# Are Provider Communication Constructs and Structural Relationships the Same Across English and Spanish?

Gerald K. Arnold[1] and Rebecca A. Baranowski[1]
[1]American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia PA 19106

## Abstract

We compare patient ratings of primary-care physicians on three composite measures related to provider communication skills (6 items), shared decision making (2 items), and overall satisfaction (11-point rating scale). Measures are adapted from practice-based Clinician and Group (CG) CAHPS® Surveys in English or Spanish; surveys are completed anonymously (minimum of 25 responses) using an automated phone or Web system. These surveys satisfy self-evaluation of practice performance requirements for physicians engaged in maintaining board certification in internal medicine. Responses may differ among patients because measurement constructs may not translate well between languages. If latent constructs are interpreted by Spanish- and English-speaking patients similarly, then covariance structures for the same physicians would be indistinguishable after controlling for differences unrelated to language (e.g., health status, education). We compared patient ratings for 79 physicians who administered surveys to both Spanish- (301 cases) and English- (1,708 controls) speaking patients. A second control group included 2,048 ratings (English only) of another 79 physicians who were propensity-matched to each study physician based on 31 practice, demographic, training, and achievement characteristics and 9 patient demographic variables. Groups were divided into 49 training pairs and 30 testing pairs for validation. A three-group, MIMIC, structural equation model was adapted to the Spanish cases and the two English control groups. Covariance and mean structures were not equivalent ($p<0.001$), and a series of hypothesis tests suggests that language factor structures were the same for Spanish and English patients ($p=.061$), factor loadings for items differed ($p=.003$), and differences in propensity score regressions between groups were not due to selection alone ($p<.001$). This study shows that patients' ratings on three provider constructs had similar structural relationships among latent constructs but showed differential item functioning by language. There was a significant propensity score interaction with group membership. The study suggests that structural equation models may be useful in assessing whether different language groups vary with respect to what constitutes good communication skills in their health care providers, the quality of the items used to measure constructs, and whether propensity score matching works uniformly across groups.

**Key Words: Communication constructs, structural equation models, propensity scores**

## 1. Introduction

Today medical certification is more than passing an examination it has become a continual process of learning and re-evaluation of medical competencies—a process known as maintenance of certification (MOC). Implementation of the MOC program by the American Board of Internal Medicine in 1990 meant that physicians must periodically demonstrate that they are keeping up their medical knowledge. Since 2006, they must also show engagement in self-evaluation of the quality of care they provide. They do that through a web-based tool called the Practice Improvement Module (PIM). One specific area of assessment is the domain of physician-patient communication, which is assessed in the Communication PIM. Three versions are available: primary care assessment over 12 months, subspecialty care assessment over 12 months, and assessment at last office visit.

The Communication PIM uses the CAHPS® Clinician and Group (CG) surveys to measure the processes of care. Physicians use a PDSA (Plan, Do, Study, Act), Shewhart-Deming, cycle to evaluate their performance. Physicians administer surveys to patients anonymously (minimum of 25 responses) using an automated phone or Web system. Physicians review survey results, plan an improvement on one communication process, implement the plan, and then re-measure that process using 25 more patient surveys. CAHPS® Clinician and Group (CG) surveys are available in English and Spanish.

We compare patient ratings of primary care physicians on three CAHPS® composite measures related to provider communication skills (six items), shared decision making (two items), and overall satisfaction (11-point rating scale). Previous experience indicates that measures from the CG CAHPS® surveys sometimes differ between English-speaking and Spanish-speaking patients. Responses may differ because instruments or measurement constructs may not translate well across languages. We wanted to evaluate this possibility. If Spanish- and English-speaking patients interpret communication constructs similarly, then covariance structures of measures for these constructs on the same physicians would be indistinguishable after controlling for differences unrelated to language (e.g., health status, education).

Two general strategies for assessing whether latent constructs measured by questionnaires are interpreted similarly across groups include procedures related to comparing latent structures among groups or procedures related to assessing psychometric properties of items used to measure the latent variables. One way of comparing or controlling language structures in questionnaires is through the use of language experts following a translation protocol. The Spanish translation of the CAHPS® instruments followed such a protocol (Weidmer, Hurtado, Weech-Maldonado, Ngo-Metzger and Bogen, 2006 & Weidmer, Brown, & Garcia, 1999). Another approach is to have patients with different languages or cultures review sets of standardized, provider-patient encounters such as video presentations and ask reviewers to rate these encounters on a common form (Weinick, Elliot, Volandes, Lopez, Burkhart, & Schlesinger, 2011). More analytical approaches used to compare latent structures among groups include linear and nonlinear random-effects models (Meulder & Xie, 2004) and confirmatory factor analysis models such as the Multiple Indicator Multiple Cause (MIMIC) model (Bollen, 1989 & Kaplan, 2009).

