping and disenrollment from health plans (Rossiter et al., 1989; Newcomer, Preston, and Harrington, 1996; Kerr et al., 1998).

The extent to which consumer assessments of health care vary by race/ethnicity is of significance in evaluations of health plan performance. Health care organizations (HCOs) will increasingly face a more diverse patient population. As of 1999, 28% of the U.S. population was member of a racial or ethnic minority group, and it is projected that by 2030, 40% of the U.S. population will be members of a racial or ethnic minority group (U.S. Census Bureau, 1999). Cultural differences across groups can serve as a communication barrier and result in less satisfaction with health care. In a review of the literature on access to care in Hispanic communities, Valdez and colleagues (1993) found that poor patient-provider communication and lack of cultural competency were significant barriers to high quality care.

The study of racial/ethnic differences in consumer assessments of care is particularly important for Medicaid managed care populations. Increasingly, government is relying on the managed care sector to provide coverage for Medicaid and Medicare populations as a cost-containment mechanism (Halstead & Becherer, 1998). As of 1997, 32.1 million people, or 47% of Medicaid recipients were enrolled in managed care plans.

As more vulnerable populations are enrolled in managed care plans it becomes essential to develop assessments of their care. Indigent populations may have more difficulties in dealing with the complexities of managed care organizations (Sisk et al., 1996). African Americans and Hispanics are disproportionately represented in the low-income groups. Furthermore, managed care may not have the necessary competencies to manage a linguistically and culturally diverse population. Indeed, managed care oftentimes restructure services moving patient populations away from community-based, traditional providers that are more familiar with the culture and language of racial/ethnic groups (Leigh et al., 1999). There may also be a clash between the belief systems and attitudes of cultural subgroups and those of the managed care culture, since managed care values and principles focus on a healthy and young population (Lavizzo-Mourey & Mackenzie, 1996). A recent study by Phillips and colleagues (1999) shows that racial/ethnic minorities enrolled in managed care plans generally report greater continuity of care but less satisfaction with the care received. Similarly, Leigh et al. (1999) in a study comparing managed care enrollees with fee for service (FFS) enrollees in Florida, Tennessee, and Texas found that African Americans enrolled in managed care plans were more likely to report problems with access to care than African

Americans in FFS, and Hispanic managed care enrollees were more likely to be dissatisfied with the provider/patient relationship than Hispanics in FFS.

This study examines whether parent's reports and ratings of pediatric health care vary by race/ethnicity in Medicaid managed care. In addition, the study assesses whether reports and ratings of care vary by primary language for Hispanic and Asian parents.

Race/Ethnicity and Consumer Assessments of Health Care

Relatively few studies have examined the impact of race/ethnicity on satisfaction with health care. A meta-analysis of the relationship of patient sociodemographic characteristics and patients' satisfaction with medical care, found no significant relationship between patient's race/ethnicity and satisfaction (Hall and Dornan, 1990). However, more recent studies have shown differences in satisfaction across racial and ethnic groups. Meredith and Siu (1995) and Taira et al. (1997) found Asians less satisfied than other racial/ethnic groups with the care received in outpatient settings. Similarly, in a study by Snyder and colleagues (in press), Asian and Pacific Islanders reported worse access to care than other racial/ethnic groups. Harpole et al. (1996) examined patient dissatisfaction with outpatient clinics, and found African Americans more dissatisfied with the timeliness of care than Whites or Hispanics. Gross et al. (1998) reported that nonwhites were less satisfied than Whites with the amount of time spent with their family physician. Finally, a study by Morales et al. (1999) of medical group practices found Hispanics less satisfied than Whites with provider communication.

A study using the 1996 Medical Expenditure Panel Survey (MEPS) showed that racial/ethnic minorities face greater barriers to care than Whites, especially Hispanics and Asians (Phillips, Mayer, & Aday, 1999). Hispanics were twice as likely as other groups to have long waits and to perceive their provider failed to listen and provide the needed information. Asians reported more difficulties in getting appointments, dissatisfaction with the care received, and lack of confidence in the provider's abilities.

Language has been documented as a barrier to care among racial/ethnic minorities, especially for Spanish-speaking Hispanics. Hu and Covell (1986) found that outpatients whose primary language was English were more satisfied with their care in general than were patients whose primary language was Spanish. Carrasquillo et al. (1999) examined patient satisfaction with emergency departments (EDs) at five urban teaching hospitals, and showed that non-English-speaking patients were less satisfied than English-speaking patients with the care provided by the ED and less likely to visit the same ED if they needed care in the future. Non-English speakers were particularly dissatisfied with the overall care, courtesy and respect, and discharge instructions. Furthermore, language barriers have been found to affect satisfaction with care beyond cultural barriers. Studies contrasting Spanish-speaking and English-speaking Hispanics have also found Spanish-speakers to be less satisfied with the care received and with provider communication (David and Rhee, 1998; Morales et al., 1999).

Methodology

The research questions investigated in this study are:

Do parents' reports and ratings of pediatric care vary by race/ethnicity in Medicaid managed care?

Do parents' reports and ratings of pediatric care vary by primary language for Hispanics and Asians?

Data

The Consumer Assessment of Health Plans Study (CAHPS®) was established by the Agency for Healthcare Research and Quality (AHRQ) in 1995 through cooperative agreements with consortia headed by Harvard Medical School, RAND, and the Research Triangle Institute (RTI). The primary purpose of CAHPS® was to produce a set of standardized surveys and report templates that would yield comparative information about the experiences of enrollees with their health plan and health care providers. CAHPS® 1.0 was developed and tested from 1995 to 1997. The CAHPS® Child Survey includes questions about the issues covered by the adult survey and some additional issues pertinent to children's care (Shaul et al., 1999).

This study analyzes the National CAHPS® Benchmarking Database (NCBD) CAHPS® 1.0 Child Surveys. NCBD is a collaborative initiative of the Quality Measurement Advisory Service (QMAS), The Picker Institute, and Westat. Sponsors of the CAHPS® surveys voluntarily participate in the NCBD and include Medicaid agencies, health plans, and employers. The purpose of this database is to facilitate comparisons among various users of CAHPS® surveys. In this study we analyze only Medicaid Child cases, since the NCBD CAHPS 1.0 only contains Medicaid sponsors. The Medicaid Child sponsors database included 33 health plans from Arkansas, Kansas, Minnesota, Oklahoma, Vermont and Washington. The data was collected by telephone and mail, and surveys were administered in

Spanish and English. Previous research has shown the equivalence of the telephone and mail responses to the CAHPS survey (Fowler, Gallagher, and Nederend, 1999). Limitations in the NCBD CAHPS 1.0 data did not allow us to identify surveys administered either in English or Spanish. The average response rate for all plans was 42.1%, with a median of 42.4%, and a range of 30.1% to 57.1%. The field period covers 1997-1998. The original data consist of 9,870 children (<18 years of age) Medicaid cases (4,972 males, 4,662 females, and 236 missing gender information).

Measures

The dependent variables consist of CAHPS® global ratings and reports of care. Ratings consist of the personal evaluation of providers and services; as such they reflect both personal experiences as well as the standards used in evaluating care (Davies and Ware, 1988). Reports of care capture the specific experiences with care in terms of what did or did not happen from the consumer's perspective. The survey uses a fixed time interval of the past six months in framing the questions on the experiences with health care.

CAHPS® 1.0 includes four global rating items administered using a 0-10 scale: personal doctor or nurse, specialists, health care, and health plan. In addition, it contains 17 items (reports) measuring 5 domains of health plan performance, or composite reports: getting needed care (access), timeliness of care, provider communication, staff helpfulness, and plan service (Table 1). The composite reports are calculated in a two-step process: adding the items within a scale, and then linearly transforming the total to a 0-100 scale. Internal consistency reliability for each of the five scales for the composite

reports of care was estimated using Cronbach (1951) alpha coefficients: Getting needed care (access) (_ = 0.60); Timeliness of care (_ = 0.73); Provider Communication (_ = 0.82); Staff helpfulness (_ = 0.77); and Plan service (_ = 0.67). To facilitate comparison between the composite reports and the global ratings, the 0-10 ratings were linearly transformed to a 0-100 scale.

The independent variables consist of parent's race/ethnicity, Hispanic and Asian parents' language, and case-mix adjustors. Race/ethnicity constitutes a categorical variable representing the racial or Hispanic ethnicity of the respondent:

- Hispanic/Latino
- Black/African American
- Asian/Pacific Islander
- Native American/Alaskan native
- White
- Other Race/Ethnicity
- Missing Race/Ethnicity

Only Non-Hispanics are coded into a racial group. Respondents who did not indicate a race/ethnicity are placed in the "Missing" category.

Hispanic and Asian parents are further classified based on the language he or she primarily speaks at home:

- Hispanic English-speaking

- Hispanic Spanish-speaking
- Asian English-speaking
- Asian Other Language
- Asian Missing Language

Hispanics missing language information were dropped from the analysis, since they consisted of only 26 cases.

An additional set of variables is used as case-mix adjustors: parent's gender, parent's age, parent's education, and child health status. These are patient characteristics known to be related to systematic biases in survey responses (Aharony & Strasser, 1993; Cleary & McNeil, 1988; Elliot et al., 2000). Parent's gender is a dichotomous variable: *0 = female, 1= male.* Parent's age is a categorical variable consisting of three categories: *18-34; 35-54; 55 or older.* Parent's education is a categorical variable with three categories: *less than high school, high school graduate, and 1 or more years of college.* Child's health status is a categorical variable measuring how parents rate their child's overall health: *excellent, very good, good, fair, and poor.*

Analysis Plan

Bivariate statistics (chi square and one-way ANOVA) were used to examine differences in age, gender, education, health status, and CAHPS® ratings/reports of care among the racial/ethnic subgroups. Ordinary least squares regression was used to model the effect of race/ethnicity, Hispanic language, and Asian language on CAHPS® ratings and reports, controlling for parent's age, parent's gender, parent's

education, and child's health status. Standard errors for all regressions were adjusted for correlation within health plans using the Huber/White correction (White, 1980).

A small departure from normality was detected for the dependent variables (negatively skewed). As a result, the variables were transformed by dividing the square of the variable by 100, to produce an approximately normal distribution. However, given similar regression results for both the transformed and untransformed dependent variables, only the results for the untransformed variables are reported here.

Given that response rates varied across health plans, non-response weights were computed. A weight proportional to the inverse of the response rate was computed for each plan (Brick and Kalton, 1996). As a result, respondents belonging to a plan with a low response rate received a greater weight than respondents belonging to a plan with a higher response rate, and all respondents within the same plan received the same weight.

Results

Bivariate statistics indicate that there are significant differences across racial/ethnic groups in terms of case-mix adjustors (Table 2) and CAHPS® ratings/reports (Table 3).

Regression results for reports of care (Table 4) show that compared to Whites:

- Getting needed care reports were more negative for Hispanic Spanish, Asian Other, Black, American Indian, and Missing Race than for other racial/ethnic groups.

- Timeliness of care reports were more negative for Hispanic Spanish, Asian Other, Asian Missing, Black, American Indian, Other Race, and Missing Race than for other racial/ethnic groups.

- Reports of provider communication were lower for Hispanic Spanish, Asian Other, Asian Missing, American Indian, and Missing Race than for other racial/ethnic groups.

- Reports of staff helpfulness were lower for Hispanic Spanish, Asian English, Asian Other, and Asian Missing than for other racial/ethnic groups.

- Plan service reports were more negative for Hispanic Spanish, Asian Other, Black, American Indian, and Missing Race than for other racial/ethnic groups.

- Except for getting care needed, Asian Other reports of care were the lowest of all subgroups.

However, regression results for ratings of care show less variation in global ratings of care (Table 5). Compared to Whites:

- Personal doctor rating was more positive for Hispanic Spanish, but more negative for American Indian than for other racial/ethnic groups.

- Specialist rating was more positive for Hispanic Spanish and Asian English than for other racial/ethnic groups.

- Health care rating was more negative for Asian Other than for other racial/ethnic groups.

- Health plan rating was more positive for Hispanic Spanish and Asian English, but more negative for American Indian and Missing than for other racial/ethnic groups.

Conclusions

Despite significant advances in medical care in recent decades, racial and ethnic disparities in health status and quality of care still persist. One of the reasons for these continued disparities is the inadequate access to care for minorities (Andrulis, 1998). This study has examined the impact of race/ethnicity, Hispanic language, and Asian language on parents' reports and ratings of pediatric care in Medicaid managed care. Reports of care captured experiences with getting care needed, timeliness of care, provider communication, staff helpfulness, and plan service. Ratings of care included evaluations of personal doctor, specialist, health care, and health plan. This study suggests that racial and ethnic minorities still face barriers to health care. In addition, the study documents that language is an important barrier to care for Hispanics and Asians, perhaps more important than race/ethnicity.

The report for getting needed care evaluates access to medical services, such as specialists and recommended treatments. Compared to whites, Hispanic Spanish, Asian Other, Blacks, and American Indians scored lower than whites on getting needed health care. Similarly, racial/ethnic minorities fared more poorly in other dimensions of access, such as timeliness of care and health plan service. Hispanic Spanish, Asian Other, Black, and American Indian reported lower scores for timeliness of care and plan service than whites. The results of this study also suggest that racial and ethnic minorities face problems

with respect to provider communication and staff helpfulness. Compared to whites, Hispanic Spanish, Asian Other, and American Indian reported lower scores for provider communication, while Hispanic Spanish, Asian English, and Asian Other reported lower scores for staff helpfulness.

Language barriers account for a large degree of the negative effect of race/ethnicity on reports and ratings of care for Hispanics. While Hispanic Spanish had lower scores than whites for all reports of care, Hispanic English speakers did not differ significantly from whites on any of the reports of care. These findings are consistent with previous research on Hispanics showing that Spanish-speakers are less satisfied with care than English-speakers (David and Rhee, 1998; Morales et al., 1999).

Asian non-English speakers had the lowest reports of care of all racial/ethnic groups. Furthermore, an examination of the beta coefficients indicates that the negative impact of language on Asian reports of care was comparable to that of poor health status for four of the reports (timeliness of care, provider communication, staff helpfulness, and plan service) (Table 4). This finding is consistent with previous research showing that Asians have lower satisfaction with care than other racial/ethnic groups (Meredith and Siu, 1995; Taira et al., 1997). However, these studies did not account for the language effect. In our study after controlling for language differences, Asian English speakers did not differ significantly from whites on four of the reports of care. This indicates that language barriers may account for a large degree of the observed negative impact of race/ethnicity on reports and ratings of care among Asians.

The lower scores on the reports of care of racial/ethnic minorities did not translate necessarily into lower ratings of care. Compared to whites, American Indians had lower ratings for personal doctor and health plan, while Asian Other had lower ratings for health care. On the other hand, Hispanic Spanish had higher ratings than whites for personal doctor, specialist, and health plan. Previous research has shown that Spanish-speaking Hispanics have a bias towards more favorable responses in patient satisfaction surveys (Hayes & Baker, 1998).

Further research is needed to examine why the lower scores on the reports of care do not necessarily translate into lower ratings of care among racial/ethnic groups. A possible explanation is that reports of care are more objective and may capture real differences in care, whereas ratings miss the differences because they are more subjective and influenced by expectations, and racial/ethnic minorities may have lower expectations. Expectations are beliefs and attitudes with respect to the medical encounter that are shaped by previous experience with care, culture, social class, and health status (Kravitz, 1996; Handler et al., 1998).

This study has important policy implications. With the increased diversity in the population, health plans need to pay increased attention to assessments of care from racial and ethnic minorities. Consumer surveys are increasingly being used as a tool in quality assessment and improvement (Cleary, 1999). Satisfaction assessment can provide data for quality improvement efforts, such as total quality management (TQM) (Halstead & Becherer, 1998). It will become more important in quality improvement efforts to examine subpopulation

differences among ethnic groups, so that quality improvement initiatives can be more focused and efficient (Taira et al., 1997).

Furthermore, this study suggests the importance for health care organizations to provide culturally and linguistically competent health care services. Understanding the determinants of positive health care experiences in different racial and ethnic groups will facilitate the development of more culturally appropriate health care services. This will include acquiring knowledge of the health-related beliefs, attitudes, and communication patterns of the different racial/ethnic groups to improve services and programs (HRSA, 1999). Linguistically appropriate services should include bilingual providers and competent interpreter services. Previous studies have shown the importance of patient-provider language concordance for adherence to medical regimens (Manson, 1988) and patient outcomes (Perez-Stable, Naapoles-Springer, & Miramontes, 1997).

Acknowledgements

The authors wish to thank Dale Shaller of the Quality Measurement Advisory Service (QMAS) for his assistance in obtaining the data for this study.

References

Andrulis, D.P. 1998. "Access to Care Is the Centerpiece in the Elimination of Socioeconomic Disparities in Health." *Annals of Internal Medicine*, 129, 412-416.

Aharony, L. and S. Strasser. 1993. "Patient satisfaction: What we know about and what we still need to explore." *Medical Care Review*, 50(1), 49-79.

Brick, J. M. and G. Kalton. 1996. "Handling missing data in survey research." *Statistical Methods in Medical Research*, 5, 215-238.

Carrasquillo, O., E. J. Orav, T. A. Brennan, and H. Burstin. 1999. "Impact of language barriers on patient satisfaction in an emergency department." *Journal of General Internal Medicine*, 14, 82-87.

Cleary, P. D. 1999. "The increasing importance of patient surveys." *British Medical Journal*, 319, 720-1.

Cleary, P. D., and B. J. McNeil. 1988. "Patient satisfaction as an indicator of quality care." *Inquiry*, 25(1), 25-36.

Crofton, C., J. S. Lubalin, and C. Darby. 1999. Foreword. *Medical Care*, 37(3), MS1-MS9.

Cronbach, L. J. 1951. "Coefficient alpha and the internal structure of tests." *Psychometrika*, 16: 297.

David, R. A., and M. Rhee. 1998. "The impact of language as a barrier to effective health care in an underserved urban Hispanic

community." *The Mount Sinai Journal of Medicine*, 65(5-6), 393-397.

Davies, A. R. and J. E. Ware. 1988. "Involving consumers in quality of care assessment." *Health Affairs*, Spring, 33-48.

Elliott, M.N., R. Swartz, J. Adams, and R.D. Hays. 2000. "Casemix Adjustment of the National CAHPS Benchmarking Data 1.0". Paper presented at the Quality from the Consumer Perspective: Research Findings Conference, Columbia, Maryland.

Fowler, F. J., P. M. Gallagher, and S. Nederend. 1999. "Comparing telephone and mail responses to the CAHPS™ survey instrument." *Medical Care*, 37(3): MS41-MS49.

Gross, D. A., S. J. Zyzanski, E. A. Borawski, R. D. Cebul, and K. C. Stange. 1998. "Patient satisfaction with time spent with their physician." *The Journal of Family Practice*, 47(2), 133-137.

Hall, J. A., and M. C. Dornan. 1990. "Patient sociodemographic characteristics as predictors of satisfaction with medical care: A meta-analysis." *Social Science and Medicine*, 30(7), 811-818.

Halstead, D., and R. C. Becherer. 1998. "Surveying patient satisfaction amongdisadvantaged managed care customers." *Health Care Strategic Management*, 16(11), 15-19.

Handler, A., D. Rosenberg, K. Raube, and M. A. Kelley, M. A. 1998. "Health care characteristics associated with women's satisfaction with prenatal care." *Medical Care*, 36(5), 679-694.

Harpole, L. H., E. J. Orav, M. Hickey, K. E. Posther, and T. A. Brennan. 1996. "Patient satisfaction in the ambulatory setting." *Journal of General Internal Medicine,* 11, 431-434.

Hayes, R. P., and D. W. Baker. 1998. "Methodological problems in comparing English-speaking and Spanish-speaking patients' satisfaction with interpersonal aspects of care." *Medical Care,* 36(2), 230-236.

