Appendix

Analysis of Differential Item Functioning in the CAHPS Composites by Depression Status

To investigate the possibility of differential scale use by depressed and non-depressed beneficiaries, we subjected the composite measures of health care experiences to an analysis of differential item functioning (DIF). Statistical techniques for evaluating differential item functioning are based on the notion that if different groups of respondents (e.g., those with and without depressive symptoms) have roughly equivalent levels of some underlying construct (e.g., experiences with health care) then they should respond similarly to individual items measuring that construct. DIF methods allow one to judge whether items (and ultimately the measurement scale they constitute) function the same way in different groups of respondents. If DIF is present, then observed group differences cannot be attributed entirely to differences in the underlying construct.

Although there are a variety of ways to assess DIF, one of the most commonly used methods is through the use of logistic regression (Clauser and Mazor 1998; Swaminathan and Rogers 1990; Zumbo 1999). In our study, we used ordinal logistic regression to evaluate the "quality of doctor communication" composite and an "access-to-care" composite to determine whether the items on these measures function similarly across depressed and non-depressed Medicare beneficiaries. The logistic regression method of DIF detection entails conducting a regression analysis for each item wherein one tests the statistical effect of the grouping variable (i.e., depression) and the interaction of the grouping variable and a criterion variable (typically the total scale score) after conditioning on the criterion variable. Because of the need for a criterion variable, we could not perform a DIF analysis of the single-item global ratings of care. Because multi-item criterion variables are more reliable than single-item criterion variables, we

combined the items from the "getting needed care" and "getting care quickly" composites and evaluated the resulting 4-item scale ($\alpha = 0.70$) as a single measure of "access to care."

To assess the unidimensionality of the "quality of doctor communication" and "access to care" scales, we conducted confirmatory factor analyses of each using Mplus, version 6.1. A single-factor confirmatory categorical factor analysis model for the four communication items fit the data well according to practical fit indices (comparative fit index (CFI) = 0.997; root mean square error of approximation (RMSEA) = 0.028), providing support for the unidimensionality of the scale. A single-factor confirmatory categorical factor analysis model for the four access to care items also fit the data well according to practical fit indices (CFI = 0.956; RMSEA = 0.045), Modification indices indicated that the residual correlation between the two items representing getting care quickly was noteworthy ($r = 0.221$) and the fit of the model could be improved even more by adding this estimate.

Having established the unidimensionality of the communication and access composites, we fit three ordinal regression models to each item of each scale to evaluate DIF:

Model 1: $Y = b_0 + b_1TOT$

Model 2: $Y = b_0 + b_1TOT + b_2DEPRESSION$

Model 3: $Y = b_0 + b_1TOT + b_2DEPRESSION + b_3TOT*DEPRESSION$

where Y is the probability of a particular response on the item's ordinal response scale, TOT is the total scale score for each respondent, and DEPRESSION is the grouping variable (dummy coded as 0 = non-depressed, 1 = depressed). This method provides a test of DIF conditional on the relationship between the item response and the total scale score.

Preliminary analyses revealed several violations of the proportional odds assumption underlying the ordinal logit model. To identify the nature of these violations, we compared the

log odds for each cut-point of each of the four access and four communication items as recommended by Scott and colleagues (1997). Doing so showed that depression is associated with the tendency to give lower ratings to a near constant extent across the response categories "never," "sometimes," and "usually," but that depression is less strongly associated with a tendency to endorse "usually" rather than "always." With the "usually" and "always" categories collapsed, the proportional odds assumption held for all items; thus, we conducted our DIF analyses with all items scored in this manner.

A comparison of the R-squared value for Model 1 with the R-squared value for Model 3 provides a simultaneous test of uniform and non-uniform DIF. If the incremental R-squared value is not at least 0.035, there is no evidence of DIF and further model comparisons to separately evaluate uniform and non-uniform DIF are not warranted (Zumbo 1999). In analyses such as ours that are based on large sample sizes, a comparison of model R-squared values is preferable to a comparison of model chi-squared statistics as the latter can point to statistically significant findings where the effect is trivial (Kirk 1996; Zumbo 1999). In this analysis, we used the Zumbo-Thomas measure of effect size for R-squared (Zumbo 1999), which parallels effect size measures for other statistics (Cohen 1992). The results of our analyses are shown in the Table A1 below.

Table A1

Results of Ordinal Logistic Regression Models to Evaluate DIF in the Provider Communication and Access to Care Composites by Depression Status

	Model 1 (with total scale score)	Model 2 (with total score and depression)	Model 3 (with total score, depression, and their interaction)	DIF Test (comparison of Models 1 and 3)
	R^2	R^2	R^2	Difference in R^2
Provider Communication				
Gives easy-to-understand explanations	0.5704	0.5701	0.5675	0.0029
Listens carefully	0.7037	0.7033	0.7025	0.0012
Shows respect	0.6670	0.6669	0.6661	0.0009
Spends enough time	0.5713	0.5709	0.5706	0.0007
Access to Care				
Getting care when sick or injured	0.2888	0.2888	0.2858	0.0030
Getting routine care	0.2441	0.2439	0.2380	0.0061
Getting appointments with specialists	0.1853	0.1842	0.1710	0.0143
Getting care, tests, and treatment	0.1890	0.1889	0.1752	0.0138

As can be seen in the last column of Table A1, we found no evidence of DIF for either composite. The largest increment in R-squared that we observed was 0.014, which corresponds to a trivial effect size by Cohen's standards (Cohen 1992). Thus, we conclude that differences observed between depressed and non-depressed beneficiaries on the measures of provider communication and access to care reflect true differences in patient experience rather than differential scale use by the two groups of beneficiaries.