The psychometric approach for comparing latent structures among groups typically comes in the form of comparing item response characteristics such as slopes and

intercepts estimated from Item Response Theory (Hambleton, Van der Linden, & Wells, 2010). Group differences in item characteristics are called Differential Item Functioning (DIF) or in a collection of items with the same content (facets), Differential Facet Functioning (DFF). DIF can be "uniform," group differences in item intercepts only, or "non-uniform," group differences in item slopes or slopes and intercepts (Meulder & Xie, 2004). Many DIF studies utilize nonlinear models for group comparisons such as logistic regression (Scott et al. 2010). Recently latent structure models and psychometric models have been combined to study group differences in latent structures. MIMIC models are used to study uniform and non-uniform DIF (Wang & Shih, 2010; Woods, 2009; Woods & Grimm, 2011, and Ying, et al. 2012). This progression seems reasonable in that under certain conditions, item factor loadings and means used in latent variable models relate to the slopes and intercepts estimated by item response procedures.

Comparing latent variables like provider communication skills or factor structures such as factorial invariance among groups assumes that the groups are similar with respect to other factors that could confound these contrasts. If Spanish-speaking patients and English-speaking patients differ in their ratings of physician communication skills is it because these patient groups use different cultural norms for these constructs or is it due to some hidden selection bias? Most language comparison studies of questionnaires are observational in nature, so hidden biases are always a threat to any causal inferences.

One approach to control hidden biases in comparison studies of latent structures is regression in structural models with means, an analysis of covariance model, ANCOVA (Sorbom, 1978). This model assumes homogeneity of regression models among groups. An alternative is group matching through the use of propensity scores (Rosenbaum and Rubin, 1983). This model assumes that treatment assignment is strongly ignorable. Kaplan (1999 & 2009) describes a propensity score approach for comparing latent structures that involves stratifying groups in a MIMIC structural model on the basis of the propensity scores. One problem with stratification or regression by propensity scores is that estimates of group differences can be somewhat biased (Austin, 2010). In this study we develop a pseudo-ANCOVA, MIMIC model regressing latent variables on propensity scores to compare patient ratings of physician communication skills from Spanish and English versions of the CG CAHPS® surveys. Use of the ANCOVA MIMIC model in a series of four hypothesis tests enables us to demonstrate the validity of the communication constructs across languages and assess the quality of the instruments used to evaluate the constructs.

## 1.1 Hypotheses Related to Provider-Patient Communication Constructs

The first hypothesis is that covariance and mean structures of the comparison groups are the same, the Equal Structures Hypothesis. If this null hypothesis is not rejected then no further tests are necessary since equal covariance and mean structures imply equal latent structures (Joreskog, 1993).

If the first hypothesis is rejected then a series of three hypotheses follow. Tests two and three follow the testing protocol for assessing factor structure invariance among groups described by Bollen (1989). Test two is an assessment of the Equal Forms Hypothesis among groups. A common MIMIC factor structure is fit across all groups. Next the factor structure for the group in question is altered. The model remains a nested structure within the equal forms model. A goodness-of-fit chi-squared test comparing the equal forms to the altered group structure model is a test that the group in question has the same latent structure as other groups. Note this is the test of forms and not of equal model

parameters. Further testing following the Bollen protocol would be required to assess equivalence of model parameters. The common form model for the CAHPS® measures used in this study is shown in Figure 1 below. Each group compared in the MIMIC model would have the same structure.

Test three is a test of equal model parameters for the measurement (confirmatory factor analysis) portion of the MIMIC model. This test is called the Equal IRT Hypothesis. The equal form model is restricted so that the factor loadings for the measures and the measure mean vectors are equal. Recall that latent variable factor loading and item mean vectors are related to the slope and intercept estimates that come from item response theory applicable to the normal ogive model. The Equal IRT model may be compared with the equal form model but is typically used to compare nested models within groups to test whether factor loadings and mean vectors differ across groups (i.e., differential item functioning). Note if the equal forms hypothesis is rejected, then the equal IRT model should not assessed because the factor structures for groups are not equivalent.

The fourth test is of the Equal Propensity Score Regression Hypothesis. In this model the equivalence of the regression parameters for the propensity scores on the latent variables across groups is assessed. This test came from a propensity matching protocol described by Kaplan (1999 & 2009). Here the structural coefficients and the propensity score means (intercepts) are set to be equal across groups. Again this model may be compared with the equal forms model, but typically it is used to assess whether group-specific changes in the regression models provide better fit. If not, then the assumption that group differences due to selection only is appropriate. That is, the propensity matching worked. If the changes in regression parameters are significantly better than the equal regression model, then the group differences are not entirely due to selection issues. However, if the equal forms model is a significantly better fit than the restricted regressions, then group comparisons are not justified.