Health Resources and Services Administration (HRSA). 1999. *Cultural Competence: A Journey.* Bethesda, MD: HRSA.

Hu, D. J., and R. M. Covell. 1986. "Health care usage by Hispanic outpatients as a function of primary language." *West Journal of Medicine,* 144(4), 490-493.

Kerr, E. A., R. D. Hays, M. L. Lee, and A. L. Siu. 1998. "Does dissatisfaction with access to specialists affect the desire to leave a managed care plan?" *Medical Research and Review,* 55(1), 59-77.

Kravitz, R.L. 1996. "Patients' expectations for medical care: An expanded formulation based on review of the literature." *Medical Care Research and Review,* 53(1), 3-27.

Lavizzo-Mourey, R., and E. R. Mackenzie. 1996. "Cultural competence: Essential measurements of quality for managed care organizations." *Annals of Internal Medicine,* 124(10), 919-920.

Leigh, W. A., M. Lillie-Blanton, R. M. Martinez, and Karen S. Collins. 1999. "Managed care in three states: Experiences of low-income African Americans and Hispanics." *Inquiry,* 36, 318

331.

Manson, A. 1988. "Language concordance as a determinant of patient compliance and emergency room use in patients with asthma." *Medical Care,* 26, 1119-1128.

Meredith, L. S., and A. L. Siu. 1995. "Variation and quality of self-report health data: Asians and Pacific Islanders compared with other ethnic groups." *Medical Care,* 33(11), 1120-1131.

Morales, L. S., W. E. Cunningham, J. A. Brown, H. Liu, and R. D. Hays. 1999. "Are Latinos less satisfied with communication by health care providers?" *Journal of General Internal Medicine,* 14,409-417.

Newcomer, R., S. Preston, and C. Harrington. 1996. "Health plan satisfaction and risk of disenrollment among social/HMO and fee-for-service recipients." *Inquiry*, 33, 144-154.

Perez-Stable, E. J., A. Naapoles, and J. M. Miramontes. 1997. "The effects of ethnicity and language on medical outcomes of patients with hypertension or diabetes." *Medical Care,* 35(12), 1212-1219.

Phillips, K. A., M. Mayer, and L. A. Aday. 1999. "Access to care for racial/ethnic groups under managed care." Presentation at the Association for Health Services Research Annual Meeting, Chicago, IL.

Rossiter, L.F., K. Langwell, T. T. H. Wan, and M. Rivnyak. 1989. "Patient satisfaction among elderly enrollees and disenrollees in Medicare health maintenance organizations." *Journal of the American Medical Association*, 262(1): 57-63.

Shaul, J.A., F. J. Fowler, A. M. Zaslavsky, C. J. Homer, P. M. Gallagher, and P.D. Cleary. 1999. "The impact of having parents report about both their own and their children's experiences with health insurance plans." *Medical Care,* 37(3): MS59-MS68.

Sisk, J. E., S. A. Gorman, A. L. Reisinger, S. A. Glied, W. H. DuMouchel, and M. M. Hynes. 1996. "Evaluation of Medicaid managed care." *Journal of the American Medical Association,* 276(1), 50-55.

Snyder, R., W. Cunningham, T. T. Nakazano, and R. D. Hays. In press. "Access to medical care reported by Asians and Pacific Islanders in a West Coast Physician Group Association." *Medical Care Research and Review.*

Taira, D. A., D. G. Safran, T. B. Seto, W. H. Rogers, M. Kosinski, J. E. Ware, N. Lieberman, and A. R. Tarlov. 1997. "Asian-American patient ratings of physician primary care performance." *Journal of General Internal Medicine,* 12, 237-242.

U.S. Census Bureau, Population Estimates Program, Population Division. 1999. *Resident population estimates of the United States by sex, race, and Hispanic origin: April 1, 1990 to November 1, 1999.* Washington, DC: U.S. Census Bureau.

Valdez, R. B., A. Giachello, H. Rodriguez-Trias, P. Gomez, and C. De La Rocha. 1993.

"Improving access to health care in Latino communities." *Public Health Reports,* 108(5), 534-539.

White, H. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica*, 48, 817-830.

Zastowny, T. R., K. J. Roghmann, and G. L. Cafferata. 1998. "Patient satisfaction and the use of health services." *Medical Care*, 27(7), 705-723.

Table 1

CAHPS® 1.0 Child Global Ratings and Reports of Care

Ratings/Composite measure	Survey Items	Response scale
Personal doctor or nurse rating	How would you rate your child's personal doctor or nurse now? (C8)	0-10 Scale
Specialist rating	How would you rate your child's specialist? (C13)	0-10 Scale
Health care rating	How would you rate all your child's health care? (C42)	0-10 Scale
Health plan rating	How would you rate your child's health insurance plan now? (C64)	0-10 Scale
Getting needed care (composite): assess access to care	1) Was it easy to find a personal doctor or nurse for your child you are happy with? (C3) 2) Was it always easy to get a referral when your child needed one? (C12)	Yes (1) No (0)
	1) How often did your child get the tests or treatment you thought your child needed? (C41) 2) How often did your child's health insurance plan deal with approvals or payments without taking a lot of your time and energy? (C59)	1 Never 2 Sometimes 3 Usually 4 Always
Timeliness of care (composite): assess getting care promptly	How often did you get the medical help or advice you needed for your child when you phoned the doctor's office or clinic during the day Monday to Friday? (C15) 1) When you tried to have your child seen for an illness or injury, how often did you see a doctor or other health professional as soon as you wanted? (C18) 2) When your child needed regular or routine health care, how often did your child get an appointment as soon as you wanted? (C20)	1 Never 2 Sometimes 3 Usually 4 Always

	3) How often did your child wait in the doctor's office or clinic more than 30 minutes past the appointment to see the person you went to see? (C24)	
Provider communication (composite): assess communication of provider with patients	1) How often did your child's doctors or other health professionals listen carefully to you? (C27) 2) How often did your child's doctors or other health professionals explain things in a way you could understand? (C29) 3) How often did your child's doctors or other health professionals show respect for what you had to say? (C30) 4) How often did your child's doctors or other health professionals spend enough time with your child? (C36)	1 Never 2 Sometimes 3 Usually 4 Always
Staff helpfulness (composite): whether the staff treats the customer with courtesy and respect	1) How often did office staff at your child's doctor's office or clinic treat you and your child with courtesy and respect? (C25) 2) How often were office staff at your child's doctor's office or clinic as helpful as you thought they should be? (C26)	1 Never 2 Sometimes 3 Usually 4 Always
Plan service (composite): assess calls to customer service	1) How often did you have more forms to fill out for your child's health insurance plan than you thought was reasonable? (C57) 2) How often did you get all the information or other help you needed when you called the health insurance plan's customer service? (C62) 3) How often were the people at the health insurance plan's customer service as helpful as you thought they should be? (C63)	1 Never 2 Sometimes 3 Usually 4 Always

Table 2
Case-Mix Adjustors by Race/Ethnicity

	Hispanic -English	Hispanic -Spanish	Black	Asian-English	Asian-Other	Asian Missing	Amer. Indian	White	Other	Missing	Chi Square	P-value*
N	395	471	1,413	95	123	82	339	6,509	114	301		
Parent's Age (%)												
18-34	64.2	65.7	56.8	64.2	67.8	42.3	51.8	54.9	50.9	53.5	99.08	0.00
35-54	32.2	32.5	35.1	31.6	29.6	55.1	38.9	40.0	39.5	42.5		
55+	3.6	1.7	8.1	4.2	2.6	2.6	9.3	5.1	9.7	4.0		
Parent's Gender (%)												
Female	94.9	81.5	94.3	80.9	56.7	65.0	92.0	92.1	84.1	89.2	350.33	0.00
Male	5.1	18.5	5.7	19.2	43.3	35.0	8.0	7.9	15.9	10.8		
Parent's Education (%)												
<High School	29.3	66.7	24.6	9.7	38.8	29.1	29.3	16.3	16.8	17.7	750.94	0.00
High school	40.3	22.6	41.9	46.2	30.2	21.5	34.3	42.8	32.7	35.0		
College	30.4	10.7	33.5	44.1	31.0	49.4	36.4	40.9	50.4	42.3		
Child's Health Status												
Excellent	43.7	33.3	35.8	50.0	30.6	27.9	38.7	42.7	38.1	46.5	212.63	0.00
Very good	31.0	29.2	32.7	26.6	32.2	45.6	33.3	34.8	26.6	33.2		
Good	18.2	23.0	21.3	17.0	29.8	22.8	18.8	18.2	25.7	16.2		
Fair	5.9	13.9	9.6	5.3	7.4	3.8	8.6	4.0	8.9	3.7		
Poor	1.3	0.6	0.6	1.1	0.0	0.0	0.6	0.4	0.9	0.4		

Table 3

CAHPS® 1.0 Global Ratings/Reports of Care by Race/Ethnicity/Language

	Hispanic -English	Hispanic -Spanish	Black	Asian-English	Asian-other	Asian Missing	Amer. Indian	White	Other	Missing	F-value	P-value*
Personal doctor or nurse rating	87.04	88.49	86.14	84.29	80.00	81.06	82.39	86.36	85.33	83.44	3.11	.001
Specialist rating	81.92	87.07	76.86	92.14	68.42	70.00	70.00	80.08	80.81	78.87	2.78	.003
Health care rating	85.94	84.87	83.98	83.24	76.54	75.80	82.49	85.08	82.02	82.65	4.10	.000
Health plan rating	82.17	89.16	81.43	85.33	79.37	79.36	77.76	82.04	76.04	76.52	10.25	.000
Satisfying health care needs	80.70	75.08	78.04	83.84	70.71	84.80	72.84	83.04	76.00	75.85	11.47	.000
Timeliness of care	74.64	64.94	73.36	72.55	57.09	59.36	74.75	79.32	73.27	71.55	33.41	.000
Provider Communication	83.56	77.61	83.96	80.11	65.06	69.81	80.95	85.05	84.62	80.30	16.78	.000
Staff helpfulness	84.11	77.81	84.74	80.39	63.46	66.67	84.87	86.73	86.53	85.96	20.37	.000
Plan service	86.92	78.48	83.01	83.92	75.30	85.83	83.95	88.35	81.81	82.43	19.78	.000

* Statistical significance determined by using one-way ANOVA.

Table 4

Regression Results for Reports of Care by Race/Ethnicity and Language

	Getting Care Needed	Timeliness of Care	Provider Comm.	Staff Helpfulness	Plan Service
Parent's Race/Ethnicity[1]					
Hispanic English	-2.18	-3.86	-.93	-2.17	-1.21
Hispanic Spanish	-6.06*	-9.92**	-4.64*	-6.33**	-7.12**
Asian English	1.00	-5.85	-4.76	-6.08*	-4.02
Asian Other	-8.02*	-19.58****	-18.41****	-21.48****	-11.92**
Asian Missing	4.10	-18.69****	-14.72****	-17.98****	-.48
Black	-3.57**	-4.45***	.26	-.92	-4.58**
American Indian	-8.88*	-3.47*	-3.38*	-1.08	-4.15**
Other	-5.90	-3.79*	.62	1.35	-5.02
Missing	-9.34**	-8.91****	-4.76***	-1.77	-6.97**
Parent's Age[2]					
35-54	.11	2.60****	2.08**	3.92****	2.04**
55+	-.80	5.65****	6.49****	8.13****	6.72**
Parent's Gender[3]					
Male	-.70	-2.31	-1.39	-1.59	-3.61**
Parent's Education[4]					
< High School	-2.21*	-2.96***	-2.39**	-2.43**	-1.14
College	-1.27	-.41	-1.57**	-2.55****	-1.12
Child's Health Status[5]					
Very good	-4.08**	-3.78****	-4.54****	-4.13****	-2.63**
good	-8.56**	-8.52****	-9.25****	-7.79****	-6.34**
Fair	-18.77**	-15.26****	-13.30****	-10.45****	-9.03**
Poor	-27.70**	-18.85***	-18.24****	-21.85**	-16.35**
R^2	0.04	0.07	0.64	0.06	0.04

Reference groups: [1] Whites; [2] 18-34; [3] Female; [4] High School; [5] Excellent.
*p < .05; ** p < .01; *** p < .001; **** p < .0001

Table 5
Regression Results for Ratings by Race/Ethnicity and Language

	Personal Doctor	Specialist	Health Care	Health Plan
Parent's Race/Ethnicity[1]				
Hispanic English	.68	2.27	1.14	.14
Hispanic Spanish	3.46**	9.73*	2.13	7.75****
Asian English	-1.53	12.40**	-1.97	3.74*
Asian Other	-2.89	4.06	-3.66*	.98
Asian Missing	-4.41	-7.59	-5.75	-2.32
Black	.46	-2.33	-.04	-.04
American Indian	-2.86*	-9.83	-1.48	-4.02*
Other	.71	1.69	-1.57	-4.46
Missing	-1.70	.21	-2.12	-6.48***
Parent's Age[2]				
35-54	.49	.16	1.27*	1.26*
55+	2.73*	4.97	6.27****	6.10****
Parent's Gender[3]				
Male	-2.70**	2.00	-2.62*	-2.14*
Parent's Education[4]				
< High School	-.77	.58	-1.91*	-.67
College	-1.24*	.83	-2.32****	-3.23****
Child's Health Status[5]				
Very good	-3.22****	-2.48	-4.51****	-4.51****
Good	-6.18****	-6.60****	-9.16***	-8.47****
Fair	-7.92****	-12.04**	-14.93***	-12.66****
Poor	-14.46	-12.03	-24.36**	-21.40***
R^2	0.02	0.04	0.06	0.05

Reference groups: [1] Whites; [2] 18-34; [3] Female; [4] High School; [5] Excellent.
*p < .05; ** p < .01; *** p < .001; **** p < .

7. EVALUATING THE EQUIVALENCE OF HEALTH CARE RATINGS BY WHITES AND HISPANICS

Abstract

Purpose: To assess the equivalence of a health care ratings scale administered to non-Hispanic white and Hispanic survey respondents.

Methods: Eighteen thousand eight hundred forty questionnaires were sent to a random sample of patients receiving medical care from a physician group association concentrated in the western United States; 7,093 were returned (59% adjusted response rate). Approximately 90% of survey respondents self-identified as white/Caucasian (n = 5,508) or Hispanic/Latino (n = 713). Interpersonal and technical aspects of medical care were assessed using 9 items, all administered with a 7-point response format: *The Best, Excellent, Very Good, Good, Fair, Poor,* and *Very Poor,* with a *Not Applicable* option. Item response theory procedures were used to test for differential item functioning (DIF) between White and Hispanic respondents.

Results: Hispanics were found to be significantly more dissatisfied with care than whites (effect size=0.27; $p<0.05$). Two of 9 test items had statistically significant DIF ($p<0.05$): (1) *Reassurance and support offered by your doctors and staff* and (2) *Quality of examinations received.* However, summative scale scores and test characteristic curves for whites and Hispanics were similar whether or not these items were included in the scale.

Conclusions: Despite some differences in item functioning, valid satisfaction-with-care comparisons between whites and Hispanics are possible. Thus, disparities in satisfaction ratings between whites and Hispanics should not be ascribed to measurement bias, but should be viewed as arising from actual differences in experiences with care.

Introduction

As the health care system continues to evolve, consumers have increasingly turned to cost and quality-of-care information to guide their health care choices. Demand for such information, in turn, has fueled the number of consumer surveys conducted each year. Although such surveys can provide important information about how well health plans and clinicians are meeting the needs of their various patient populations Edgman-Levitan & Cleary, 1996; Crofton, Luliban, & Darby, 1999) a number of researchers have raised methodological concerns about their use in culturally and linguistically diverse patient populations. In addition to concerns about providing adequate translations into multiple languages (Weidmer, Brown, & Garcia, 1999) there is concern that cultural differences in the interpretation of questions (Johnson, O'Rourke, Chavez, Sudman, Warnecke et al, 1996; Angel & Thoits, 1987; Liang, Van Tran, Krause, & Markides, 1989; Dick, Beals, Keane, & Manson, 1994; Weissman, Sholomskas, Pottenger, Prusoff, & Locke, 1977) and in response styles (Hayes & Baker, 1998) may limit direct comparisons between members of different racial/ethnic groups. As a result, the quality of care provided to members of vulnerable population groups may prove difficult to monitor, evaluate, and improve. Hence, there is a need to determine the equivalence of patient satisfaction measures in different cultural and linguistic groups.

This paper addresses the comparability of ratings by Hispanic and white consumers. In a prior study, we reported greater dissatisfaction with provider communication among Hispanics than among whites, and raised the concern that undetected measurement bias may have affected our results (Morales, Cunningham, Brown, Lui, & Hays, 1999). In this study, we assess the equivalence of satisfaction with care questions administered to white and Hispanic respondents in that study (Hays, Brown, Spritzer, Dixon, & Brook,

1998). More specifically, we test for the measurement equivalence of a 9-item satisfaction with care scale using multigroup item response theory (IRT) procedures. Because no prior empirical work has addressed the comparability of patient satisfaction with care ratings for whites and Hispanics, we had no a priori hypotheses regarding particular items that might be expected to display bias. Thus, this research is exploratory in nature.

Methods

Setting

This study was based on survey data obtained from randomly selected patients receiving medical care from an association of 48 physician groups. The survey asked individuals about their satisfaction with care, health status, and use of health services during the past 12 months. Sixty-three physician group practices located primarily in the Western United States participated in the study.

Patients at least 18 years of age who made at least one provider visit during the 365 days prior to the study were eligible for the study. The field period began in October 1994 and ended in June 1995. Each patient selected was mailed both Spanish and English language versions of the survey along with a $2 cash payment and a return envelope. Survey nonrespondents were followed up with reminder postcards and telephone calls. Eighteen thousand eight hundred forty surveys were mailed out and 7,093 returned, for an overall response rate of 59%, adjusted for undeliverable surveys, ineligible respondents, and deceased individuals. Response rates across medical groups ranged from 46% to 73% and were not significantly associated with ratings of health care (Hays, Brown, Spritzer, Dixon, & Brook, 1998).

Survey Instrument

A detailed description of the survey, including a full description of its contents and psychometric properties, has been reported elsewhere (Hays, Brown, Spritzer, Dixon, & Brook, 1998). Briefly, the survey included 153 items and took approximately 27 minutes to complete. The Spanish language version of the survey was created through a process of independent forward (English to Spanish) and back (Spanish to English) translation followed by reconciliation.

This study evaluates the 9 survey items relevant to ratings of interpersonal and technical aspects of care. Five items asked about interpersonal care (*Medical staff listening; Answers to your questions; Explanations about prescribed medications; Explanations about tests and medical procedures;* and *Reassurance and support offered*) and 4 items asked about technical care (*Quality of examinations; Quality of treatment; Thoroughness and accuracy of diagnosis;* and *Comprehensiveness of exams*). All 9 survey items were asked using a 7-point response format (*The Best, Excellent, Very Good, Good, Fair, Poor, Very Poor*), with a *Not Applicable* response option.

Seventy-nine percent of respondents were white/Caucasian (white) (n = 5,508) and 10% were Hispanic/Latino (Hispanic) (n = 713). The remaining 11% were either Asian/Pacific Islander, African-American/Black, Native American/American Indian or other ethnic groups. Because precise item parameter estimation using IRT requires a large number of respondents across the trait level continuum (Hambleton, Swaiminathan, & Rogers, 1991), we retained only white and Hispanic respondents for this analysis. Although the white and Hispanic groups were similar with respect to gender and health status, Hispanics were significantly younger ($p<0.01$), more likely to be

married ($p<0.01$), and less likely to have graduated from high school ($p<0.01$) (Table 1).