## 2. Data and Methods

### 2.1 Populations, Samples, and Study Design

The study population included 1,407 US physicians, certified in Internal Medicine, who completed a primary care Communication PIM between spring 2009 and fall 2012. The baseline CG CAHPS® surveys from these practice performance assessments included ratings from 37,272 patients. Among these physicians, 79 case practices administered both the Spanish and English versions of the survey. The remaining 1,486 physicians administered CG CAHPS® to English-speaking patients only and made up the pool of control practices for the study. From the control pool, 79 practices (i.e., physicians with their patients) were matched one-to-one to the case practices, making 79 case-control pairs. The 158 physician practices with the baseline ratings from 4,057 patients made up the study sample. Because the study involved modelling, the sample was divided into two sub-samples of 49 learning practice pairs for model building and 30 testing pairs for cross-validation. The entire study consisted of six patient groups. Within the learning sample, there were 204 Spanish-speaking patients and 1,071 English-speaking patients from the case practices and 508 English-speaking patients from the control practices. Within the testing sample, there were 97 Spanish-speaking patients and 637 English-speaking patients from the case practices and 1,468 English-speaking patients from the control practices. The 49 case practices in the learning sample were used in a previous

language study so the data from the learning sample preceded that of testing sample chronologically.

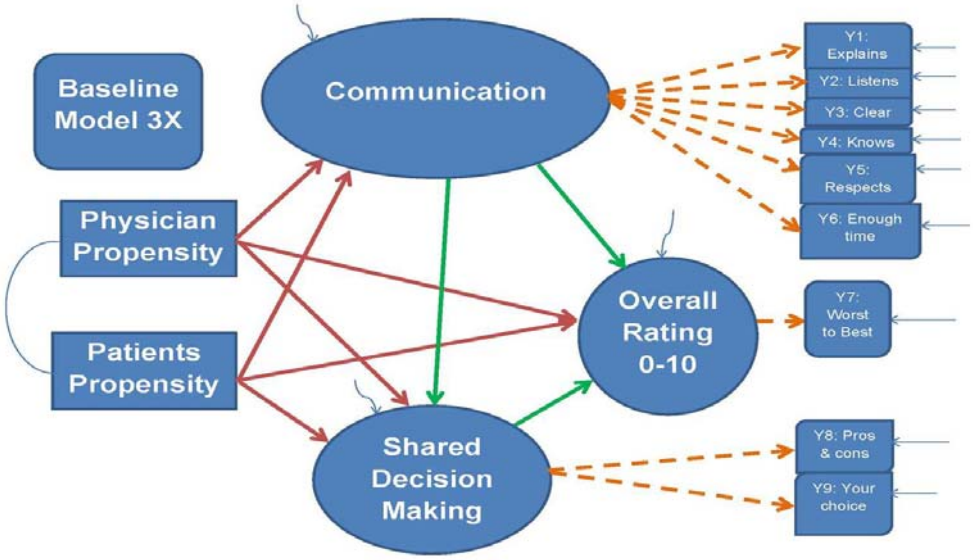


**Figure 1:** The Equal Forms MIMIC model for three composite measures from the CG CAHPS® patient surveys. Each study group would have the model shown. This study includes three groups: Spanish-speaking patients from the case practices, English-speaking patients from the case practices, and English-speaking patients from propensity matched control practices. The squares on the left represent the exogenous physician and patient propensity scores; the ovals in the middle represent the three endogenous, latent variables measured in the survey instrument; and the nine smaller squares on the right represent the questionnaire items used to measure each of the latent variables.

## 2.2 Instruments and Measures

The instruments used in the Communication PIM were based on the CAHPS® Clinician and Group (CG) Surveys, Adult Primary Care Questionnaire 1.0 version, six-point response scale in English and Spanish. The questionnaires are formatted to the web system that administers the Communication PIM. Copies of the English and Spanish versions of the questionnaires may be found at the website:
https://www.cahps.ahrq.gov/clinician_group/

The provider communication composites included six items related to provider communication skills, two items on shared decision-making, and an overall quality rating of the physician based on an 11-point scale ranging from 0 to 10. The six communication items are rated on a common six-point scale regarding frequency of a physician's behavior ranging from (1) Never to (6) Always. The decision-making items are also about a physician's behavior but are binary scored as Yes or No. For analyses, the six-point items were scored from (0) Never to 5 (Always), and the decision-making items were scored 1 for Yes and 0 for No. Answering the decision-making items was conditional on the patients indicating that their physician told them that there was more than one choice for their treatment or health care.

## 2.3 Statistical Methods

Dual propensity score matching formed the case control pairs of practices. The physician-level propensity scores calculated by logistic regression included 31 test performance, practice characteristic, demographic, and PIM version information variables. The second score included nine patient demographic and PIM version variables and all two-way interactions among these variables. Patients are nested within physicians, so the physician level score was a weighted average of patient responses adjusted for within physician clustering. Practice pairs were greedy matched using shortest Mahalanobis distances between case and control practices.