Unidimensionality

Because the typical IRT model assumes sufficient unidimensionality (Widaman & Reise, 1997), we evaluated the dimensionality of our 9-item scale. First, we conducted principal components factor analysis for the white and Hispanic groups separately using the SAS FACTOR procedure (SAS Institute, Inc., 1989). For both whites and Hispanics, we examined the magnitude of the eigenvalues, the ratio of the first and second eigenvalues, the component loadings, the Tucker and Lewis coefficient (Tucker & Lewis, 1973), the average residual correlations (absolute values), and the standard deviation of the residual correlations. In addition, we computed item-scale correlation coefficients and internal consistency reliability for the white and Hispanic groups.

Overview of IRT Models

IRT models posit a nonlinear monotonic function to account for the relationship between the examinee's position on a latent trait (Θ) and the probability of a particular set of item responses (Lord, 1980). In this study, Θ refers to a respondent's level of satisfaction with care. The curves specified by this function are referred to as category response curves (CRC). We used the generalized partial credit model as implemented in Parscale 3.5 (Muraki & Block, 1997) to estimate the relationship between Θ and the item response probabilities. This model was developed for scales composed of items with polytomous response formats and defines the CRCs for each item (i) and response category (k) as follows:

$$P_{ik} = \frac{\exp\left[\sum_{v=1}^{k} a_i(\theta - \lambda_i + \tau_k)\right]}{\sum_{c=1}^{K} \exp\left[\sum_{v=1}^{c} a_i(\theta - \lambda_i + \tau_k)\right]} \qquad (1)$$

where each item is represented by 3 parameters $(a_i, \ddot{e}_i, \hat{o}_k)$ and the examinee trait level is represented by 1 parameter, Θ. For identification purposes, the latent trait scale is specified to have a mean of 0 and a standard deviation of 1.0. The τ_k parameters are called category intersection parameters; there are 6 such parameters for an item with 7 response options. The generalized partial credit model requires that the category intersection parameters be estimated for the scale as a whole (i.e., remain constant across items).

The λ_i parameter is called an item location parameter. It indicates the difficulty of an item and can be thought of as shifting the intersection parameters up and down the latent trait scale. Large positive values of λ_i indicate a difficult item in which few examinees respond in the highest categories. Negative λ_i values indicate an easy item in which many examinees respond in the highest category. The slope parameter, a_i, indicates how fast the probability of responding in a higher category changes as a function of increases in the trait level. Items with large a_i are more discriminating than items with smaller slopes.

Assessing Goodness-of-Fit

There is no widely accepted goodness-of-fit statistic or index available for polytomous IRT models. To assess fit, we computed the difference between the observed and expected response frequencies by item and response category for whites and Hispanics. Parscale 3.5 does produce an

item-fit chi-square statistic based on these cell frequencies, but this test is too sensitive to sample size to produce a good gauge of model fit (Orlando & Thissen, *in press*; Reise, 1990).

Assessing Measurement Invariance with IRT

Measurement invariance (no bias) occurs when the CRCs for each item of a scale are identical for the groups of examinees in question (e.g., whites and Hispanics)(Kok, 1988). Conversely, when particular item CRCs are not identical, measurement invariance is not obtained. The IRT literature uses the term differential item functioning (DIF) to describe items with nonidentical CRCs across groups.

In this study, DIF is determined by contrasting the item parameters (i.e., a_i and λ_i parameters) that determine the CRCs for whites and Hispanics (Thissen, Steinberg, & Wainer, 1993). Since the CRCs are completely determined by their corresponding item parameters, CRCs can only be identical if the item parameters that determine them are equal.

To guard against finding item DIF by chance alone, we conducted our analyses in a stepwise fashion. First, we contrasted a multigroup model in which the slope and location parameters were freely estimated between groups (unconstrained model) with a multigroup model in which the slope and location parameters were constrained to equality across groups (fully constrained model). A significant difference in the likelihood function value for the 2 models was interpreted as indicating the presence of DIF without identifying the particular items accounting for it (Thissen, 1991).

Subsequently, we fit 2 additional multigroup models to test individual items for DIF. In the first model, we freely estimated the slope parameters across ethnic groups while constraining the location parameters to equality.

Then, we compared the slope parameters for each item using the following effect size statistic:

$$SDIF = DIF \Big/ \sqrt{Var\,\hat{a}_{i(White)} + Var\,\hat{a}_{i(Hispanic)}}\ , \quad (2)$$

where *DIF* = $\hat{a}_{i(white)} - \hat{a}_{i(Hispanic)}$. *SDIF* refers to standardized differential item functioning and is evaluated as chi-square with 1 degree of freedom.[21]

In the second model, we freely estimated the location parameters across ethnic groups while constraining the slope parameters to equality. We computed a similar statistic that contrasted the location parameters for each item:

$$SDIF = DIF \Big/ \sqrt{Var\hat{e}_{i(White)} + Var\hat{e}_{i(Hispanic)}}\ , \quad (3)$$

where DIF = $\hat{\lambda}_{i(White)} - \hat{\lambda}_{i(Hispanic)}$. Note that in both models, the category intersection parameters (τ_k) are constrained to equality across ethnic groups. For this study, an item was considered to display DIF if its test-statistic chi-square value was significant at the 0.05 level.

Results

Descriptive Results and Unidimensionality of Scale

Table 2 shows the raw score descriptive statistics (i.e., means and standard deviations) and inter-item correlation coefficients for each ethnic group. Also shown is the ethnic group effect size (the group mean difference divided by the pooled standard deviation) for each item and for the scale. A total scale score was computed by summing across the 9 items (possible 0-100 range). The total score was 67.86 (SD=16.11) for whites (n=5,508) and 63.54 (SD=16.34) for Hispanics (n=713). The difference between the mean

scores was significant ($t=6.74$, $p<0.01$) and resulted in an effect size of 0.27 (pooled standard deviation of 16.14). Thus, assuming no item bias (measurement invariance), Hispanics scored nearly one third of a standard deviation lower than whites on this satisfaction with care scale.

The inter-item correlation coefficients ranged from 0.66 to 0.83 for whites and from 0.69 to 0.84 for Hispanics (Table 2). The results of the principal components analysis of the 9 items indicated 1 dimension for whites and Hispanics. For both whites and Hispanics, only one eigenvalue was greater than 1; it accounted for 78% of the total variance for whites and 77% of the total variance for Hispanics. The ratio of the first and second eigenvalues was 7.1/0.4 = 17.8 for whites and 6.9/0.5 = 17.3 for Hispanics. The mean residual correlation (absolute value) after extraction of one factor was 0.03 (SD=0.03) for whites and 0.03 (SD=0.03) for Hispanics. The Tucker and Lewis coefficient for a one factor solution was 0.96 and 0.94 for whites and Hispanics, respectively. Principal components loadings were 0.83 or larger for both whites and Hispanics, and item-scale correlation coefficients (corrected for overlap) ranged from 0.81 to 0.89 for whites and from 0.79 to 0.89 for Hispanics (Table 3). Alpha coefficients for both whites and Hispanics were 0.96. By any standard factor analytic/psychometric criterion, this 9-item scale is unidimensional (McDonald, 1967).

Goodness-of-Fit

Table 4 shows the difference between the observed and expected response frequencies by item and response category for whites and Hispanics as evidence of data-model fit. The mean discrepancy (absolute values) across all items and all response categories was 0.04 (SD=0.03) for whites and 0.02 (SD=0.02) for Hispanics. The item fit chi-square statistics generated by Parscale were significant ($p<0.05$) for both groups across all items.

Item Response Theory Results

The mean score difference between whites and Hispanics on the latent trait scale was 0.27 (SD=0.99), which is consistent with the raw score effect size noted above. The difference in likelihood function value between the unconstrained model and the fully constrained model was statistically significant at the $p<0.05$ level, indicating the presence of item-level DIF.

Table 5 shows item slope parameter estimates and the slope parameter DIF statistics. It is worth noting that the mean item slopes were 2.86 for whites and 2.88 for Hispanics, indicating good model fit at the item level, and that the items in the scale are highly discriminating. Slope parameter values greater than 2.0 are generally regarded as high.[12] The DIF test results show that the slope parameter estimates for Items 5 (_= 4.11, p = 0.04) and 6 (_ = 11.94, p < 0.01) were statistically different between the two groups. The slope parameter estimates for Item 5 were 2.84 for whites and 2.53 for Hispanics. Similarly, the slope parameter estimates for Item 6 were 3.09 for whites and 3.70 for Hispanics.

Table 6 shows item location parameter estimates and the location parameter DIF statistics. The DIF statistics indicate that no items demonstrated DIF with respect to item location. Only Item 6, for which the location parameter estimates were -0.83 for whites and -0.76 for Hispanics, had a nearly significant DIF statistic (_ = 3.65, p = 0.05).

Assessing the Impact of Items with DIF

To evaluate the impact of the item-level DIF on raw scale scores, we dropped the biased items from the scale and recomputed the effect size for whites' versus Hispanics' satisfaction ratings. The effect sizes were computed based on a summative scale (0-100 possible range). After dropping

Item 5, we obtained scale scores of 67.8 (SD = 16.0) for whites and 63.6 (SD = 16.3) for Hispanics, and an effect size of 0.26 (pooled SD = 16.0). After dropping Item 6, we obtained scale scores of 67.7 (SD = 16.5) for whites and 63.4 (SD = 16.5) for Hispanics, and an effect size of 0.26 (pooled SD = 16.5). Finally, dropping Items 5 and 6 from the scale simultaneously, we obtained scale scores of 67.7 (SD =16.4) for whites and 63.5 (SD = 16.6) for Hispanics, and an effect size of 0.26 (pooled SD = 16.4). Recall that with all 9 items, the effect size was 0.27.

To further assess the effect of the detected item bias on our measure of satisfaction with care, we compared test response curves for whites and Hispanics using the following procedure. (The test response curves show the relationship between the underlying level of satisfaction and the expected raw score on the 9-item scale.) First, we estimated the IRT item parameters for the 9-item satisfaction scale independently for whites and Hispanics. This is equivalent to estimating a simultaneous multigroup model without between-group constraints on any of the parameters. However, because the 2 sets of item parameters may not be on the same scale, we rescaled the item parameter estimates for Hispanics to those for whites by estimating linking constants and performing the appropriate transformations. Using the 2 sets of commonly scaled item parameters, we then computed the test response curves for whites and Hispanics.

Figure 1 shows the test response curves for whites and Hispanics. Deviations between the test response curves for whites and Hispanics show the degree of differential scale functioning due to Items 5 and 6. Figure 2 shows the results of subtracting the Hispanics' test response curve from the whites' test response curve. At low satisfaction levels, whites tend to score higher than Hispanics, whereas at middle levels of satisfaction

Hispanics tend to score higher than whites. However, the largest differential scale functioning (bias) is 1.5, which occurs at the -2.0 satisfaction level. A differential of 1.5 (on the 0-100 score range) represents less than one-tenth of a standard deviation difference between whites and Hispanics with the same latent trait level.

Discussion

This study examined a satisfaction with care scale for equivalence among 2 demographically important groups in the United States - whites and Hispanics. Our study found that valid comparisons between whites and Hispanics are possible, despite detection of statistically significant differences in the slope parameters for 2 of 9 scale items. More specifically, we found that Item 5 (*Reassurance and support*) and Item 6 (*Quality of examinations*) showed statistically significant DIF ($p<0.05$), but that the DIF did not have a meaningful impact on the expected scores of whites and Hispanics responding to these items. As a result, Hispanics' significantly lower rating of care in this study should be viewed as representing actual differences in experiences with care and should not be attributed to biased measurement.

Previous methodological studies of survey questions have found evidence that whites and Hispanics may not respond similarly. Johnson et al. (1998) found qualitative differences in whites' and Hispanics' interpretation of health status questions from widely used health surveys. Hayes and Baker (1998) found that the reliability and validity of a Spanish version of a patient satisfaction with communication scale differed significantly from that of the English language version. Aday and colleagues (1980) noted that Hispanics were more likely to respond "yes" to patient satisfaction questions than non-Hispanics, regardless of whether the question indicated greater

satisfaction or dissatisfaction, providing support for the contention that Hispanics are prone to more acquiescent responses than non-Hispanics or are biased toward more favorable responses (Hayes & Baker, 1998).

Unlike many prior studies, we conducted analyses to assess the effect of differences in scale functioning among whites and Hispanics on comparisons between the groups. Specifically, we examined the effect of the 2 biased items on the group mean scale scores and computed the effect size with and without including the items showing DIF. When all 9 items were included in the scale, the effect size was 0.27, with whites rating care significantly more positively than Hispanics ($p<0.05$). When the biased items - Items 5 and 6 - were dropped from the scale, the effect size changed to 0.26 and the mean scale scores remained significantly different ($p<0.05$).

Further, we examined the test response cures for whites and Hispanics. These curves plot the expected raw scale scores of each group over the underlying satisfaction with care continuum. At worst, our nine-item scale resulted in a 1.5 raw score differential (bias) between whites and Hispanics. Together, these results show that at all levels of satisfaction, whites and Hispanics have nearly identical expected raw scale scores despite 2 items with statistically significant DIF.

Our study uses a relatively new procedure for detecting DIF that is based on polytomous IRT model procedures. Prior studies have primarily relied on classical psychometric methods (e.g., reliability, validity, and item-scale correlation), exploratory factor analysis (EFA), and confirmatory factor analysis (CFA) for the identification of item and survey bias in multiethnic settings. Although these methods can yield useful information about item and scale bias, IRT models are theoretically more appropriate for survey scales that use categorical response formats. While EFA and CFA

models typically assume continuous indicators that have linear relationships with the latent variable(s), IRT models do not make these assumptions. Furthermore, IRT models do not assume multivariate normality, which is an assumption made by most CFA estimation routines.

IRT models also offer practical approaches to quantifying the effect size of statistically significant DIF. As other studies have illustrated (Smith & Reise, 1998) and as we have demonstrated in this study, statistically significant DIF does not necessarily invalidate comparisons between groups of interest. EFA and CFA models do not offer a similarly practical approach to assessing the impact of DIF when it is detected. For more detailed discussions of IRT and factor analytic approaches to item and scale bias detection, the reader is referred to McDonald (1999) and Reise, Widaman and Pugh (1993).

Explaining why the items asking about *quality of examinations* and *reassurance and support* demonstrated DIF is beyond the scope of this study, and thus remains speculative. Item bias occurs when an instrument measures one thing for one group and something else for the other group. Items 5 and 6 may have been interpreted differently by white and Hispanic respondents because of between-group differences in age, gender, income, education, or cultural background. Although we found significant differences in the sociodemographic characteristics of the whites and Hispanic respondents in our study, our purpose was not to identify factors that explain the DIF we detected. Based on the results of this study, we cannot attribute the DIF in these items to ethnicity per se or to any other particular background or health status variable. Future studies may be needed to explain the influence of background characteristics on differences in item functioning.

The moderate response rate (59%) in this study may pose some risk of non-response bias. To threaten the validity of this study, however, respondents and non-respondents would have to differ with respect to their interpretations of the meanings of the survey questions. This might occur, for example, if Hispanic respondents were more acculturated than Hispanic nonrespondents. (Acculturation refers to the processes of acquisition the host culture by an ethnic minority (Berry, 1998). In this scenario, the Hispanic respondents in our study would be culturally more similar to the white respondents than a truly representative sample of the Hispanic patients would be; therefore, our study would be less likely to find measurement bias than a study with a more representative Hispanic sample. Unfortunately, our data sources do not allow us to compare respondents and nonrespondents along such dimensions as acculturation.

Based on the available data, the differences between the sampling frame and those responding to the survey were minimal. Specifically, those returning the questionnaire had a mean age of 51 years (median=49 years), whereas the mean age of the sampling frame was 46 years (median=43 years). Sixty-five percent of the responders were women; 58% in the sampling frame were women. The last medical visit for the study participants was, on average, 119 days (median=88 days) before the beginning of the study. For those in the sampling frame, the average was 130 days (median=112 days)(Hays, Brown, Spritzer, Dixon, & Brook, 1998). Unfortunately, our data sources prevented us from computing ethnic-group-specific response rates.

In sum, this study addressed the validity of comparisons of satisfaction with care across ethnic groups. We found that lower ratings of care among Hispanics relative to whites were not attributable to item or scale bias and therefore reflect actual differences in experiences with care

between the 2 groups. These results support the findings of other researchers that Hispanics are not as well served by the current health care system as whites (Morales, Cunningham, Brown, Lui, & Hays, 1999; Andersen, Lewis, Giachello, Aday, & Chiu, 1981; 7Baker, Parker, Williams, Coates, & Pitkin, 19996; Hu & Covell, 1986; Harpole, Orav, Hickey, Posther, & Brennan, 1996; Molina, Zambrana, & Aguirre-Molina, 1997; Villa, Cuellar, Gamel, & Yeo, 1993). More generally, our findings suggest that when disparities in patient ratings of care are detected across ethnic groups, they should not be attributed to biased measurement unless significant DIF (in the statistical and practical sense) can be demonstrated.

Acknowledgements

Support for this research was received from the Agency for Health Care Policy Research to RAND (U18HS09204) (RD Hays, PI; LS Morales, Co-PI) and an unrestricted research grant from The Medical Quality Commission to RAND (RD Hays, PI). The authors express appreciation to Tamara Breuder for her assistance in preparing the manuscript and Gail Della Vedova and the staff members of The Medical Quality Commission for their cumulative input. The views expressed herein are those of the authors and do not necessarily reflect the views of The Medical Quality Commission, RAND or UCLA.

References

Aday LA, Chiu GY, Andersen R. Methodological issues in health care surveys of the Spanish heritage population. Am J Public Health. 1980;70:367.

Angel R, Thoits P. The impact of culture on the cognitive structure of illness. Culture, Med Psychiatry. 1987;11:465-94.

Andersen R, Lewis SZ, Giachello AL, Aday LA, Chiu G. Access to medical care among the Hispanic population of the southwestern United States. J Health Soc Behav. 1981;22:78-89.

Baker DW, Parker RM, Williams MV, Coates WC, Pitkin K. Use and effectiveness of interpreters in an emergency department. JAMA. 1996;275:783-8.

Berry JW. Acculturative stress. In: Organista PB, Chun KM, Marin G, eds. Readings in ethnic psychology. London, England: Routledge; 1998.

Crofton C, Luliban JS, Darby C. Foreword. Med Care. 1999;37:MS1-MS9.

Dick RW, Beals J, Keane EM, Manson SM. Factorial structure of the CES-D among American Indian adolescents. J Adolesc. 1994;17:73-79.

Edgman-Levitan S, Cleary PD. What information do consumers want and need? Health Affairs. 1996;Winter:42.

Hambleton RK, Swaiminathan H, Rogers HJ. Fundamentals of item response theory. Thousand Oaks, CA: Sage; 1991.

Harpole LH, Orav EJ, Hickey M, Posther KE, Brennan TA. Patient satisfaction in the ambulatory setting. Influence of data collection methods and sociodemographic factors. J Gen Intern Med. 1996;11:431-4.

Hayes RP, Baker DW. Methodological problems in comparing English-speaking and Spanish-speaking patients' satisfaction with interpersonal aspects of care. Med Care. 1998;36:230-6.

Hays RD, Brown JA, Spritzer KL, Dixon WJ, Brook RH. Member ratings of health care provided by 48 physician groups. Arch Intern Med. 1998;158:785-90.

Hu DJ, Covell RM. Health care usage by Hispanic outpatients as a function of primary language. West J Med. 1986;144:490-3.

Johnson TP, O'Rourke D, Chavez N, Sudman S, Warnecke RB, Lacey L, et al. Cultural variations in the interpretation of health questions. Health Survey Research Methods: Conference Proceedings. 1996: 57-62.

Liang J, Van Tran T, Krause N, Markides KS. Generational differences in the structure of the CES-D scale in Mexican Americans. J Gerontol. 1989;44:S110-20.