Because measures were categorical, Muthen (1993) recommends that correlations among measures be based on polytomous correlations, which assume an underlying normal distribution to the response categories. This procedure was followed; however, the correlation matrices for the Spanish-speaking patients in both the testing and learning samples were not positive definite so the modelling program would fail. To get the modelling program to run, the correlation matrices were re-calculated using ranks. These matrices worked and were used as the analysis data. All models reported are based on the learning samples. The testing samples were used for cross validation testing only. The correlation matrices and summary statistics for each of the three study groups are given in Table 1 for the learning sample and Table 2 for the testing sample.

The SAS® procedure CALIS in LISMOD mode was used to estimate the ANCOVA MIMIC models and to calculate the test statistics for the four language constructs hypothesis tests. Fit statistics for models include the likelihood ratio chi-squared statistic, the Root Mean Square Error of Approximation (RMSEA), Akaike Information Criterion (AIC) and the Comparative Fit Index (CFI). Guidelines for use suggest that RMSEA $\leq$ .05 and CFI $\geq$ .95 suggest good model fit (Kaplan, 2009). All analyses were completed using SAS® version 9.3.

# 3. Results

## 3.1 Propensity Score Matching

Table 1 compares the variables used in the physician-level logistic regression to predict practice cases that used both Spanish and English versions of the CAHPS®. Only one of the physician-level variables (PIM version) differed significantly between cases and controls. The case data came primarily from older versions of the Communication PIM.

Table 2 compares the variables used in the patient within physician logistic regression to predict practice cases that administered both Spanish and English versions of the CAHPS®. Though patient distributions across the demographic variables are similar, all variables except gender were significantly different between case and control practices. Cases patients were younger, less educated, and had shorter tenures with their physicians than control patients. The propensity matching did not work well at controlling patient differences in the samples.

The Equal Structures Hypothesis was rejected: $\chi^2_{(df)} = 714.3$ (108), $p<.001$; RMSEA = .12; & AIC = 822.3. Spanish and English cases have different covariance and mean structures, and they both have different structures from the controls.

Figure 2. shows that the Equal Forms (least restrictive) model fits the three groups best. A modification restricting the Spanish communication constructs to two factors while leaving the English constructs at three factors was not a significant improvement in model fit: $\chi^2_{(df)}$ = 5.6 (2), p=.061 indicating that factor forms across the three study groups are similar. The Equal IRT test indicated that the Spanish factor loadings and item means were significantly different from the English samples. This suggests non-uniform differential item functioning. Closer inspection of the data showed that Spanish responses about physicians showing respect for patients were more positive than their English counterparts but that they were more critical of their physicians in terms of them providing easy to follow instructions about their care. The Equal Propensity Score Regression Hypothesis was rejected. A modification allowing the Spanish and English cases to have common slopes but different intercepts (homogeneity of slopes) was a significant improvement in model fit $\chi^2_{(df)}$ = 644.3 (10), p<.001 indicating that group differences along cannot be explained by selection biases alone. The modification however was not significantly different from the Equal Form model suggesting that the groups may not be comparable.

Figure 3 shows the fit statistics for the testing sample for same models using the parameters estimated from the learning sample. Model fit is significantly worse indicating that the structural models are sensitive to sample characteristics.

| Fit | =Form | =IRT | =PSReg |
|---|---|---|---|
| $\chi^2_{(df)}$ | 223.5 (112) | 353.5 (135) | 875.0 (127) |
| RMSEA: (90% CI) | .049 (.040, .059) | .063 (.055, .071) | .12 (.112, .127) |
| AIC | 461.5 | 545.5 | 1083.0 |
| CFI | .974 | .949 | .827 |
| **Modifications** | | | |
| $\Lambda_1=\Lambda_2\neq\Lambda_3$ $\chi^2_{(df)}=229.1_{(114)}$ | 2 factors for Spanish vs. 3 for English $\chi^2_{(2)=}$ 5.6, p=.061 | | |
| $\Lambda_1=\Lambda_2\neq\Lambda_3$ & $\nu_1=\nu_2\neq\nu_3$ $\chi^2_{(df)}=337.4_{(131)}$ | DIF for Spanish Com $\chi^2_{(4)=}$ 16.1, p=.003 | | |
| $\Gamma_1=\Gamma_3\neq\Gamma_2$ & $K_1\neq K_3\neq K_2$ $\chi 2(df)=231$ (117) | Differences not due to selection only $\chi^2_{(10)=}$644.3, p<.001 | | |

**Figure 2:** Fit Statistics for Testing Hypotheses Regarding Physician Communication Constructions in Both Spanish and English Using Data from the Learning Sample.

| Fit | 1 = Form | 2 = IRT | 3 = PS Reg. |
|---|---|---|---|
| $\chi^2_{(df)}$ | 223.5 (112) | 353.5 (135) | 875.0 (127) |
| RMSEA: (90% CI) | .049 (.040, .059) | .063 (.055, .071) | .12 (.112, .127) |
| AIC | 461.5 | 545.5 | 1083.0 |
| CFI | .974 | .949 | .827 |
| **Cross-Validation** | | | |
| $\chi^2_{(df)}$ | 3563.0 (231) | 3571.4 (231) | 4787.9 (231) |
| RMSEA: (90% CI) | .177 (.172, .182) | .177 (.172, .182) | .207 (.202, .212) |
| AIC | 3563.0 | 3671.4 | 4787.9 |
| CFI | .418 | .416 | .203 |

**Figure 3:** Sensitivity of Fit Statistics for Testing Hypotheses Regarding Physician Communication Constructions in Both Spanish and English Using Data from the Testing Sample With Model Parameters From the Learning Sample.