Lord FM. Applications of item response theory to practical testing problems. Hillsdale, N. J.: Erlbaum; 1980.

McDonald R. The dimensionality of tests and items. Br J Math Stat Psychol. 1967;34:100-117.

McDonald RP. Test theory: a unified treatment. Mawah, New Jersey: Lawrence Erlbaum Associates, Inc.; 1999.

Molina CW, Zambrana RE, Aguirre-Molina M. The influence of culture, class, and environment on health care. In: Molina CW, Aguirre-Molina M, eds. Latino health in the U.S.: a growing challenge. Washington, D. C.: American Public Health Association; 1997:23-43.

Morales LS, Cunningham WE, Brown JA, Lui H, Hays RD. Are Latinos less satisfied with communication by health care providers? A study of 48 medical groups. J Gen Intern Med. 1999;14:409-17.

Muraki E, Block RD. Parscale IRT item analysis and test scoring for rating scale data. Chicago: Scientific Software International, Inc.; 1997.

Orlando M, Thissen D. New fit indices for dichotomous item response theory models. Applied Psychological Measurement. In press.

Reise SP. A comparison of item- and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement. 1990;14:127-37.

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull. 1993;114:552-66.

SAS Institute, Inc. SAS/STAT user's guide, version 6. 4th ed. Cary, NC: SAS Institute Inc.; 1989:773-823.

Smith LL, Reise SP. Gender differences on negative affectivity: an IRT study of differential item functioning on the multidimentional personality questionnaire stress reaction scale. Journal of Personality and Social Psychology. 1998;75:1350-62.

Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holand PW, Wainer H, eds. Differential item functioning. Hillsdale, N. J.: Erlbaum; 1993.

Thissen D. MULTILOG: Mulitple, categorical item analysis and test scoring using item response theory (version 6). Chicago: Scientific Software, Inc.; 1991.

Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika. 1973;38.

Villa ML, Cuellar J, Gamel N, Yeo G. Aging and health: Hispanic American elders. . Stanford, CA: Stanford School of Medicine: Stanford Geriatric Education Center; 1993.

Weidmer B, Brown J, Garcia L. Translating the CAHPS 1.0 survey instrument into Spanish. Med Care. 1999;37:MS89-MS97.

Weissman MM, Sholomskas D, Pottenger M, Prusoff BA, Locke BZ. Assessing depressive symptoms in five psychiatric populations: a validation study. Am J Epidemiol. 1977;106:203-14.

Widaman KF, Reise, S. P. Exploring the measurement invariance of psychological instruments: applications in the substance abuse domain. In: Bryant KJ, Windle M, West SG, eds. The science of prevention: methodological advances from the alcohol and substance abuse research. Washington, DC: American Psychological Association; 1997.

Table 1. Sample Description

	Whites (n=5,508)	Hispanics (n=713)	p-value for difference
Age (mean, (SD))	51.9 (17.52	41.7 (15.22	<0.01
% Male	34.9	37.5	0.12
% Married	73.7	78.1	<0.01
% Graduated High School	69.3	46.8	<0.01
General Health Status (mean score on 0-10 scale; 10 = best, (SD))	7.3 (1.73)	7.2 (1.79)	0.24

Note. Two-sided t-tests were applied to continuous variables (age, health status) and chi-square tests to proportions (% male, % married and % graduated HS).

Table 2. Raw Score Descriptive Statistics and Inter-Item Correlations.

Item	Whites (n=5,508)		Hispanics (n=713)		Effect Size	Inter-Item Correlations								
	Mean	SD	Mean[1]	SD		1	2	3	4	5	6	7	8	9
1	5.01	1.34	4.72	1.35	0.21		0.83	0.70	0.72	0.77	0.83	0.79	0.77	0.69
2	5.00	1.29	4.70	1.30	0.23	0.83		0.75	0.76	0.77	0.80	0.80	0.78	0.69
3	4.97	1.35	4.65	1.37	0.23	0.68	0.72		0.76	0.73	0.71	0.74	0.73	0.69
4	4.76	1.36	4.47	1.38	0.21	0.70	0.74	0.77		0.77	0.73	0.76	0.76	0.72
5	4.76	1.38	4.43	1.35	0.24	0.74	0.74	0.74	0.76		0.76	0.76	0.76	0.76
6	5.09	1.23	4.70	1.34	0.31	0.83	0.79	0.67	0.73	0.75		0.83	0.81	0.72
7	5.05	1.28	4.70	1.35	0.27	0.77	0.81	0.73	0.76	0.77	0.82		0.84	0.76
8	4.85	1.35	4.47	1.38	0.28	0.74	0.77	0.73	0.74	0.74	0.79	0.82		0.74
9	4.59	1.45	4.26	1.44	0.23	0.66	0.66	0.67	0.68	0.77	0.68	0.74	0.71	
Total	67.86	16.11	63.54	16.34	0.27									

Note. Individual item scores range from 1-7 (7 = The Best). The total score was computed by summing across the 9 item scores, then transforming to a 0-100 scale, where 100 is the highest possible rating. Effect size was computed as the difference in means divided by the pooled standard deviation. Inter-item

correlation coefficients for Hispanics are shown above the diagonal; those for whites are shown below the diagonal.

[1]All means differences between whites and Hispanics were statistically significantly ($p<0.05$).

Table 3. Principal Component Loadings and Item-Scale Correlations for White, Hispanic and Merged Samples

Item	Principal Component Loadings		Item-Scale Correlations	
	White	Hispanic	White	Hispanic
1	0.89	0.88	0.86	0.84
2	0.90	0.89	0.87	0.86
3	0.85	0.85	0.82	0.81
4	0.88	0.87	0.84	0.83
5	0.89	0.89	0.86	0.86
6	0.90	0.89	0.87	0.86
7	0.92	0.91	0.89	0.89
8	0.90	0.89	0.87	0.86
9	0.85	0.83	0.81	0.79

Note. Loadings were derived from a single factor principal components model. The ratio of the 1st and 2nd eigenvalues was 7.1/0.4 = 17.8 for whites and 6.9/0.5 = 17.3 for Hispanics. Item-scale correlations were corrected for overlap. Cronbach's alpha is 0.96 for both whites and Hispanics.

Table 4. Difference between observed and expected response frequencies (absolute values) by item and response category for whites and Hispanics.

	Response Category							p-value
	1	2	3	4	5	6	7	
Whites								
Item 1	0.00	0.00	0.02	0.06	0.05	0.06	0.08	<0.05
Item 2	0.00	0.00	0.02	0.08	0.05	0.07	0.08	<0.05
Item 3	0.00	0.00	0.02	0.06	0.05	0.05	0.07	<0.05
Item 4	0.00	0.00	0.03	0.07	0.03	0.07	0.07	<0.05
Item 5	0.00	0.01	0.03	0.07	0.02	0.07	0.07	<0.05
Item 6	0.00	0.00	0.01	0.07	0.06	0.06	0.08	<0.05
Item 7	0.01	0.00	0.03	0.07	0.06	0.06	0.09	<0.05
Item 8	0.00	0.00	0.03	0.08	0.03	0.07	0.07	<0.05
Item 9	0.00	0.01	0.04	0.06	0.01	0.06	0.06	<0.05
Hispanics								
Item 1	0.00	0.00	0.01	0.04	0.05	0.02	0.06	<0.05
Item 2	0.00	0.00	0.01	0.04	0.06	0.03	0.05	<0.05
Item 3	0.00	0.00	0.00	0.04	0.04	0.03	0.05	<0.05
Item 4	0.00	0.01	0.00	0.05	0.04	0.03	0.05	<0.05
Item 5	0.01	0.00	0.01	0.04	0.03	0.03	0.05	<0.05
Item 6	0.00	0.00	0.01	0.04	0.05	0.03	0.06	<0.05
Item 7	0.00	0.00	0.00	0.03	0.05	0.03	0.06	<0.05
Item 8	0.00	0.00	0.00	0.04	0.04	0.04	0.05	<0.05
Item 9	0.00	0.00	0.00	0.05	0.02	0.03	0.03	<0.05

Note. The mean difference (absolute values) between the observed and expected response frequencies across all items and all response categories

was 0.03 (SD=0.04) for whites and 0.02 (SD=0.02) for Hispanics. The reported p-values are based on the item-fit χ^2 reported by Parscale 3.5.

Table 5. Slope Parameters and Differential Item Functioning Tests for Whites and Hispanics

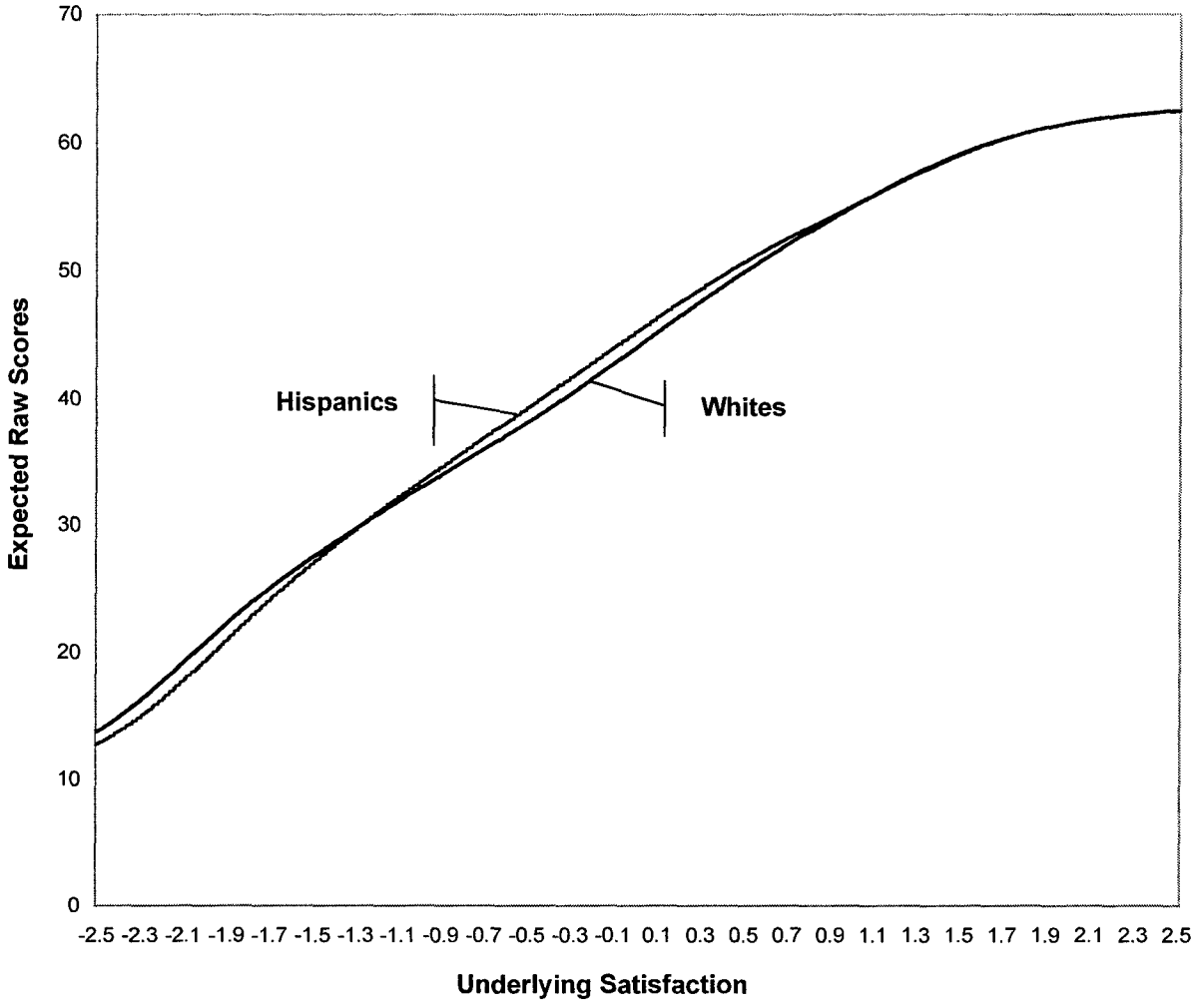
Item	White		Hispanic		SDIFF	Chi-square (df=1)	p-Value
	Slope	SE	Slope	SE			
1	2.99	0.15	2.90	0.05	0.55	0.30	0.59
2	3.52	0.15	3.32	0.06	1.28	1.64	0.20
3	2.09	0.10	2.00	0.03	0.81	0.65	0.43
4	2.39	0.09	2.34	0.04	0.46	0.21	0.65
5	2.84	0.14	2.53	0.04	2.03	4.11	0.04*
6	3.09	0.16	3.70	0.07	-3.46	11.94	<0.01*
7	3.97	0.23	4.08	0.08	-0.47	0.23	0.64
8	3.11	0.15	3.28	0.05	-1.07	1.13	0.29
9	1.77	0.08	1.78	0.03	-0.10	0.01	0.88

* p Value < 0.05.

Table 6. Location Parameters and Differential Item Functioning Tests for Whites and Hispanics

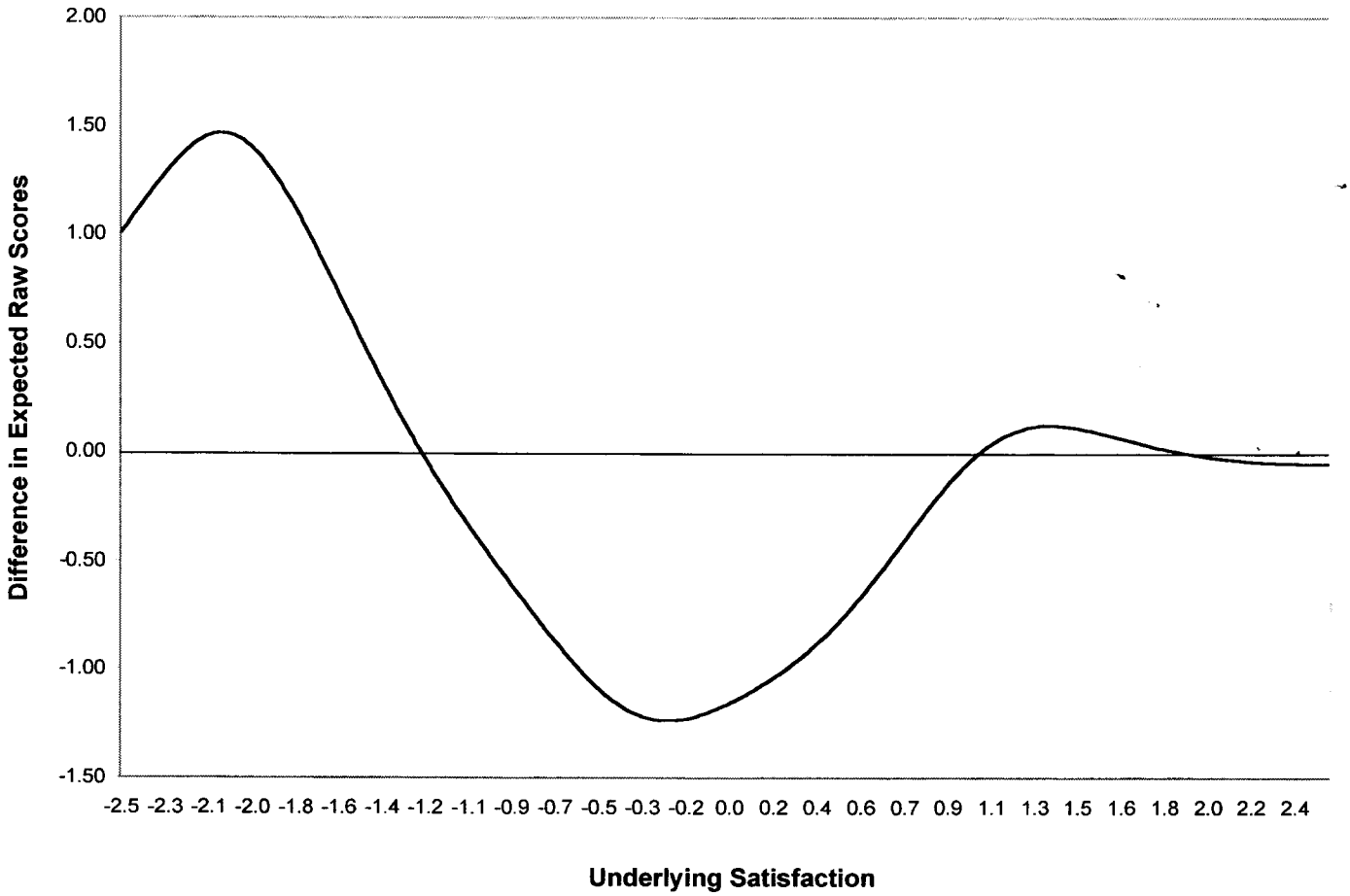
Item	White Location	SE	Hispanic Location	SE	SDIFF	Chi-square (df=1)	p-Value
1	-0.71	0.01	-0.76	0.04	-1.15	1.31	0.25
2	-0.74	0.01	-0.77	0.03	-0.75	0.56	0.46
3	-0.72	0.01	-0.72	0.04	-0.07	0.01	0.90
4	-0.58	0.01	-0.61	0.04	-0.87	0.76	0.39
5	-0.55	0.01	-0.54	0.04	0.00	0.00	0.95
6	-0.83	0.01	-0.76	0.03	1.91	3.65	0.05
7	-0.77	0.01	-0.76	0.03	0.43	0.19	0.67
8	-0.61	0.01	-0.58	0.03	0.84	0.70	0.41
9	-0.43	0.01	-0.44	0.04	-0.11	0.01	0.88

Figure 1. Expected raw scores for whites and Hispanics on 9-item satisfaction with care scale.



Note. Each item score from 1 to 7 resulting in a 9-63 scale range.

Figure 2. Difference between white and Hispanic expected raw scores on 9-item satisfaction with care scale.



Note. Values greater than zero indicate Hispanic scores exceed white scores. Values less than zero indicate the converse.

8. Confirmatory Factor Analysis of the Consumer Assessment of Health Plans Study (CAHPS®) 1.0 Core Survey

Abstract

The National Consumer Assessment of Health Plans Survey (CAHPS®) Benchmarking Database was used to assess the factor structure and invariance of the CAHPS® 1.0 Core Survey in four different samples of consumers. Four analytic samples (total N = 15,092) consisted of Hispanic and white health care consumers receiving care in commercial and Medicaid sectors. As hypothesized, results demonstrated that the 23 CAHPS® 1.0 report items capture consumer experiences with five aspects of health plan performance: *Access to Care*, *Timeliness of Care*, *Provider Communication*, *Health Plan Consumer Service*, and *Office Staff Helpfulness*. The CAHPS® instrument also contains four items assessing *Global Ratings* of care. Analyses revealed an identical pattern of pattern of fixed (i.e., zero) and free (i.e., non-zero) factor loadings across all samples. In addition, considerable evidence of stability in the magnitude of factor loadings was found for consumers receiving care within a common health service sector. A higher order factor analysis of report and rating items revealed evidence of concordance across different methods of measuring overall assessments of health care.

Confirmatory Factor Analysis of the Consumer Assessment of Health Plans Study (CAHPS®) 1.0 Core Survey

A confluence of interrelated social forces, including the rise of consumerism, the advent of managed care, and an increasingly competitive health care marketplace has spurred interest in consumer evaluations of health services. Consumer judgments about health care have emerged as an essential index of health care quality, and a significant determinant of marketplace viability. Consumer evaluations have been associated with various health-related behaviors including the initiation of malpractice litigation (Penchansky & McGee, 1994; Vaccarino, 1977), disenrollment from health plans and providers (Allen & Rogers, 1997; Newcomer, Preston, & Harrington, 1996; Schlesinger, Druss, & Thomas, 1999), and poor adherence to medical regimens (Hall, Roter, & Katz, 1988). Perhaps not surprisingly, consumer evaluations have been implicated as both an antecedent and consequence of good health (e.g., Hall, Roter, & Milburn, 1999; Marshall, Hays, & Mazel, 1996).