## 4. Summary

Are provider communication constructs the same in English and Spanish? Probably but some items show differential item functioning in the physician communication factor. Specifically Spanish patients rate their physicians much higher on their provider showing respect for patients but significantly lower on physicians providing understandable instructions about their care. This study shows that patients' ratings on three provider constructs had similar structural relationships among latent constructs but showed differential item functioning by language.

There was a significant propensity score interaction with group membership. Spanish and English patients within case practices had similar propensity to latent structure regressions but the propensity regression for the matched controls was quite different. This suggests that there are likely real group differences beyond selection biases. However the restricted model showing heterogeneity of the regressions between cases and controls was not a significantly better fit that the equal form model indicating that the propensity matching likely failed especially with the patient matching.

The study suggests that structural equation models may be useful in assessing whether different language groups vary with respect to what constitutes good communication skills in their health care providers, the quality of the items used to measure constructs, and whether propensity score matching works uniformly across groups.

## Acknowledgements

The authors would like to thank Halyna Didura for her assistance with data preparation. We also thank Drs. Rebecca Lipner, Senior Vice President, and Robin Guille, Vice President, of the Evaluation, Research and Development Department at the American Board of Internal Medicine for their critical but helpful comments. Both authors contributed to the study conception and design; to the acquisition, analysis, and interpretation of the data; and drafted these proceedings. There are no funders for this paper but the authors are employees of the American Board of Internal Medicine. There were no prior presentations of this work.

## References

Austin P. C. (2010) The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Statistics in Medicine. 29 2137–2148

Bollen, K. A. (1989) Structural equations with latent variables. Wiley, New York pages 355-369.

Joreskog, K. G. (1993) Testing structural equation models. In Testing Structural Equation Models (Kenneth A. Bollen & J. Scott Long eds.). Sage, Newbury Park, CA Pages 294-316.

Kaplan D. (2009) *Structural equation modeling, 2nd.* Ed. Sage, Thousand Oaks, CA Pages 82-83.

David Kaplan (1999): An Extension of the propensity score adjustment method for the analysis of group differences in MIMIC models, Multivariate Behavioral Research, 34:4, 467-492

Hambleton, R. K., Van der Linden, W. J. & Wells, C. J. (2010) *IRT models for the analysis of polytomously scored data: brief and selected history of model building advances*. IN Handbook of Polytomous Item Response Theory Models (Michael L. Nering & Remo Ostini eds) Routledge, Taylor and Francis Group, NY .) Pages 21-42.

Meulder M. & Xie, Y. (2004) *Person-by-item predictors.* In Explanatory Item Response Models (Paul De Boeck & Mark Wilson eds.) Pages 213-240. Springer-Verlag, NY

Muthen, B. O. (1993) Goodness of fit with categorical and other non-normal variables. In Testing Structural Equation Models (Kenneth A. Bollen & J. Scott Long eds.). Sage, Newbury Park, CA Pages 205-234.

Neil, S., Scott, W., Fayers, W. M., Aaronson, N. K., Bottomley, A.,de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., Sprangers, AG, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health and Quality of Life Outcomes 2010 8:81.

Rosenbaum, P. R & Rubin, D. B. (2006) The Central role of the propensity score in observational studies for causal effects. In Matched Sampling for Causal Effects 2006 Cambridge University Press NY reprinted from Biometrika, 1983, 70 Pages 41-55.

Sorbom, D (1978). An alternative to the methodology of analysis of covariance. Psychometrika, 43, 381-396.

Tatsuoka, M. M. (1988) Multivariate Analysis: Techniques for Educational and Psychological Research, 2nd. Ed. Macmillian NY Pages 210-266.

Weidmer, B. Hurtado, M. Weech-Maldonado, R. Ngo-Metzger, Q.and Bogen, K. (2006). Guidelines Translating CAHPS® Surveys https://www.cahps.ahrq.gov/translating-surveys.htm

Weidmer B, Brown J, Garcia K. "Translating CAHPS 1.0 Survey Instruments into Spanish," Medical Care, March 1999. 37 (3).

Weinick, Elliot, Volandes, Lopez, Burkhart, & Schlesinger, (2011) Using standardized encounters to understand reported racial/ethnic disparities in patient experiences of care. Health Services Research 46 (2) pages 491-509.

Wang, W.C. and Shih, C. L. (2010) MIMIC methods for assessing Differential Item Functioning in polytomous items Applied Psychological Measurement 34(3) 166–180

Woods, C. M. and Grimm, K. J. (2011) Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models Applied Psychological Measurement 35(5) 339–361

Woods, C. M. (2009) Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis, Multivariate Behavioral Research, 44:1, 1-27

Ying J., Nicholas, D., Myers, N. D. Soyeon, A. and Penfield, R. D. (2012) A Comparison of uniform DIF effect size estimators under the MIMIC and Rasch models. Educational and Psychological Measurement 73(2) 339–358

Table 1: Demographic, Training, Certification, Practice, Performance Characteristics of Physicians in Study

| Propensity Matching Variables | Cases n=79 | Controls n=79 | P-values |
|---|---|---|---|
| Personal Demographics | | | |
| Gender: % Female | 35% | 34% | 0.867 |
| Age in Years: Mean (SD) | 46.9 (6) | 47.3 (6) | 0.65 |
| County of Origin: % US/Canada | 44% | 42% | 0.748 |
| Medical Training & Certification | | | |
| County of MS Training: % US/Canada | 52% | 47% | 0.525 |
| Pass 1st Attempt at IM Exam: % Pass | 77% | 81% | 0.557 |
| Cert. Exam score imputed?: % Yes | 4% | 3% | 0.65 |
| Years Since 1st IM Exam: Mean (SD) | 14.6 (5) | 15.0 (5) | 0.615 |
| Initital IM Cert Score: Mean (SD) | 430.4 (130) | 439.2 (125) | 0.663 |
| Practice Characteristics | | | |
| Region of Practice | | | 0.669 |
| Northeast | 42% | 38% | |
| Midwest | 11% | 16% | |
| South | 25% | 20% | |
| West | 22% | 25% | |
| Practice Setting % category | | | 0.884 |
| Solo/Group practice | 56% | 62% | |

| | | | |
|---|---|---|---|
| Group HMO | 11% | 15% | |
| Academic Practice | 6% | 4% | |
| Hospital Office | 5% | 5% | |
| Hospital Inpatient | 4% | 4% | |
| Military/Government | 1% | 1% | |
| Community Health Center | 11% | 6% | |
| Residency Clinic or Other Settings | 5% | 3% | |
| Specialty Practice: % Multispecialty | 38% | 42% | 0.847 |
| # Physicians in Practice | | | 0.975 |
| Solo | 28% | 28% | |
| 2-5 | 31% | 34% | |
| 6-10 | 9% | 14% | |
| 11-15 | 5% | 6% | |
| 16-25 | 3% | 4% | |
| 26-50 | 9% | 6% | |
| 51 or more | 14% | 13% | |
| Not applicable | 1% | 1% | |
| Major Conditions Treated in practice | | | 0.725 |
| Arthitis | 3% | 4% | |
| Cardiovascular Disease | 10% | 8% | |
| Depression | 3% | 1% | |
| Diabetes | 51% | 46% | |
| Hypertension | 30% | 35% | |
| Other | 1% | 0% | |
| None | 3% | 6% | |
| Weekly Hrs. in Patient Care: % Category | | | 0.864 |
| 1-20 hrs. | 10% | 8% | |
| 21-40 hrs. | 5% | 5% | |
| 41-60 hrs. | 19% | 16% | |
| 61-80 hrs. | 8% | 7% | |
| > 80 hrs. | 57% | 66% | |
| Skipped | 1% | 0% | |
| Patient Zip Codes - 1st digit: % category | | | 0.999 |
| 0-1 | 43% | 38% | |
| 2-3 | 15% | 11% | |
| 4-5 | 4% | 5% | |
| 6-7 | 16% | 20% | |
| 8-9 | 22% | 25% | |
| Practices in 2 or More States: % Yes | 5% | 2% | 0.405 |
| Paticipate in Pay-for-Performance: % Yes | 20% | 29% | 0.321 |
| % Activities in Primary Care: Mean (SD) | 77.6% (29) | 76.1% (30) | 0.76 |

| | | | |
|---|---|---|---|
| % Pts Insurance Medicare: Mean (SD) | 41.0% (27) | 45.0%(28) | 0.358 |
| % Pts Insurance Medicaid: Mean (SD) | 24.2% (25) | 21.7% (25) | 0.536 |
| % Time in Pts Care Paperwork: Mean (SD) | 14.8% (13) | 13.6% (12) | 0.385 |
| % Time in Practice Office: Mean (SD) | 72.1%(31) | 71.1%(29) | 0.848 |
| % Time Treating Cardiology: Mean (SD) | 21.7% (13) | 20.9% (15) | 0.704 |
| % Time Treating Endo: Mean (SD) | 17.1% (12) | 15.7% (11) | 0.454 |
| % Time Treating Geriatric: Mean (SD) | 8.8% (11) | 8% (8) | 0.588 |
| MOC Practice Improvement Module Performance Characteristics | | | |
| Versions of Module (chronological order) | | | 0.023 |
| ABIM 2008 | 41% | 22% | |
| CME360 v1 | 32% | 29% | |
| CME360 v2 | 19% | 35% | |
| CME360 v3 | 9% | 14% | |
| Days to Complete PIM: Mean (SD) | 198.2 (166) | 207.7 (162) | 0.717 |
| Days on QI in PIM: Mean (SD) | 87.6 (118) | 83.0 (105) | 0.799 |
| Total PIM Systems Score: Mean (SD) | 55.5 (21) | 59.4 (23) | 0.265 |