The Consumer Assessment of Health Plans Survey (CAHPS®; Cleary & Edgman-Levitan, 1997; Hays et al., 1999) is perhaps the most widely used measure of consumer evaluations of health care, having been adopted by the National Committee for Quality Assurance as part of its accreditation process (NCQA, 1998). In 1999, CAHPS® data were available to assist the decision making of an estimated 90 million Americans (Cleary, 1999).

The CAHPS® 1.0 Core Survey possesses several desirable features. First, the CAHPS® 1.0 assesses both reports about specific health care

experiences (e.g., delays in gaining access to care) as well as global ratings of care received (e.g., overall judgments concerning one's health plan). Incorporation of reports of health care experiences can help to target specific domains in need of quality improvement, thereby addressing the criticism that general ratings are too nebulous to yield practically meaningful information (e.g., Williams, 1994). By including items about both specific experiences and global impressions, it is possible to examine empirically the association between the two types of information.

In addition, although many preceding instruments have required respondents to provide answers without regard to a specific time frame (e.g., Kippen, Strasser, & Joshi, 1997), the CAHPS® 1.0 items have been expressly anchored to a specific reference period (i.e., the past six months). Given the interpretative difficulties inherent in attempting to evaluate change over time when items are answered without respect to a specific time frame, incorporation of a temporal reference frame offers clear advantages (Cleary et al., 1998).

Despite the apparent promise of the CAHPS® 1.0 survey instrument as a tool for measuring consumer appraisals of health care and its increasingly widespread adoption for this purpose, key questions have yet to be answered. First, although some information concerning the psychometric properties of the CAHPS® 1.0 instruments has been documented (e.g., Hays et al., 1999; Zaslavsky, Beaulieu, Landon, & Cleary, 2000), fundamental assumptions underlying the development of the instrument have not yet been demonstrated. In particular, although the CAHPS® 1.0 was explicitly designed to assess consumer experiences with respect to multiple dimensions of care,

there has—as yet—been no evaluation of the factor structure of the instrument at the level of the individual.[1] Thus, the multi-factorial structure of the CAHPS® 1.0 has been assumed but not demonstrated.

Additionally, although supplemental items have been developed to tailor the CAHPS® to the assessment needs of various vulnerable populations including persons on Medicaid or Medicare, and children (Brown, Nederend, Hays, Short, & Farley, 1999; Schnaier et al., 1999; Shaul et al., 1999), the core set of CAHPS® 1.0 items was constructed to measure the same constructs across diverse populations. As yet, this assumption of measurement equivalence across diverse groups has not been tested explicitly. Thus, demonstration of the equivalence of the core instrument across persons from different ethnic groups and health service sectors would provide much needed information concerning the validity of the instrument for use with diverse groups receiving care in different health settings.

An additional issue concerns the degree of concordance between alternative methods of summarizing experiences with health care (Marshall, Hays, Sherbourne, & Wells, 1993). The CAHPS® instrument incorporates two strategies. Although conceptually distinct, each approach would be expected to yield comparable information. One approach assesses global perceptions of care by asking consumers to indicate the extent to which they agree with statements reflecting beliefs about broad domains of care (e.g., *overall rating of health care provider*). This strategy has two principal strengths. First, it requires comparatively few items, thus minimizing respondent burden. Second, by measuring evaluations of overall care directly, fewer

assumptions about the specific content domains of potential relevance are required.

An alternative measurement strategy calls for consumers to provide reports about the frequency with which specific experiences took place within multiple dimensions of care. Consumer views regarding overall health care can be inferred by combining domain-specific judgments to yield a single summary index. The major strength of the latter approach is that it offers considerable detail about specific features of care that may be otherwise obscured in more global ratings. There are, however, at least two significant disadvantages associated with this strategy. First, explicit assumptions are required concerning the number of relevant domains. Moreover, multiple items are required to assess these domains, thereby increasing respondent burden. Given the trade-offs associated with these two strategies, direct comparison of the two approaches would provide valuable insight into the conceptualization and measurement of overall ratings of health care.

In summary, the goals of this study were to confirm: (a) the individual-level factor structure of the CAHPS® 1.0 adult survey; (b) the invariance of the factor structure across persons from white and Hispanic ethnic groups as well as individuals receiving care within commercial and Medicaid health service sectors; and (c) the concordance between direct and indirect methods of summarizing experiences with health care.

Method

Data Source

The data were drawn from adult surveys of the National CAHPS® Benchmarking Database (NCBD 1.0). The purpose of the NCBD 1.0 is to provide a frame of reference against which to compare the results obtained by other CAHPS® users. Health care plans and sponsors participated voluntarily in this data pooling effort, and included Medicaid as well as commercial participants.

The Medicaid database included information from 29 health maintenance organizations (HMO) and two primary care case management programs located in the District of Columbia, Arkansas, Kansas, Minnesota, Oklahoma, and Washington. The database of commercially insured persons included information from 27 HMOs, eight physician provider organizations, three point-of-service plans, one fee-for-service plan, and 15 other unspecified health plans in the District of Columbia, Florida, Kansas, New Jersey, Oklahoma and Washington.

Data were collected, using a mix of telephone interviews and mail surveys, from 28,354 enrolled health service consumers between 1997 and 1998. Previous research attests to the equivalence of CAHPS® information obtained via these two methods (Fowler, Gallagher, & Nederend, 1999). Surveys were administered in either English or Spanish as necessary. Records were not available, however, as to the language version used by a given respondent.

Participants

Participants completing the survey did so voluntarily, and were guaranteed of the confidentiality of the data. The overall response rate for the survey was 52%, with rates of 66% and 34% for commercial and Medicaid participants, respectively. For the purposes of the current study, participants included Hispanic and white (non-Hispanic Caucasian) enrollees who were at least 18 years of age. Four separate analytic samples were created: Hispanics drawn from the commercially-insured (N = 1,020) and Medicaid sectors (N = 609) and whites drawn from the commercially-insured (N = 7,983)[2] and Medicaid sectors (N = 5,480). Age and gender for each of the four analytic samples is shown in Table 1. As anticipated, Medicaid participants were somewhat younger and disproportionately female (Health Care Financing Administration, 1999a, 1999b).

Table 1

Sample Demographic Characteristics

Characteristic	**Sample**			
	Hispanic **Medicaid**	White Medicaid	Hispanic Commercial	White Commercial
Gender (% male)	**11**	**9**	**42**	**41**
Age (% in range)				
18-24	28	22	8	4
25-34	39	40	27	18
35-44	24	27	33	29
45-54	6	7	21	32
55-64	2	3	10	15
65-74	1	1	1	**2**

Measures

The CAHPS® 1.0 Adult Core Survey includes 23 items measuring reported experiences concerning specific aspects of health plan performance, derived from a review of the literature, focus groups, cognitive interviews, and field tests (Harris-Kojetin, Fowler, Brown, & Schnaier, 1999). Twenty-one of the 23 report items are assessed using a 4-point scale, with response options ranging from *never* (1) to *always* (4). Two additional items are answered using a 2-point (*yes* = 1, *no* = 2) response format. The instrument also contains 4 items assessing global ratings of health care (i.e., health plan, quality of care, personal doctor, and specialists). Items assessing ratings differ from those measuring reports in that the former refer to global assessments rather than judgments about specific experiences. Each of the four rating items is answered using a 11-point scale with anchor points reflecting either the *worst* (0) or the *best* (10) possible care. All 27 items are answered with respect to the past six months.

Based on a priori considerations (McGee et al., 1999), we hypothesized that the 23 reporting items about specific aspects of care represented 5 underlying dimensions: *Access to Care* (5 items), *Timeliness of Care* (5 items), *Provider Communication* (6 items), *Health Plan Consumer Service* (5 items), and *Office Staff Helpfulness* (2 items). Descriptive statistics for individual items for each of the four analytic samples, organized by a priori scale membership, are shown in Table 2. Actual item content can be found in the Appendix.

Table 2

CAHPS® 1.0 Univariate Statistics for Each Sample

	Mean				Standard Deviation			
	H-Med	W-Med	W-Com	W-Com	W-Med	W-Med	W-Com	W-Com
Access to Care								
AC-1	1.20	1.21	1.18	1.17	0.46	0.40	0.38	0.38
AC-2	2.71	2.78	1.84	1.77	1.22	1.18	1.06	1.06
AC-3	2.69	2.76	2.86	2.91	1.14	1.17	1.15	0.98
AC-4	1.23	1.28	1.26	1.26	**0.43**	0.44	0.44	0.44
AC-5	3.30	3.45	3.32	3.37	1.03	0.84	0.96	0.93
Timeliness of Care								
TC-1	3.34	3.32	3.34	3.41	0.88	0.87	0.91	0.84
TC-2	2.91	2.80	2.89	2.91	1.01	1.01	1.06	1.02
TC-3	3.16	3.16	3.21	3.20	0.94	0.96	0.98	0.96
TC-4	3.12	3.14	3.15	3.15	**0.96**	0.97	1.01	0.98
TC-5	3.01	3.10	3.03	3.06	1.02	0.98	1.09	1.05
Provider Communication								
PC-1	3.51	3.38	3.45	3.41	0.74	0.81	0.77	0.75
PC-2	3.43	3.48	3.51	3.51	0.84	0.77	0.72	0.69

PC-3	3.50	3.41	3.50	3.46	0.75	0.82	0.74	0.73
PC-4	3.21	3.23	3.28	3.26		0.88	0.82	0.83
					0.89			
PC-5	3.23	3.24	3.28	3.24	0.92	0.91	0.86	0.83
PC-6	3.47	3.45	3.50	3.47	0.84	0.84	0.77	0.78
			Health Plan Consumer Service					
	3.51	3.70	3.59	3.57	0.80	0.66	0.85	0.92
HP-1								
HP-2	2.89	2.91	2.84	3.02	1.17	1.15	1.18	1.10
HP-3	2.82	2.81	2.81	2.83	1.11	1.09	1.10	1.06
HP-4	3.19	3.12	3.13	3.11		1.01	0.98	0.99
					0.98			
HP-5	3.28	3.19	3.20	3.16	0.96	0.96	0.94	0.95
Office Staff Helpfulness								
	3.58	3.59	3.64	3.64	0.71	0.69	0.66	0.62
OS-1								
OS-2	3.39	3.35	3.39	3.36	0.81	0.81	0.78	0.75
Global Ratings								
	7.98	7.76	8.06	8.02	2.58	2.58	1.97	1.83
GR-1								
GR-2	7.94	7.58	7.95	7.96	2.68	2.82	2.37	2.26
GR-3	5.38	7.68	7.94	7.83	4.20	2.42	3.44	1.91
GR-4	7.84	7.45	7.65	7.31		2.00	2.12	2.11
					2.59			

Note. H = Hispanic sample; W = white (Non-Hispanic); Med = Medicaid; Com = Commercially-insured.
Item content can be found in the Appendix.

Data Analysis

Treatment of missing data. As noted above, the CAHPS® item battery yields consumer reports (23 items) and ratings (4 items) concerning health care received during a specific time period. Although incorporation of a specific frame of reference for responses has certain advantages (Cleary, Lubalin, Hays, Short, Edgman-Levitan, & Sheridan, 1998), the use of a six-month time frame rendered some items irrelevant for certain respondents (e.g., if a respondent had not needed to see a specialist in the past 6 months, then it would not be meaningful for that person to rate the adequacy of the care received). For this reason, the use of a specific time frame resulted in a significant amount of *appropriately* missing data. The amount of missing data for the 27 items ranged from 2% to 66%, with a median of 36%. Two items were missing less than 10% of data; two items were missing between 10-25%; 15 items were missing between 25-50%, and 8 items were missing more than 50%. Because listwise deletion of cases yielded small and unique subsets each comprising no more than 7% of total possible cases, missing values were imputed for all cases using a hot-deck imputation strategy (Rubin, 1987; for an extended discussion, see Brick & Kalton, 1996). Specifically, respondents were grouped into quintiles based on the average of their responses to two survey items (i.e., GR-1: *global rating of health care provider*, and GR-4: *global rating of health plan*). For each missing item, a randomly-selected value was drawn without replacement from among respondents in the same quintile who had answered that item. The randomly-selected value was then used in lieu of the missing data point.

Overview of analytic method. Confirmatory factor analysis (CFA) of latent variables using the EQS structural equation modeling (SEM) program,

constituted the principal method of data analysis (Bentler, 1995). Within SEM, hypotheses are translated into a series of regression equations that can be solved simultaneously to generate an estimated covariance matrix. By means of various goodness-of-fit indexes including the normed (NFI) and non-normed fit (NNFI) indexes (Bentler & Bonett, 1980), the comparative fit index (CFI; Bentler, 1990), the incremental fit index (IFI; Bollen, 1989) and the root mean squared error of approximation (RMSEA; Browne & Cudeck, 1993), the estimated matrix can be evaluated against the observed sample covariance matrix to determine whether the hypothesized model is an acceptable representation of the data. In general, incremental fit indexes (i.e., NFI, NNFI, CFI, IFI) above 0.90 signify good model fit. RMSEA values lower than .08 signify acceptable model fit, with values lower than .05 indicative of good model fit (Browne & Cudeck, 1993).

Analytic plan. To determine the extent to which a common factor structure accurately characterized White and Hispanic sub-samples in both health care sectors, a series of multi-sample confirmatory factor analyses were conducted. The first analysis simultaneously modeled all four analytic samples. Two subsequent parallel series of multi-sample analyses were conducted, focusing separately on respondents receiving care within the commercial and Medicaid sectors.

Our initial hypothesized representation of the pattern of fixed and free factor loadings underlying latent constructs was based on a priori considerations. A series of models were built, each adopting a progressively more stringent set of assumptions about measurement invariance drawn from classical measurement theory (Lord & Novick, 1968). As an initial starting point, we tested a 5-factor *configurally-invariant* model (Horn, McArdle, &

Mason, 1983) in which each item was posited to contribute to one, and only one, factor. The essential assumption of configural invariance is that the pattern of fixed and free factor loadings is constant across groups. In other words, although the pattern of freely estimated and fixed factor loadings was identical across groups, the estimated parameters were free to vary.

After determining that the reporting data conformed to assumptions of configural invariance, we then assessed the degree to which the data fit the requirements of so-called *weak factorial invariance* (Meredith, 1993). The core assumption underlying weak factorial invariance is that analogous factor loading across multiple groups are invariant with respect to their contribution to the latent factor. In other words, not only is the pattern of fixed and freely estimated loadings the same, but also the actual values for the estimated loadings are constrained to be equal across groups. For example, the contribution of a specific item to its respective factor (e.g., the loading for item AC-1 on *Access to Care*) is constrained to be equal across samples. As an additional step, we then tested whether the data were consistent with assumptions underlying a cross-sample extension of *parallelism* (Lord & Novick, 1968). Parallelism is predicated on a more stringent definition of invariance that incorporates the additional restriction that error terms for analogous items are equivalent across groups. After establishing the measurement model that provided the best fit for the report data for both the commercial and Medicaid samples, we then examined whether reports regarding specific health care experiences in various domains formed empirically, as well as conceptually, distinct constructs. At issue in the latter analysis is whether variation among the items can be explained with fewer than five latent constructs.

In an additional series of analyses, we first examined the degree to which these reports of health care experiences converge, at a higher level of analysis, to form a single—indirectly measured—index reflecting global evaluations of health care. We then developed a measurement model for the four items that served as a direct index of global ratings of care and assessed the degree to which this measurement model was stable across ethnicity and service sectors. Finally, we determined the extent to which the direct and indirect methods of measuring global evaluations yielded concordant information.

Results

Factorial Invariance

As discussed above, an initial four-group multi-sample analysis examined the degree to which the structure of the report data was invariant across ethnicity and service sector. A simultaneous confirmatory factor analysis of the 23 report items assessing specific aspects of care was conducted. As noted above, the 23 report items were posited to represent 5 factors: *Access to Care, Timeliness of Care, Provider Communication, Health Plan Consumer Service,* and *Office Staff Helpfulness*. Correlation's among the 5 factors were freely estimated. In the initial multi-sample—configurally-invariant—model, all items were allowed to load freely on their hypothesized factors but were not allowed to load on other factors. The variances for each of the factors were fixed at unity to identify the model. No cross-group equality constraints were imposed. This model fit the data well, indicating that the pattern of fixed and free factor loadings was substantially identical across all four groups, chi-square (df = 880) =

4417.52, $p < .001$, NFI = .947, NNFI = .951, CFI = .957, IFI = .957, RMSEA = .016 (confidence interval = .015, .016).

Factor loadings and correlation coefficients for this model are shown in Tables 3 and 4, respectively. As shown in Table 3, with a single exception, all factor loadings were statistically significant in all four subgroups. One item (i.e., HP-1; "too many forms to fill out") failed to contribute significantly to the *Health Plan Consumer Service* factor in the Hispanic Medicaid sample. Interestingly, this item contributed only weakly, albeit significantly, to this dimension in the remaining 3 analytic samples. Nonetheless, because this loading was statistically significant in 3 of the 4 subsamples, the item was retained in all four samples to facilitate comparison across groups.

Close inspection of the sign of factor loadings also revealed a single item (i.e., AC-2; "having to see someone other than personal doctor") for which the sign differed as a function of service sector. Specifically, this item contributed positively to *Access to Care* for the commercial sector but loaded negatively on this factor within the Medicaid sector. Because this finding suggested the possibility that other minor differences might exist within service sectors, subsequent analyses focused separately on the commercial and Medicaid sectors. That is, separate two-group multisample analyses were conducted to compare responses across ethnic groups within the commercial and Medicaid sectors.

Table 3
Five-Factor Model of CAHPS® Report Items: Factor Loadings

Item and Dimension	Sample			
	H-Med	W-Med	H-Com	W-Com
Access to Care				
AC-1	-.43	-.51	-.51	-.46
AC-2	.12	.15	-.24	-.17
AC-3	.36	.39	.32	.32
AC-4	-.49	-.52	-.43	-.45
AC-5	.61	.66	.55	.58
Timeliness of Care				
TC-1	.51	.58	.54	**.54**
TC-2	.42	.52	.42	**.51**
TC-3	.54	.58	.49	**.52**
TC-4	.57	.56	.48	**.51**
TC-5	.41	.46	.34	**.27**
Provider Communication				
PC-1	.75	.81	.68	**.73**
PC-2	.65	.63	.60	**.66**
PC-3	.75	.81	.68	**.72**
PC-4	.73	.77	.65	**.69**
PC-5	.64	.66	.61	**.65**
PC-6	.50	.53	.43	**.53**
Health Plan Consumer Service (n.s)				
HP-1	.08	.20	.15	**.13**
HP-2	.28	.29	.22	**.35**
HP-3	.46	.52	.51	**.54**
HP-4	.70	.67	.69	**.68**
HP-5	.66	.68	.68	**.71**
Office Staff Helpfulness				
OS-1	.71	.72	.61	**.65**
OS-2	.79	.81	.76	**.76**

Note. All items significant at p < .01, except as noted. H = Hispanic sample; W = white (Non-Hispanic) sample; Med = Medicaid; Com = Commercially-Insured. Item content can be found in the Appendix.