Table 2: Demographic, Health & Survey Characteristics of Patient Respondents

| Propensity Matching Variables | Spanish/ English Cases n=2,009 | English Controls n=2,048 | P-values |
|---|---|---|---|
| Personal Demographics | | | |
| Gender % Female | 59% | 58% | 0.757 |
| Age in years | | | <.001 |
| 18-44 years old | 31% | 23% | |
| 45-64 years old | 44% | 44% | |
| 65 years and older | 25% | 33% | |
| Education | | | <.001 |
| < High School | 18% | 10% | |
| High School grad or GED | 20% | 22% | |
| Some College | 28% | 30% | |
| College Grad. | 31% | 35% | |
| Skipped | 4% | 3% | |
| Race | | | <.001 |
| white | 47% | 57% | |
| black | 15% | 16% | |
| asian | 6% | 8% | |
| native American | 1% | 1% | |
| Other | 18% | 9% | |
| Skipped | 13% | 9% | |
| Office visits, health status, & tenure with physician | | | |

| | | | |
|---|---|---|---|
| Office visits/yr. | | | 0.005 |
| 1/yr. | 12% | 8% | |
| 2-4/yr. | 38% | 36% | |
| 5-6/yr. | 16% | 19% | |
| 7-12/yr. | 14% | 14% | |
| 13 or more/yr. | 4% | 5% | |
| Major Conditions Treated in practice | | | <.001 |
| Arthitis | 3% | 4% | |
| Cardiovascular Disease | 9% | 8% | |
| Depression | 2% | 1% | |
| Diabetes | 52% | 45% | |
| Hypertension | 30% | 36% | |
| Other | 1% | 0% | |
| None | 2% | 6% | |
| Tenure with physician | | | <.001 |
| ≥ 1 yr. & < 3 yrs. | 34% | 28% | |
| ≥ 3 yrs. & < 5 yrs. | 26% | 26% | |
| ≥ 5 yrs. | 41% | 46% | |
| Health Status | | | 0.033 |
| Fair-Poor | 21% | 18% | |
| Good | 33% | 35% | |
| Very good-Excellent | 48% | 46% | |
| Skipped | 1% | 1% | |
| Survey Characteristics | | | |
| Versions of Module (chronological order) | | | <.001 |
| ABIM 2008 | 40% | 21% | |
| CME360 v1 | 32% | 30% | |
| CME360 v2 | 19% | 36% | |
| CME360 v3 | 9% | 14% | |
| Data collection Mode | | | |
| IVR | 20% | 23% | 0.042 |
| Web | 80% | 77% | |

Table 3: Rank Correlation Matrices and Summary Statistics for the Three Study Groups Used in the Learning Sample