Table 4
Five-Factor Model of CAHPS® Report Items: Correlations among Factors

Factor and Sample	Factor I	II	III	IV	V
I. Access to Care					
H-Med	- - -	.72	.77	.64	.62
C-Med	- - -	.84	.83	.69	.70
H-Com	- - -	.85	.89	.71	.69
C-Com	- - -	.78	.80	.63	.63
II. Timeliness of Care					
H-Med		- - -	.80	.62	**.82**
C-Med		- - -	.78	.63	**.77**
H-Com		- - -	.80	.62	**.70**
C-Com		- - -	.73	.57	**.71**
III. Provider Communication					
H-Med			- - -	.60	**.83**
C-Med			- - -	.52	**.81**
H-Com			- - -	.56	**.79**
C-Com			- - -	.52	**.76**
IV. Health Plan Consumer Service					
H-Med				- - -	**.48**
C-Med				- - -	**.47**
H-Com				- - -	**.54**
C-Com				- - -	**.45**
V. Office Staff Helpfulness					
H-Med					- - -
C-Med					- - -
H-Com					- - -
C-Com					- - -

Note. All correlations significant at p < .001.

As an initial reference point for subsequent two-group analyses, configural invariance was replicated within both the commercial and Medicaid sectors. Not surprisingly, as shown in Table 5, this model (Model 1) fit the data well for both the commercial and Medicaid sectors. In the next step, cross-group equality constraints were imposed on analogous factor loadings to assess the degree to which the data conformed to weak factorial invariance. As shown in Table 5 (Model 2), this model also fit the data well, although a chi-square difference test indicated that the initial model provided a statistically more adequate account of the data in both the commercial, chi-square difference (df = 23) = 40.07, $p < .001$; and Medicaid sectors, chi-square difference (df = 23) = 40.19, $p < .001$.

Post hoc inspection of the Lagrange Multiplier univariate tests for releasing constraints indicated that 18 of 23 cross-sample equality constraints in the commercial sector were statistically reasonable. Releasing equality constraints on the remaining 5 pairs of factor loadings, involving items AC-1, AC-2, TC-5, PC-6, and HP-2, resulted in a model (Model 3) that fit the data as well as the initial model, chi-square difference (df = 18) = 11.30, $p > .05$. Similar post hoc inspection of the Lagrange Multiplier univariate tests for releasing constraints in the Medicaid sector indicated that 20 of 23 equality constraints were statistically reasonable. Releasing equality constraints on the remaining 3 pairs of factor loadings, involving items PC-1, PC-2, and PC-3, resulted in a model (Model 3) that fit the data as well as the initial model, chi-square difference (df = 20) = 16.64, $p > .05$.

Thus, these results indicate that the structure of the reports of care was substantially similar across Hispanic and white sub-samples within the

commercial and Medicaid health care sectors. That is, the essential pattern of fixed and free parameter estimates were identical within sectors, irrespective of ethnicity. The data also met requirements of weak factorial invariance in the strictest sense that all absolute model fit criteria were quite good, indicating that factor loadings were of comparable magnitude across samples. Close inspection did reveal, however, that releasing constraints on a few pairs of loadings did result in improved model fit. Specifically, 18 of 23 cross-group pairs of factor loadings in the Medicaid sector and 20 of 23 cross-group pairs of factor loadings in the commercial sector were equivalent. Thus, these results indicate that reports of care were substantially similar across ethnic groups within the two health care sectors with respect to the pattern of fixed and free loadings and—in most instances—the actual magnitude of factor loadings.[3]

Table 5
Multi-group Model Fit Indexes

Model Number and (Description)	Chi-Square	df	NFI	NNFI	CFI	IFI	RMSEA	(CI)
Commercial Samples								
1. Configurally-Invariant (No cross-group equality constraints)	3137.89	440	.930	.930	.939	.939	.026	(.025, .026)
2. Basic Factorial Invariance (Cross-group equality constraints imposed on analogous factor loadings)	3177.96	463	.924	.933	.939	.939	.026	(.025, .026)
3. Modified Basic Invariance (Cross-group constraints released on five pairs of loadings)	3149.19	458	.930	.933	.940	.940	.026	(.025, .026)
4. Higher Order Model	3413.09	468	.924	.928	.934	.934	.026	(.025, .026)
5. Higher Order Model with Global Ratings Factor Incorporated	6852.22	657	.896	.898	.905	.905	.032	(.032, .033)
Medicaid Samples								
1. Configurally-Invariant (No cross-group equality constraints)	1279.60	440	.967	.975	.978	.978	.018	(.017, .019)
2. Basic Factorial Invariance (Cross-group equality constraints imposed on analogous factor loadings)	1319.79	463	.966	.975	.978	.978	.017	(.016, .019)
3. Modified Basic Invariance (Cross-group constraints released on three pairs of loadings)	1296.29	460	.966	.976	.978	.978	.017	(.016, .018)
4. Higher Order Model	1650.44	470	.957	.967	.969	.969	.020	(.019, .021)
5. Higher Order Model with Global Ratings Factor Incorporated	3800.31	658	.926	.934	.938	.938	.028	(.027, .029)

Notes: NFI=Normed Fit Index; NNFI=Non-Normed Fit Index; CFI=Comparative Fit Index; IFI=Incremental Fit Index; RMSEA=Root Mean Squared Error of Approximation; CI=Confidence Interval for RMSEA.

Factor Independence

As shown in Table 4, and consistent with much prior research on the measurement of satisfaction with health care (e.g., Marshall et al., 1993), correlations among all five domain-specific factors in each analytic sample were statistically significant. Correlations between analogous pairs of factors were of generally comparable magnitude across the four subsamples, indicating substantial cross-group stability. Perhaps of greater importance, 10 of 40 correlations equaled or exceeded 0.80 in magnitude, consistent with the possibility that one or more of these factors are redundant.

To examine whether fewer than five factors were needed to account for variation among the data, alternative factor structures were examined. Within each service sector, 10 alternative models were tested in which the correlation between a different pair of factors was fixed at 1.0—rather than freely estimated—to reflect the possibility that the two factors tap a single dimension. In all other respects, the alternative models were identical to one another. Fit indexes and chi-square difference tests revealed that the five-factor model was statistically superior to all other models.[4] That is, although substantially correlated in some instances, the five dimensions are, in the strictest sense, empirically differentiable with each providing unique information.

Higher Order Structure of Consumer Evaluations

The substantial covariation of the five primary reporting dimensions implies that these dimensions may converge, at a more abstract level of analysis, to form a single overarching domain reflecting overall judgments regarding heath care. To examine the tenability of this higher order model,

two additional multi-sample models were evaluated using the commercial and Medicaid samples. In both models, each of the five domain-specific reporting dimensions were represented as stemming from a single broad evaluative domain. To fix the scale of these models, the variance of the higher order factor was fixed at unity, as was a single factor loading for each of the primary reporting dimensions. All other first order factors were allowed to load freely on the higher order factor.

Although chi-square difference tests indicate that the original five-factor model offered a statistically more adequate account of the data for both the commercial [chi-square (df = 10) = 269.90, $p < .001$] and Medicaid samples [chi-square (df = 10) = 377.16, $p < .001$], both of the higher order models fit the data well in an absolute sense (see Model 4, Table 5). For each of the four subgroups, higher order factor loadings, reflecting the contribution of each primary dimension to overall evaluations of care, were statistically significant at $p < .001$.[5]

Concordance Between Methods of Assessing Overall Health Care

For the specific purpose of examining the correspondence between direct and indirect methods of measuring overall judgments concerning health care, both methods were modeled simultaneously. Prior to assessing the comparability of information obtained using the two methods, separate multi-sample models of global ratings were developed using the 4 items discussed above (i.e., ratings of health plan, quality of care, personal doctor, and specialists). In each instance, the model was identified by fixing the variance of the factor to unity. Within each sample, the four global ratings items were allowed to load freely on a single factor reflecting overall

appraisals of care. Cross-group equality constraints were imposed on analogous factor loadings to determine whether the model met the assumptions underlying weak factorial invariance (Meredith, 1993).

This model fit the data well for both the commercial [chi-square (df = 8) = 70.38, p < .001, NFI = .992, NNFI = .989, CFI = .993, IFI = .993, RMSEA = .029] and Medicaid samples [chi-square (df = 8) = 88.01, p < .001, NFI = .989, NNFI = .985, CFI = .990, IFI = .990, RMSEA = .034]. As expected, all factor loadings were statistically significant. Close inspection revealed, however, that some of the cross-group equality constraints on factor loadings were not statistically tenable. In the commercial sample, one of four equality constraints, involving item GR-1, was subsequently released as were two of the four constraints, involving items GR-1 and GR-4, in the Medicaid sample. After releasing these constraints, the final models proved slightly superior to the more constrained models for both the commercial [chi-square difference (df = 1) = 4.40, p < .05] and Medicaid samples [chi-square difference (df = 2) = 24.24, p < .001].

To evaluate the concordance between direct and indirect methods of measuring global ratings of health care, we included both methods in a single model. The direct method of measuring global ratings of care was modeled as described in the preceding paragraph. As discussed previously, global ratings of care were operationalized *indirectly* as the covariation of the five consumer reporting dimensions: *Access to Care*, *Timeliness of Care*, *Provider Communication*, *Health Plan Consumer Service*, and *Office Staff Helpfulness*. The correlation between the two methods of measuring global satisfaction was freely estimated.

As shown in Table 5, this model (Model 5) proved to be a good fit to the data in both the commercial and Medicaid samples. The magnitudes of the correlations between direct and indirect methods of measuring global ratings of health care were quite high in all four samples (H-Com: $r = 0.95$; C-Com: $r = 0.98$; H-Med: $r = 0.98$; and C-Med: $r = 0.97$), attesting to the apparent concordance of information yielded using direct and indirect methods. Once again, however, chi-square difference tests revealed that constraining the correlation between the two methods to 1.0 led to significant degradation in the fit of the models for both the commercial [chi-square difference ($df = 2$) = 45.25, $p < .001$] and Medicaid samples [chi-square difference, $df = 2$) = 55.77, $p < .001$]. Thus, although substantially correlated, the direct and indirect methods of measuring global satisfaction are, in the strictest statistical sense, empirically distinguishable.

Discussion

This research employed multi-sample structural equation modeling to examine the CAHPS® 1.0 Core Survey in Hispanic and white health care consumers receiving care in either commercial or Medicaid service sectors. The study had three objectives. The first goal was to examine the extent to which the factor structure of the instrument conformed to a priori expectations. As anticipated (McGee et al., 1999), results demonstrate that the CAHPS® 1.0 Core Survey does, in fact, measure consumer reports of experiences in five key spheres of health care: *Access to Care, Timeliness of Care, Provider Communication, Health Plan Consumer Service,* and *Office Staff Helpfulness.*

The second major aim of this study was to examine the degree to which the CAHPS® 1.0 Core Survey instrument factor structure is invariant across individuals of differing ethnicity as well as persons receiving care in different health care settings. Three models of invariance were evaluated, each of which incorporated the assumptions of the preceding definition, and thus were progressively more stringent: cross-sample stability in the pattern of fixed (i.e., zero) and free (i.e., non-zero) factor loadings; cross-sample equivalence in the actual magnitude of analogous factor loadings; and cross-sample equivalence in the magnitude of analogous error variances.

Results revealed substantial evidence of factorial invariance. Notably, in all four samples, the CAHPS® 1.0 Core Survey showed configural stability with respect to the pattern of fixed and free factor loadings. Stated differently, the pattern of zero and non-zero factor loadings for the instrument was identical, irrespective of either the ethnicity of respondents or the health service sector in which they received care. Analyses also revealed evidence of considerable stability in the actual magnitude of analogous factor loadings across Hispanics and whites receiving care within a common health service sector. That is, the size of analogous factor loadings for Hispanics and whites receiving care with the commercial sector were, in most instances, virtually identical in magnitude. A similar finding held true for persons of different ethnicity receiving care within the Medicaid sector. The most stringent assumption of equivalence across analogous error variances was not, however, borne out by the data.

In arriving at these conclusions, we note that structural equation models positing equality of cross-sample loadings for analogous items provided a good fit to the data. At the same time, an examination of individual equality constraints imposed on specific item pairs revealed a small number in each model that were statistically untenable. Stated precisely, 20 of 23 pairs of analogous factor loadings were equivalent in the commercial sector as were 18 of 23 pairs in the Medicaid sector. Thus, in the strictest sense, the data did not meet criteria for equality with respect to each pair of analogous factor loadings. Whereas the five non-equivalent items in the Medicaid sector were dispersed across reporting dimensions, it is perhaps noteworthy for future research that the three non-equivalent items in the commercial sector were confined to the factor assessing *Provider Communication*. We also note that relevant analyses were conducted within—rather than across—service sectors due to an inconsistency across Medicaid and commercial samples with respect to the sign of the loading for a single item, AC-2. Thus, these findings do not directly attest to the equality of factor loadings of persons receiving care in different health service sectors.

The third broad objective of this study was to examine the extent of concordance between two alternative methods of assessing overall consumer judgments regarding health care services. As discussed above, the CAHPS® 1.0 provides two distinct methods of assessing global judgments. Overall impressions can be quantified directly by asking consumers to rate their care on a small number of key domains (e.g., rating of health care provider). Global judgments can also be summarized by aggregating across separate reports of specific experiences in distinct health care realms (e.g., access

to care). Although the two methods have counterbalancing strengths and weaknesses, these data indicate from a practical standpoint that the alternative approaches yield highly comparable information. Thus, a small number of items assessing consumer ratings of health care may offer adequate precision as a brief, minimally burdensome, stand-alone measure of overall judgments regarding care. Of course, where more comprehensive assessment of specific consumer experiences is desired, reference to overall ratings alone is not sufficient.

While providing empirical support for several key assumptions underlying the development of the CAHPS® 1.0 Core Survey, these results do suggest, however, that minor modifications might be possible to further refine the instrument. In particular, a single item reflecting consumer reports regarding burden associated with completing health plan paperwork (i.e., HP-1) proved to be a rather weak marker of *Health Plan Consumer Service*. This item did not load significantly for the sample composed of Hispanics receiving Medicaid care and contributed only modestly, albeit significantly, to this factor for the three other subsamples. Thus, although at first glance, the poor performance of this item might be partially attributed to the lack of salience of paperwork within the Medicaid health service sector, the general pattern of findings suggests that the item may not be a good marker of *Health Plan Consumer Service*. Similarly, a second item, i.e., AC-2, emerged as a relatively weak marker of *Access to Care* in all samples. This item assessed the reported frequency with which respondents in the commercial and Medicaid sectors were able to see their preferred health care practitioner. Moreover, this item contributed positively to the *Access to Care* construct in the commercial sector but

negatively in the Medicaid sector. The latter finding is not entirely unexpected inasmuch as persons receiving care within the Medicaid sector may have less control over their choice of care provider than do persons receiving care within the commercial sector. Although both items in question were retained for the present purposes, further research is warranted to further evaluate their utility.

Additional research is also needed to examine the extent to which CAHPS® 1.0 Core Survey responses might vary as a function of acculturation or language in which the survey was completed. In the current study, no data were available concerning whether a given individual of Hispanic descent completed an English or a Spanish version of the survey. Thus, although these data provide evidence that the instrument performs quite similarly across persons of differing ethnicity, future research is needed to clarify possible differences in the responses of persons of Hispanic origin completing different language versions of the instrument.

In conclusion, these findings indicate that the CAHPS® 1.0 instrument generally performs in a similar fashion irrespective of the ethnic background of respondents (i.e., white vs. Hispanic) or the health service sector (i.e., commercial vs. Medicaid) within which they received care. As such, these data are consistent with recent research (Morales, Reise, & Hays, in press), suggesting that valid comparisons regarding judgments about health services can be drawn across whites and Hispanics. The latter point is important insofar as some prior research has suggested that whites and Hispanics may respond differently to questions about health services (e.g., Aday, Chiu, & Anderson, 1980; Hayes & Baker, 1998). Additional research is needed to

determine the suitability of the instrument for use with persons from other racial or ethnic groups.

References

Aday, L. A., Chiu, G. Y., & Andersen, R. (1980). Methodological issues in health care surveys of the Spanish heritage population. *American Journal of Public Health, 70,* 367-374.

Allen, H. M., & Rogers, W. H. (1997). The Consumer Health Plan Value Survey: Round two. *Health Affairs, 16,* 156-166.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M. (1995). *EQS Structural Equations Program manual.* Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17,* 303-316.

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5,* 215-238.

Brown, J. A., Nederend, S. E., Hays, R. D., Short, P. F., & Farley, D. O. (1999). Special issues in assessing care of Medicaid recipients. *Medical Care, 37,* MS79-MS88.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230-239.

Cleary, P. D. (1999). The increasing importance of patient surveys. *British Medical Journal, 319,* 720-721.

Cleary, P. D., & Edgman-Levitan, S. (1997). Health care quality: Incorporating consumer perspectives. *Journal of the American Medical Association, 278,* 1608-1612.

Cleary, P. D., Lubalin, J., Hays, R. D., Short, P. F., Edgman-Levitan, S., & Sheridan, S. (1998). Debating survey approaches. *Health Affairs, 17,* 265-266.

Fowler, F. J., Gallagher, P. M., & Nederend, S. (1999). Comparing telephone and mail responses to the CAHPS survey instrument. *Medical Care, 37,* MS41-MS49.

Hall, J. A., Roter, D. L., & Katz, N. R. (1988). Meta-analysis of correlates of provider behavior in medical encounters. *Medical Care, 26,* 657-675.

Hall, J. A., Roter, D. L., & Milburn, M. A. (1999). Illness and satisfaction with medical care. *Current Directions in Psychological Science, 8,* 96-99.

Hays, R. D., Shaul, J. A., Williams, V. S. L., Lubalin, J. S., Harris-Kojetin, L. D., Sweeny, S. F., & Cleary, P. D. (1999). Psychometric properties of the CAHPS 1.0 Survey Measure. *Medical Care, 3,*MS22-MS31.

Hayes, R. P., & Baker, D. W. Methodological problems in comparing English-speaking and Spanish-speaking patients' satisfaction with interpersonal aspects of care. *Medical Care, 36,* 230-236.

Health Care Financing Administration (1999a). *Medicaid recipients and vendor payments by age.* Table 6. HCFA-2082 Report: Center for Medicaid and State Operations. Department of Health and Human Services, Washington, D.C.

Health Care Financing Administration (1999b). *Medicaid recipients as a percentage of population by sex.* Table 10. HCFA-2082 Report: Center for Medicaid and State Operations. Department of Health and Human Services, Washington, D.C.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant?: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 4,* 179-188.

Harris-Kojetin, L. D., Fowler, F. J., Brown, J. A., Schnaier, J. A., & Sweeny, S. F. (1999). The use of cognitive testing for developing and evaluating CAHPS 1.0 Core survey items. *Medical Care, 37,* MS10-MS21.

Kippen, L. S., Strasser, S., & Joshi, M. (1997). Improving the quality of the NCQA (National Committee for Quality Assurance) Annual Member Health Care Survey Version 1.0. *American Journal of Managed Care, 3,* 719-730.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marshall, G. N., Hays, R. D., & Mazel, R. (1996). Health status and satisfaction with medical care: Results from the Medical Outcome Study. *Journal of Consulting and Clinical Psychology, 64,* 380-390.

Marshall, G. N., Hays, R. D., Sherbourne, C. D., & Wells, K. B. (1993). The structure of patient satisfaction with outpatient medical care. *Psychological Assessment, 5,* 477-483.

McGee, J., Kanouse, D. E., Sofaer, S., Hargraves, J. L., Hoy, E., & Kleimann, S. (1999). Making survey results easy to report to consumers: How reporting needs guided survey design in CAHPS™. *Medical Care, 37,* MS32-MS40.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525-543.

Morales, L. S., Reise, S. P., & Hays, R. D. (In press). Evaluating the equivalence of health care ratings by whites and Hispanics. *Medical Care.*

National Committee for Quality Assurance. (1998). Accreditation '99: Standards for the accreditation of managed care organizations, 11-17.

Newcomer, R., Preston, S., & Harrington, C. (1996). Health plan satisfaction and risk of disenrollment among social/HMO and fee-for-service recipients. *Inquiry, 33,* 144-154.

Penchansky, R., & MacNee, C. (1994). Initiation of medical malpractice suits: A conceptualization and test. *Medical Care, 32,* 813-821.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Schlesinger, M., Druss, B., & Thomas, T. (1999). No exit? The effect of health status on dissatisfaction and disenrollment for health plans. *Health Services Research, 34,* 547-576.

Schnaier, J. A., Sweeny, S. F., Williams, V. S. L., Kosiak, B., Lubalin, J. S., Hays, R. D., & Harris-Kojetin, L. D. (1999). Special issues addressed in the CAHPS Survey of Medicare managed care beneficiaries. *Medical Care, 37,* MS69-MS78.

Shaul, J. A., Fowler, F. J., Zaslavsky, A. M., Homer, C. J., Gallagher, P. M., & Cleary, P. D. (1999). The impact of parents reporting about their own and their children's experiences with health insurance plans. *Medical Care, 37,* MS59-MS68.

Vaccarino, J. M. (1977). Malpractice: The problem in perspective. *Journal of the American Medical Association, 238,* 861-863.

Williams, B. (1994). Patient satisfaction: A valid construct? *Social Science & Medicine, 38,* 509-516.

Zaslavsky, A. M., Beaulieu, N. D., Landon, B. E., & Cleary, P. D. (2000). Dimensions of consumer-assessed quality of Medicare managed care health plans. *Medical Care,* 38, 162-74.

Author Notes

The Consumer Assessment of Health Plans Study (CAHPS®) is funded by the Agency for Health Care Quality and Research (AHQR) and the Health Care Financing Administration through cooperative agreements with Harvard Medical School, RAND, and the Research Triangle Institute. User support is provided through a contract with Westat. Additional information about the study can be obtained by calling the AHQR Clearinghouse at 800-358-9295. The authors thank project officers Chris Crofton, Chuck Darby, and Beth Kosiak active participation and thoughtful feedback throughout the project.

Footnotes

[1]We note that a recent exploratory factor analysis of the CAHPS® 1.0 has examined the associations among dimensions using health plans, rather than individuals, as the unit of analysis (Zaslavsky et al., 2000). In other words, the responses of individuals receiving care within a given health plan were averaged, and the resulting plan level ratings were factor analyzed. Although this approach addresses important research questions, the resulting factor structure would not necessarily correspond to that which would emerge from an individual level analysis.

[2]This sample was randomly drawn from a total sample of 14,643 cases. Current versions of EQS for Windows (Bentler, 1995) can only accommodate sample sizes of less than 8,000 cases.

[3]Additional tests were conducted to assess whether these data fit the requirements of cross-group parallelism. As would be expected, imposition of additional equality constraints on cross-group error terms for both the commercial and Medicaid sectors revealed that the data did not satisfy these more stringent criteria.

[4]Complete fit information for all models is available from the first author.

[5]Higher order factor loadings for the four samples are listed below. Hispanic Medicaid: AC = 0.87, TC = 0.88, PC = 0.90, HP = 0.64, OS = 0.85. white Medicaid: AC = 0.91, TC = 0.90, PC = 0.91, HP = 0.63, OS = 0.85; Hispanic Commercial: AC = 0.95, TC = 0.83, PC = 0.86, HP = 0.70, OS = 0.75; white commercial: AC = .95, TC = 0.87, PC = 0.87, HP = 0.70, OS = 0.75.

APPENDIX

CAHPS® 1.0 Survey Items

Access to Care

AC-1 *With the choices your health insurance plan gives you, was it easy to find a personal doctor or nurse you are happy with?*[i]

AC-2 *How often did you have to see someone else when you wanted to see your personal doctor or nurse?*

AC-3 *How often did you see a specialist when you thought you needed one?*

AC-4 *Was it always easy to get a referral when you needed one?*

AC-5 *How often did you get the tests or treatment you thought you needed?*

Timeliness of Care

TC-1 *How often did you get the medical help or advice you needed when you phoned the doctor's office or clinic during the day Monday through Friday?*

TC-2 *How often did you get the help during the day Monday through Friday without a long wait?*

TC-3 *When you tried to be seen for an illness or injury, how often did you see a doctor or other health professional as soon as you wanted?*

TC-4 *When you needed regular or routine healthcare, how often did you get an appointment as soon as you wanted?*

TC-5 *How often did you wait in the doctor's office or clinic more than 30 minutes past your appointment time to see the person you went to see?*

Provider Communication

PC-1 *How often did doctors or other health professionals listen carefully to you?*

PC-2 *How often did doctors or other health professionals explain things in a way you could understand?*

PC-3 *How often did doctors or other health professionals show respect for what you had to say?*

PC-4 *How often did doctors or other health professionals spend enough time with you?*

PC-5 *How often did doctors or other health professionals know what you thought they should know about your medical history?*

PC-6 *How often were you involved as much as you wanted in these decisions about your health?*

APPENDIX

CAHPS® 1.0 Survey Items

(cont.)

Health Plan Consumer Service

HP-1 *How often did you have more forms to fill out for your health insurance plan than you thought was reasonable?*

HP-2 *How often did your health insurance plan deal with approvals or payments without taking a lot of your time and energy?*

HP-3 *How often were calls to the health insurance plan's customer service taken care of without a long wait?*

HP-4 *How often did you get all the information or other help you needed when you called the health insurance plan's customer service?*

HP-5 *How often were the people at the health insurance plan's customer service as helpful as you thought they should be?*

Office Staff Helpfulness

OS-1 *How often did office staff at a doctor's office or clinic treat you with courtesy and respect?*

OS-2 *How often were office staff at a doctor's office or clinic as helpful as you thought they should be?*

Global Rating

GR-1 *How would you rate your personal doctor or nurse now?*

GR-2 *How would you rate the specialist (you saw most often)?*

GR-3 *How would you rate all of your health care (from all doctors and other health professionals)?*

GR-4 *How would you rate your health insurance plan now?*

[1]Except as follows, all items were answered on a four-point scale (1 = never, 2 = sometimes, 3 = usually, 4 = always): Items AC-1 and AC-2 (2-point scale, 1 = yes, 2 = no); items GR-1 to GR-4 (11-point scale; 0 = worst possible, 10 = best possible).

9. Assessing Racial and Ethnic Differences in Patient Evaluations of Care: Summary and Implications for Health Policy and Future Research.

Introduction

This thesis has two central themes, racial/ethnic differences in consumer evaluations of care and methodological concerns related to making those comparisons. The first theme, racial/ethnic differences in patient evaluations of care, is addressed in chapters 4 to 6 of this thesis. The second theme, methodological concerns related to making racial/ethnic comparisons of evaluations of care, is addressed in Chapters 2, 3, 7.

The two principal purposes of this chapter are to summarize the results from the main substantive and methodological chapters of this thesis and to present the implications of those results for health policy and future research. With these objectives in mind, the remainder of this chapter is organized into three sections. In the next section, the main results from the substantive chapters are summarized. In the third section of this chapter, the main results from the methodological chapters are summarized. In the fourth section, the implications of the main substantive and methodological results for health policy and future research are presented.

Summary of Substantive Results

Chapter 4: Are Latinos Less Satisfied with Communication by Health Care Providers?

In Chapter 4, patient ratings of communication by medical providers among non-Hispanic whites, Hispanics who completed their surveys in English, and Hispanics who completed their surveys in

Spanish are compared. The survey data was collected from patients receiving care from the United Medical Group Association (UMGA), an association of physician groups located in the western United States. Of the 6,211 surveys analyzed in this study, 713 were completed by Hispanics, of which 181 were completed in Spanish. The overall survey response rate for the study was 59%.

Patient's evaluation of communication by medical providers was assessed using a 5-item scale. All five items were administered using an identical 7-point response format (*Very Poor, Poor, Fair, Good, Very Good, Excellent,* and *The Best*) along with the option, *Does Not Apply to Me*. The five items asked survey respondents to rate the following aspects of communication: medical staff listening to what you have to say; answers to your questions; explanations about prescribed medications; explanations about medical tests and procedures; and reassurance and support from your doctor and support staff.

Each rating question was modeled separately using ordinal logistic models. The main independent variables in the study were ethnicity (white versus Hispanic) and language of survey response (English versus Spanish). These two variables were combined to produce three mutually exclusive categories of respondents: whites who responded to the English survey version (whites), Hispanics who responded to the English survey version (E-Hispanics), and Hispanics who responded to the Spanish survey version (S-Hispanics).

The case-mix variables included in all models were age and gender. Other case-mix variables including health status, education, and a Spanish language response variable were included in preliminary models. However, because the results from models using these

additional case-mix variables did not differ from models that only included age and gender, the more parsimonious models were selected for presentation. Huber and White corrected standard errors were computed to correct for intra physician group clustering of survey respondents.

The five final models showed similar patterns of results. Across all five models, S-Hispanics rated communication with their medical providers the lowest, followed by E-Hispanics and whites. Specifically, there were statistically significant differences in ratings between whites and S-Hispanics across all models at the 0.05 level. There were also statistically significant differences between S-Hispanics and E-Hispanics across all questions at the 0.05 level. In addition, there were statistically significant differences between whites and E-Hispanics on three of five questions.

These results suggest that Hispanics who primarily communicate in Spanish are at increased risk for poor communication with their medical providers and subsequent sub-optimal outcomes of care. For example, Spanish speaking patients receiving unsatisfactory explanations about how to take their prescribed medications may inadvertently take them inappropriately, resulting in less than optimal outcomes including medication toxicities, regardless of whether or not the prescriptions were technically appropriate.

The results from this study are consistent with prior research on quality of care for Hispanic patients. Baker et al. (1996) found poor communication between Spanish speaking patients and their medical providers in emergency room settings (Baker, Parker, Williams, Coates, & Pitkin, 1996). Perez-Stable et al. (1999) found worse outcomes of care among Spanish speaking Hispanic patients with non-Spanish speaking

providers than English speaking Hispanic patients (Perez-Stable, Naapoles-Springer, & Miramontes, 1997). The results presented in Chapter 4 are consistent with language barriers faced by all patients with poor English language skills accessing the medical care system (Woloshin, Bickell, Schwartz, Gany, & Welch, 1995b).

Chapter 5: Differences in CAHPS® Adult Survey Ratings and Reports by Race and Ethnicity: An Analysis of the National CAHPS® Benchmarking Data 1.0.

In Chapter 5, a study of adult racial/ethnic differences in consumer evaluations of care using the National CAHPS® Benchmarking Database 1.0 (NCBD 1.0) is presented. The NCBD 1.0 is an aggregation of CAHPS® 1.0 survey results collected by health plans located across the United States. The NCBD project is administered by Quality Measurement Advisory Service (QMAS) with finding from the Agency for Healthcare Quality and Research (AHRQ). Health plans administering the CAHPS® surveys were asked to voluntarily contribute their survey results to the NCBD project for purposes of benchmarking and research. Both adult and child survey results from Medicaid and commercial settings are included in NCBD 1.0.

Although different versions of the CAHPS® surveys have been developed for adults and children, Medicaid and commercial settings, and mail and telephone administration, all versions of the surveys contain the a core set of evaluations. These evaluations can be divided into reports and ratings. Reports ask survey respondents about the frequency with which certain events take place. In the CAHPS® paradigm, there are five reporting domains, each addressed by several

questions. The five reporting domains are access to needed care (four questions), provider communication (four questions), office staff helpfulness (three questions), promptness of needed care (four questions), and health plan customer service (two questions). With the exception of two questions which are asked using a *Yes/No* format, all questions in the reports domains are asked using a *Never, Sometimes, Usually, Always* format.

Ratings, in contradistinction to reports, are single items that ask consumers to make summary evaluations of various aspects of care. The CAHPS® 1.0 surveys include four rating items, which ask consumer to rate their personal medical provider, specialists, health care, and health plan.

All four rating questions are administered using an 11-point numeric scale ranging from 0 to 10. Statements like "The Best Possible Personal Doctor" and "The Worst Possible Personal Doctor" anchor the extremes of each scale.

Each report and rating was modeled separately using OLS regression. The main independent variables were indicators of race/ethnicity. Based on two questions on ethnicity (Hispanic or non-Hispanics) and race (White, Black/African American, Asian/Pacific Islander, American Indian/Native Alaskan, Other Race, Multiple Races), seven mutually exclusive respondent categories were created: Hispanics regardless of race and non-Hispanic White, Black/African American, Asian/Pacific Islander, American Indian/Native Alaskan, other or multiple race, and missing race data. Other independent variables included in all models were age, gender, education, and health status.

Because of the skewness of the distributions of the dependent variables, a linear transformation was used to achieve greater normality. However, the regression results varied little whether or not the transformed dependent variables were used, thus only results based on the untransformed variables were presented. Huber and White corrected standard errors were computed to adjust for intra-plan clustering. Weights were derived to adjust for differences in response rates across health plans.

Overall, whites reported better care than other racial/ethnic groups. Compared to whites, Hispanics reported worse access to care, promptness of care, and health plan customer service. Asian/Pacific Islanders reported worse care than whites across all five reporting domains. Persons in the Other/Multiracial category reported worse access to care, promptness of care, provider communication, and health plan customer than whites. Persons in the Missing category also reported worse care than whites across most reports domains. American Indians/Alaskan natives were the only group to report care similar to whites.

African Americans, unexpectedly, reported better care than whites in two domains: provider communication and office staff helpfulness. Interestingly, the office staff reporting composite includes a question about the frequency with which the customer was treated with courtesy and respect. If African Americans felt they had experienced more racial discrimination by office staff than whites, it is likely that this question would have captured that experience.

Ratings revealed fewer differences between whites and other race/ethnic groups than reports, and in some cases, ratings

contradicted reports. For example, Hispanics rated their personal doctor, specialists and their health care similar to whites, and rated their health plans higher than whites despite reporting worse access to care and promptness of care than whites. Asian/Pacific Islanders gave ratings similar to whites across all five rating questions despite reporting worse care across all report domains. American Indian/Alaskan Natives rated their personal doctors and specialists lower than whites despite similar reports about care. Persons in the Multiracial/Other and Missing categories rated their care lower than whites, consistent with worse reports about care given by both groups compared with whites. African Americans rated their health plans higher than whites, consistent with better reports about care than whites.

Chapter 6: Racial and Ethnic Differences in Parents' Assessments of Pediatric Care in Medicaid Managed Care.

In Chapter 6, a study of consumer evaluations of pediatric care based on the NCBD 1.0 is presented. Although this study is based on the same database as the previous study, several important differences in the data resulted in the decision to conduct separate studies. First, pediatric CAHPS® 1.0 data is collected by proxy; the adult parent or guardian of a child respondent is asked to complete the survey. Second, only the child versions of the CAHPS® 1.0 surveys collect information about the main language spoken at home. This variable enabled further analysis of the child data not possible with the adult data. Third, it was the judgment of the investigators that the amount of information generated by simultaneous presentation of the adult and child results would be overwhelming.

As in the analysis of the adult NCBD 1.0 data, the core reports and ratings were chosen as the main dependent variables in this study. Also as in the adult study, seven racial/ethnic categories served as the main independent variables in this study. The categories included: Hispanics regardless of race, and non-Hispanic Whites, Black/African Americans, Asian/Pacific Islanders, American Indian/Native Alaskans, other or multiple races, missing race data. The adult proxy's race/ethnicity was used rather than the child's race/ethnicity because the adult proxy's completed the surveys.

In addition, Hispanics and Asian/Pacific Islanders were subcategorized by language spoken at home. Specifically, Hispanics were subcategorized as speaking Spanish (Hispanic-S) or English (Hispanic-E) at home and Asian/Pacific Islanders were categorized as speaking English (Asian-E) or another language (Asian-O) at home. A small group of Hispanics, who did not indicate what language they spoke at home (N=26), was dropped from the study. Asian/Pacific Islanders who did not indicate what language they spoke at home formed a third Asian/Pacific Islander group (Asian-M).

The case-mix variables used in this analysis were similar to those used in the study of adult evaluations. They included the proxy's age, the proxy's gender, the proxy's education, and the child's health status.

Separate OLS models were used for each dependent variable. Because results based on transformed and untransformed dependent variables were similar, only results based on the untransformed dependent variables are presented. Huber and White corrected standard errors were computed to correct for intra-plan clustering and

analytical weights were used to account for variations in response rates across plans.

The first main result of this study is that whites reported better experiences with care and higher ratings of care than members of other race/ethnic groups. Hispanics-S and Asian-O reported worse access to care, promptness of care, provider communication, staff helpfulness, and health plan customer service than whites. Asian-M reported worse promptness of care, provider communication, and staff helpfulness. African Americans reported worse access to care, promptness of care, and health plan customer service than whites. American Indians reported worse access to care, promptness of care, provider communication, and health plan customer service than whites. Finally, persons in the Missing and Multiracial/Other categories reported worse access to care, promptness of care, provider communication, and health plan customer service than whites.

Ratings of care closely mirrored reports of care, with whites rating their experiences with providers and services more highly than other racial/ethnic groups. As in the study of the adult data, there were fewer differences between whites and the other groups based on ratings than reports. However, unlike the adult study, there were no intra-racial/ethnic group inconsistencies between reports and ratings (e.g., worse reports but higher ratings).

The second major finding in this study is that language barriers may account for the lower ratings and worse reports given by Hispanics and Asians compared with whites. In all comparisons between whites and Hispanics, Hispanics who reported speaking English at home gave reports and ratings of care similar to whites, while Hispanics who reported

speaking Spanish at home gave lower reports and ratings than whites. Similarly, in all comparisons between whites and Asians, Asians who reported speaking English at home gave reports and ratings of care similar to whites, while Asians who reported speaking a language other than English at home gave lower reports and ratings than whites.

Summary of Methodological Results

Chapter 3: Readability of CAHPS® 2.0 Child and Adult Core Surveys

In Chapter 3, a readability analysis of the Adult and Child, English and Spanish versions of the CAHPS® 2.0 surveys is presented. Literacy level is an attribute of individuals, and refers to the reading ability level of an individual. Readability, on the other hand, is an attribute of written materials, and refers to the reading ability required to comprehend a text.

Concern about low literacy among potential respondents to self-administered surveys such as CAHPS®, makes this readability analysis pertinent. This concern is particularly acute regarding Medicaid respondents, who are increasingly being asked to respond to consumer surveys such as CAHPS®. According to the 1993 National Adult literacy survey (Kirsch, Jungeblur, Jenkins, & Kolstad, 1993), 75% of welfare recipients read at or below the eighth grade level and 50% read at or below the fifth grade level.

A mismatch between an intended respondent's reading ability and the survey instrument may have important implications for the validity of patient evaluations research based on self-administered surveys. Some of the consequences of a mismatch may include low response rates,

especially in vulnerable populations, and unreliable responses because of poor item comprehension.

The readability analysis presented in this chapter is based on readability formulas. Readability formulas are mathematical formulas that predict the readability level of a text. Most readability formulas predict the readability level of a text from two measures: a measure of sentence length (syntactic variable) and a measure of word difficulty (semantic variable). These two variables are each multiplied by distinct constants and linearly combined with an intercept term to form a prediction rule for readability. Mathematically, a readability formula takes the form of:

$$R = \alpha + \beta_1 S + \beta_2 W,$$

where R is the readability level, α is the intercept term, β_1 is the syntactic parameter, and β_2 is the semantic parameter. The syntactic and semantic parameters in readability formulas are estimated based on studies of criterion passages of varying but known levels of reading difficulty.

For this study, five readability formulas were applied to the CAHPS® 2.0 surveys. The *Fry Readability Graph* is adapted for Spanish and English language documents. The *FRASE Graph* is applicable to Spanish language documents only. The *Fog Index*, *SMOG Grading Formula*, and the *Flesch Reading Ease Score* are only applicable to English language text.

Based on applying these formulas, the English versions of the CAHPS® 2.0 surveys had an estimated 6^{th} to 8^{th} grade readability level,

while the Spanish versions of the CAHPS® 2.0 surveys had an estimated 7th grade reading level. The readability estimates from this study are consistent with the opinion of an independent reading expert who assessed the readability level of the CAHPS® 1.0 surveys (Harris-Kojetin, Fowler, Brown, Schnaier, & Sweeny, 1999). They are also consistent with the observations of researchers who conducted cognitive interviews on the CAHPS® 1.0 survey instruments (Brown, Nederend, Hays, Short, & Farley, 1999).

Chapter 7: Evaluating the Equivalence of Health Care Ratings by Whites and Hispanics.

In Chapter 7, an assessment of the equivalence of ratings of care by Hispanic and non-Hispanic white survey respondents to the United Medical Group Association (UMGA) study survey is presented.

The purpose of this study is to assess nine rating questions asked in the survey for bias using statistical methods--more specifically item response theory (IRT) procedures. Bias, in the context of this study, refers to the observation that a question (item) displays different statistical properties in each of the two groups in the study - whites and Hispanics - after controlling for group differences in ratings. Put another way, this study seeks to identify items to which equally satisfied individuals from the different groups have unequal probabilities of answering in the same way. In psychometric parlance, this is a study of differential item functioning or DIF.

The survey data analyzed in this study are from the UMGA study, which is the same survey data analyzed in Chapter 4 and discussed in

section 2 of this chapter. Briefly, the survey data was collected from adult patients receiving care from the United Medical Group Association (UMGA), an association of physician groups located in the western United States. Of the 6,211 surveys analyzed in this study, Hispanics completed 713 and whites completed 5,508. The overall survey response rate was 59%.

Five items included in this study assesses interpersonal aspects of care and four items assessed technical aspects of care. All nine items were administered using an identical 7-point response format (*Very Poor, Poor, Fair, Good, Very Good, Excellent,* and *The Best*) along with the option, *Does Not Apply to Me.* The five interpersonal care items asked survey respondents to rate: medical staff listening to what you have to say; answers to your questions; explanations about prescribed medications; explanations about medical tests and procedures; and reassurance and support from your doctor; and support staff. The four technical care items asked survey respondents to rate: quality of examinations; quality of treatments; thoroughness and accuracy of diagnosis; and comprehensiveness of exams.

Unidimendsionality is assumed by the IRT models used in this study. Cronbach's alpha of 0.96 was obtained for both whites and Hispanics. Principal components analysis was also conducted to test unidimensionality of the nine rating items. All principal components loadings were <0.83 for both the white and Hispanic groups. The ratio of the first to second eigenvalues was 17.8 and 17.3 for the white and Hispanic groups, respectively. The Tucker and Lewis coefficients for a one factor solution were 0.96 and 0.94 for the white and Hispanics groups, respectively. These results strongly suggest that there is a

single factor underlying the nine items and that the hypothesized nine-item scale is unidimensional. Previous studies have also demonstrated that evaluations of interpersonal and technical aspects of care may be represented by a single factor (Marshall, Hays, & Mazel, 1996).

The IRT analysis demonstrated that statistically significant DIF at the 0.05 level occurred in two of the nine items. The first DIF item was an interpersonal rating item and asked about reassurance and support; the second DIF item was a technical rating item and asked about quality of examinations.

To assess the clinical significance of these findings, additional analyses were conducted. First, the effect size of ethnicity - the standardized group mean difference - was assessed with and without inclusion of the DIF items in the ratings scale. Second, the test characteristic curves, based on all nine items, were computed for whites and Hispanics (test characteristic curves show the relationship between level of satisfaction and expected total test scores).

When all nine items were included in the scale, the effect size was 0.27, with whites rating care more positively than Hispanics ($p<0.05$). When the DIF items were dropped from the scale, the effect size was 0.26, with whites rating care more positively than Hispanics. Further, the test characteristic curves for whites and Hispanics are nearly identical by inspection (see Figure 1, page 277).

The results of this study show that although statistically significant DIF was detected in two of nine items used to measure patient evaluations of care, the amount of bias introduced by these items did not have a meaningful effect on a comparison of ratings

between whites and Hispanics. Thus the lower ratings of care by Hispanics should not be attributed to item bias, but rather to differences in actual experiences with care.

Chapter 8: Confirmatory Factor Analysis of the Consumer Assessment of Health Plans Study (CAHPS®) 1.0 Core Survey.

In Chapter 8, a confirmatory factor analysis of the CAHPS® 1.0 core survey is presented. The objectives of this study were:

- to confirm the hypothesized factor structure of the adult CAHPS® 1.0 core survey;
- to confirm the invariance of the adult CAHPS® 1.0 survey factor structure across white and Hispanic respondents, and commercial and Medicaid health sector respondents; and
- to examine the concordance between direct and indirect methods of summarizing consumer experiences with care.

The survey data analyzed in this study is the NCBD 1.0 data, which is the same survey data analyzed in Chapters 5 and 6 of this thesis and discussed in section 2 of this chapter. Briefly, the NCBD 1.0 is an aggregation of CAHPS® 1.0 survey results collected by health plans located across the United States. Health plans administering the CAHPS® surveys were asked to voluntarily contribute their survey results to the NCBD project for purposes of benchmarking and research. Although both adult and child survey results from Medicaid and commercial settings are included in NCBD 1.0, only the adult data were used in this study.

As in the previous study, only the core questions - 23 questions common to all versions of the surveys - were included in this study. These core evaluations can be divided into reports (indirect assessments) and ratings (direct assessments). Reports ask survey respondents to make judgements about the frequency of certain events. The CAHPS® 1.0 surveys include five reporting domains, each addressed by several questions.

The five reporting domains are access to needed care (four questions), provider communication (four questions), office staff helpfulness (three questions), promptness of needed care (four questions), and health plan customer service (two questions). With the exception of two questions which are asked using a *Yes/No* format, all questions in the reports domains are asked using a *Never, Sometimes, Usually, Always* format.

The CAHPS® 1.0 surveys also include four global rating questions. Each global rating item asks respondents to evaluate an aspect of care. The four rating items ask consumers to rate their personal medical provider, specialists, health care, and health plan. All four rating questions are administered using an 11-point numeric scale ranging from 0 to 10, anchored by statements like "The Best Possible Personal Doctor" and "The Worst Possible Personal Doctor."

Multiple group confirmatory factor analysis (CFA) of latent variables was used to address the objectives of this study. In confirmatory factor analysis, also know as structural equation modeling (SEM), hypotheses are translated into a series of mathematical equations that can be solved simultaneously to generate an estimated covariance matrix. By means of various goodness of fit indexes (e.g.,

CFI, NFI, NNFI), the estimated matrix can be evaluated against the observed covariance matrix to determine whether the hypothesized model is a good representation of the data. In general fit index values above 0.90 indicate acceptable model fit.

In the first set of analyses, the extent to which a common factor structure accurately characterized whites and Hispanics sub-samples in both Medicaid and commercial health sectors was tested. To address this objective, a four-group model, with the same patterns of item-factor relationships within each group but without any cross-group constraints, was estimated. These analyses confirmed that the five-factor data structure hypothesized in CAHPS® is well represented among all four subgroup; all goodness of fit indexes evaluated had values of greater 0.90. However, because of small differences in the valence of factor loadings between the health sectors, the subsequent analyses were conducted separately for the Medicaid and commercial subgroups.

In the second set of analyses, the extent of factorial invariance between whites and Hispanics was assessed. To test for weak factorial invariance, cross-group constraints were placed on analogous factor loadings between the white and Hispanic subgroups within each sector. Thus, not only were the patterns of item-factor relationships the same across ethnic groups as in the previous analyses, but the magnitudes and valances of those relationships were constrained to equality across groups. The results of these analyses showed that the model chi-square values increased to a statistically significant level ($p<0.05$) by adding the cross-group constraints in both the Medicaid and commercial sectors. However, by releasing 5 of 23 constraints in the commercial group, and 3 of 23 constraints in the Medicaid group, the change in

model chi-square did not reach statistically significance. These results indicate substantial but not complete weak factorial invariance between whites and Hispanics.

The third main set of analyses addressed the comparability of using direct and indirect evaluations of global evaluations care. To address this question, a secondary factor, accounting for the covariance between the five reports factor, and a common primary factor, accounting for the covariance between the four rating questions, were hypothesized. To test the comparability of the direct and indirect measures, models with the correlation between the secondary factor (reports) and the primary factor (ratings) freely estimated and constrained to 1.0, were compared. In the unconstrained models, the correlation between the two factors was quite high in all groups (Hispanic/commercial: r=0.95; white/commercial: r=0.98; Hispanic/Medicaid: r=0.98; white/Medicaid: r=0.97). However, when the model chi-squared for the unconstrained and constrained were compared there were significant increases in both the Medicaid and commercial samples, indicating significant degradations in model fit. These results indicate that even though the indirect and direct approaches of measuring global satisfaction are highly correlated, they are not completely comparable, in a strict statistical sense.

Implications for Health Policy

This thesis provides evidence that significant racial/ethnic inequalities in patient's evaluations of care exist. Specifically, the studies presented in Chapters 4, 5 and 6 show that Hispanics, Asians/Pacific Islanders, American Indians/Native Alaskans, persons of multiracial/other backgrounds, and persons who speak a language other

than English at home, report worse experiences with care and rate the care they receive lower than whites. Chapters 5 and 6 also show that although African American adults report better experiences with care and rate the care they receive higher that whites, African Americans report worse experiences with care and rate the care their child receive lower than whites. These findings suggest that efforts to improve the delivery of health services to racial and minorities, particularly those who encounter language barriers, are needed.

In a recent policy paper in the Journal of the American Medical Association, Fiscella et al. outlined five principles for reducing racial/ethnic and socioeconomic inequalities of healthcare quality (Fiscella, Franks, Gold, & Clancy, 2000). First, disparities in healthcare must be recognized as a significant quality problem. As the results of the research in this thesis shows and prior research has demonstrated, the healthcare system is not serving all members of society equally. Variation in quality of care generated by factors other than medical need or risk pose a critical challenge to quality in addition to raising questions of distributive justice.

The second principle proposed by Fiscella et al. (2000) is that health plans and other providers of health care need to collect relevant and reliable data to address racial/ethnic disparities. Health plans and other providers should collect racial/ethnic demographic data as part of the core data on patients. Without these data, efforts to illuminate racial/ethnic differences in quality of care are greatly hampered. Racial/ethnic group specific response rates cannot be computed. As a result, it is impossible to determine the representativeness of each sub-group in a survey sample.

Health plans and other providers need to ask about racial/ethnic demographic information on consumer surveys in a consistent manner. Although many health surveys now routinely collect this information, some do not. Further, standardized approaches for collecting this information are needed. Many surveys combine racial and ethnic backgrounds in a single question, forcing respondents to choose from a highly limited set of descriptors. More detailed information can be collected using two part questions. The 1999 current population survey questions about racial/ethnic background are a good example of this approach (www.bls.census.gov/cps/bqestair.htm).

Items about language preference and interpreters should also be included in health-related consumer surveys. Prior research (Woloshin, Bickell, Schwartz, Gany, & Welch, 1995a) as well as evidence presented in this thesis point to the importance of a patient's primary language in determining his/her experiences with healthcare. Communication is a central feature of health care. Without good communication, optimal quality of care cannot be achieved (Woloshin et al., 1995a).

An optimal set of language items might include questions about language use and ability (Schneider, Riehl, Courte-Wienecke, Eddy, & Sennett, 1999), such as those in language acculturation scales (Marin, Vanoss Marin, Perez-Stable, & Vanoss Marin, 1990), and in the 1990 US Census (www.census.gov/population/www/socdemo/lang_use.htm).

In addition to defining a set of items for studying disparities, the items must produce reliable and valid data. Collecting reliable and valid data in multicultural settings is one of the central concerns of this thesis. Chapter 2 presents a framework for producing culturally and linguistically appropriate survey instruments. The

process begins by using focus groups and cognitive interviews to develop equivalent survey domains and items. Readability assessments are recommended to ensure that the survey instruments do not exceed the literacy level of the intended target audience. Finally, the instruments are tested for equivalence using psychometric methods such as item response theory or confirmatory factor analysis. In Chapters 7 and 8, methodological studies to evaluate the psychometric equivalence of two existing instruments are presented.

The third principle proposed by Fiscella et al. (2000) is that performance measures should be stratified by socioeconomic position and racial/ethnic background. Three of the studies presented in this thesis are studies of performance measurement stratified by racial/ethnic background. These studies have been prepared for publication in peer-reviewed journals. One is published and the other two are submitted. However, publication of these results in research journals is not sufficient. Despite concerns about the complexity of performance data (Hibbard, Slovic, & Jewett, 1997), stratified public reporting of health plan performance is needed. Until there is more complete public accountability of health plans for the care they deliver to their patients, it is unlikely that the particular needs of their diverse and vulnerable patients will be addressed adequately.

The fourth principle proposed by Fiscella et al. (2000) is that population wide performance measures be adjusted for socioeconomic position and racial/ethnic background. Fiscella et al. (2000) argue that because the racial/ethnic and socioeconomic backgrounds of enrollees are correlated with current performance measures, including measures of patient's reports and ratings of care, adjustment would

facilitate more meaningful comparisons among health care providers. They further state that this step should not be undertaken until appropriate measures for monitoring care to vulnerable populations have been fully implemented to avoid institutionalizing substandard care.

The question of case-mix adjustment in general and whether to adjust comparisons among health plans for race/ethnicity specifically, has been a topic of research and discussion among the members of the CAHPS® research group at RAND (Elliott, Swartz, Adams, Spritzer, & Hays, submitted for publication). In principle, case-mix adjustment should only be used when a characteristic of survey respondents is correlated with a measure of quality of care, independent of the actual quality of care received.

Consider two alternative scenarios of race/ethnic differences in a measure of quality. In the first scenario, the racial/ethnic composition of two health plans differ because of a priori differential "assignment" to plans (or choice of plans) on the basis of race/ethnicity. This may occur because of geographic convenience, conscious choice, or other factors. Because minorities tend to rate lower than whites, the plan with the higher proportion of minorities will be rated lower. In the second scenario, pre-existing racial/ethnic differences do not exist, but plan performance in fact causes racial/ethnic difference to develop.

The first scenario constitutes the ideal situation for case-mix adjustment. Plan differences are due to the tendency of certain groups to give lower rating than others, independent of the care rendered by health plans. In the second scenario, case-mix adjustment for race/ethnicity is not advisable because it eliminates true information

on quality of care. In this case, adjustment for racial/ethnic differences prior to enrollment in the current plan is desirable, but adjustment for racial/ethnic differences since enrollment, or by extension racial/ethnic differences after enrollment in the current plan, is not desirable. In reality, probably a mixture of both scenarios exists. Thus, to the extent that the second scenario is true, case-mix adjustment for race/ethnicity is not advisable because race/ethnicity reflects true differences in quality rather than respondent bias. The methodological studies in Chapters 7 and 8 of this help determine the extent to which racial/ethnic differences in ratings are due to response bias.

The fifth and final principle Fiscella et al. (2000) suggest is that provider reimbursements include adjustments for the race/ethnicity and socioeconomic position of their patients. Presumably, these adjustments would account for the additional cost of providing care to minority patients and help offset the costs of quality improvement efforts designed to eliminate disparities.

Implications for Future Research

Future research efforts should focus on determining the appropriateness of case-mix adjustment for race-ethnicity and language use and preference. Although the two studies in this thesis address the issue of equivalence between whites and Hispanics, similar studies are needed to address equivalence among other racial/ethnic groups, including Asians, Pacific Islanders, African Americans, and American Indians. Further, cultural and linguistic differences between Asian and Hispanic subgroups may necessitate equivalence studies among these groups. For example, equivalence studies are needed between responded

answering to Spanish language versions of the CAHPS® survey and those answering to the English language version. Similar studies will be needed as the CAHPS® surveys are translated into other languages.

Although racial/ethnic differences are detectable using the current racial/ethnic categories in CAHPS® and UMGA, more refined racial/ethnic subgroup analysis is desirable. There is increasing evidence of substantial variation in access to care and satisfaction with care among Asians by country of origin (Murray-Garcia, Selby, Schmittdiel, Grumbach, & Quesenberry, 2000) and language spoken at home (see Chapter 6), and among Hispanics by language spoken at home (see Chapter 6). By aggregating up to major racial/ethic group levels (e.g., Hispanics) important subgroup differences might be overlooked. In order to look at racial/ethnic subgroup, additional patient demographic variables will be needed. These might include for immigrants: country of origin, level of acculturation, and time spent in the United States; or for the US born: generational status and level of acculturation.

Finally, existing healthcare consumer surveys are far from perfect and require ongoing refinements. Both item revisions and changes in the domains may be needed as the health care system evolves. Future translations may necessitate item revisions or item deletion to maintain instrument equivalence. To address these and other pressing issues regarding the assessment of patient's experiences with care through surveys, ongoing qualitative and quantitative research is needed.

References

Baker, D.W., Parker, R.M., Williams, M.V., Coates, W.C., & Pitkin, K. (1996). Use and effectiveness of interpreters in an emergency department. JAMA, 275(10), 783-8.

Brown, J.A., Nederend, S.E., Hays, R.D., Short, P.F., & Farley, D.O. (1999). Special issues in assessing care of Medicaid recipients. Med Care, 37(3 Suppl), MS79-88.

Elliott, M.N., Swartz, R., Adams, J., Spritzer, K.L., & Hays, R.D. (Submitted for publication). Case-mix adjustment of the National CAHPS Benchmarking Data 1.0: A comparison of case-mix models. Health Services Research,

Fiscella, K., Franks, P., Gold, M.R., & Clancy, C.M. (2000). Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. JAMA, 283(19), 2579-84.

Harris-Kojetin, L.D., Fowler, F.J. Jr, Brown, J.A., Schnaier, J.A., & Sweeny, S.F. (1999). The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. Consumer Assessment of Health Plans Study. Med Care, 37(3 Suppl), MS10-21.

Hibbard, J., Slovic, P., & Jewett, J. (1997). Informing consumer decisions in health care: Implications from decision-making research. Milbank Q, 75(3):395-414.

Kirsch, I., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). Adult Literacy in America. (ed.). Princeton, NJ: Educational Testing Service.

Marin, G., Vanoss Marin, B., Perez-Stable, E.J., & Vanoss Marin, B. (1990). Feasibility of a telephone survey to study a minority community: Hispanics in San Francisco. Am J Public Health, 80(3), 323-6.

Marshall, G.N., Hays, R.D., & Mazel, R. (1996). Health status and satisfaction with health care: results from the medical outcomes study. Journal of Consulting and Clinical Psychology, 64(2), 380-90.

Murray-Garcia, J.L., Selby, J.V., Schmittdiel, J., Grumbach, K., & Quesenberry, C.P. Jr. (2000). Racial and ethnic differences in a patient survey: patients' values, ratings, and reports regarding physician primary care performance in a large health maintenance organization. Med Care, 38(3), 300-10.

Perez-Stable, E.J., Naapoles-Springer, A., & Miramontes, J.M. (1997). The effects of ethnicity and language on medical outcomes of patients with hypertension or diabetes. Medical Care, 35(12), 1212-9.

Schneider, E.C., Riehl, V., Courte-Wienecke, S., Eddy, D.M., & Sennett, C. (1999). Enhancing performance measurement: NCQA's road map for a health information framework. National Committee for Quality Assurance. JAMA, 282(12), 1184-90.

Woloshin, S., Bickell, N.A., Schwartz, L.M., Gany, F., & Welch, H.G. (1995a). Language barriers in medicine in the United States. JAMA, 273(9), 724-8.

Figure 1. Test characteristic curves for whites and Hispanics respondents.