| | | Propensity Scores | | Communication | | | | | | Overall | Shared Decision Making | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Variable | Physician | Patients | Explains | Listens | Clear | Knows | Shows | Enough | Overall | Rx Pros & | Asks for |
| _TYPE_ | Names | Score | Score | things | Carefully | Instructions | Pt. Hx | Respect | Time | Rating | Cons | choice |
| **Spanish Cases: Learning Sample** | | | | | | | | | | | | |
| MEAN | | 2.404 | -1.980 | 4.700 | 4.814 | 4.286 | 4.679 | 4.936 | 4.714 | 9.671 | 0.907 | 0.871 |
| STD | | 2.296 | 0.654 | 0.707 | 0.557 | 1.114 | 0.742 | 0.299 | 0.660 | 0.948 | 0.291 | 0.336 |
| N | | 140 | 140 | 140 | 140 | 140 | 140 | 140 | 140 | 140 | 140 | 140 |
| CORR | xbeta_doc | 1.000 | 0.302 | -0.045 | -0.107 | 0.003 | -0.074 | -0.184 | -0.014 | 0.058 | -0.063 | -0.009 |
| CORR | xbeta_pts | 0.302 | 1.000 | -0.185 | -0.200 | -0.145 | -0.195 | -0.096 | -0.102 | 0.127 | -0.083 | -0.115 |
| CORR | c1 | -0.045 | -0.185 | 1.000 | 0.631 | 0.394 | 0.452 | 0.327 | 0.296 | 0.214 | 0.041 | 0.103 |
| CORR | c2 | -0.107 | -0.200 | 0.631 | 1.000 | 0.362 | 0.601 | 0.530 | 0.511 | 0.200 | 0.046 | 0.127 |
| CORR | c3 | 0.003 | -0.145 | 0.394 | 0.362 | 1.000 | 0.496 | 0.215 | 0.291 | 0.077 | 0.030 | 0.114 |
| CORR | c4 | -0.074 | -0.195 | 0.452 | 0.601 | 0.496 | 1.000 | 0.372 | 0.469 | 0.138 | 0.215 | 0.218 |
| CORR | c5 | -0.184 | -0.096 | 0.327 | 0.530 | 0.215 | 0.372 | 1.000 | 0.395 | 0.126 | 0.155 | 0.201 |
| CORR | c6 | -0.014 | -0.102 | 0.296 | 0.511 | 0.291 | 0.469 | 0.395 | 1.000 | 0.156 | 0.231 | 0.259 |
| CORR | doctor_scale | 0.058 | 0.127 | 0.214 | 0.200 | 0.077 | 0.138 | 0.126 | 0.156 | 1.000 | 0.019 | 0.074 |
| CORR | z2 | -0.063 | -0.083 | 0.041 | 0.046 | 0.030 | 0.215 | 0.155 | 0.231 | 0.019 | 1.000 | 0.612 |
| CORR | z3 | -0.009 | -0.115 | 0.103 | 0.127 | 0.114 | 0.218 | 0.201 | 0.259 | 0.074 | 0.612 | 1.000 |
| | | | | | | | | | | | | |
| **English Cases: Learning Sample** | | | | | | | | | | | | |
| MEAN | | 2.701 | -2.372 | 4.696 | 4.765 | 4.713 | 4.700 | 4.815 | 4.700 | 9.497 | 0.876 | 0.811 |
| STD | | 2.892 | 0.459 | 0.729 | 0.626 | 0.634 | 0.600 | 0.540 | 0.658 | 0.979 | 0.330 | 0.392 |
| N | | 710 | 710 | 710 | 710 | 710 | 710 | 710 | 710 | 710 | 710 | 710 |
| CORR | xbeta_doc | 1.000 | 0.389 | -0.105 | -0.034 | -0.134 | -0.078 | -0.085 | -0.091 | -0.033 | -0.115 | -0.074 |
| CORR | xbeta_pts | 0.389 | 1.000 | -0.127 | -0.022 | -0.062 | 0.004 | -0.023 | -0.046 | -0.022 | -0.136 | -0.117 |
| CORR | c1 | -0.105 | -0.127 | 1.000 | 0.556 | 0.574 | 0.506 | 0.506 | 0.475 | 0.369 | 0.218 | 0.150 |
| CORR | c2 | -0.034 | -0.022 | 0.556 | 1.000 | 0.592 | 0.539 | 0.566 | 0.525 | 0.412 | 0.211 | 0.160 |
| CORR | c3 | -0.134 | -0.062 | 0.574 | 0.592 | 1.000 | 0.575 | 0.561 | 0.554 | 0.427 | 0.273 | 0.257 |
| CORR | c4 | -0.078 | 0.004 | 0.506 | 0.539 | 0.575 | 1.000 | 0.572 | 0.518 | 0.440 | 0.193 | 0.197 |
| CORR | c5 | -0.085 | -0.023 | 0.506 | 0.566 | 0.561 | 0.572 | 1.000 | 0.584 | 0.437 | 0.196 | 0.215 |
| CORR | c6 | -0.091 | -0.046 | 0.475 | 0.525 | 0.554 | 0.518 | 0.584 | 1.000 | 0.422 | 0.195 | 0.209 |
| CORR | doctor_scale | -0.033 | -0.022 | 0.369 | 0.412 | 0.427 | 0.440 | 0.437 | 0.422 | 1.000 | 0.209 | 0.199 |
| CORR | z2 | -0.115 | -0.136 | 0.218 | 0.211 | 0.273 | 0.193 | 0.196 | 0.195 | 0.209 | 1.000 | 0.485 |
| CORR | z3 | -0.074 | -0.117 | 0.150 | 0.160 | 0.257 | 0.197 | 0.215 | 0.209 | 0.199 | 0.485 | 1.000 |
| | | | | | | | | | | | | |
| **English Controls: Learning Sample** | | | | | | | | | | | | |
| MEAN | | -0.003 | -2.901 | 4.641 | 4.779 | 4.719 | 4.622 | 4.797 | 4.659 | 9.422 | 0.862 | 0.799 |
| STD | | 1.646 | 0.468 | 0.765 | 0.551 | 0.564 | 0.716 | 0.531 | 0.655 | 0.980 | 0.345 | 0.401 |
| N | | 384 | 384 | 384 | 384 | 384 | 384 | 384 | 384 | 384 | 384 | 384 |
| CORR | xbeta_doc | 1.000 | -0.238 | -0.021 | -0.094 | -0.100 | -0.003 | -0.129 | -0.050 | 0.165 | 0.096 | 0.131 |
| CORR | xbeta_pts | -0.238 | 1.000 | -0.111 | -0.130 | -0.153 | -0.069 | -0.110 | -0.135 | -0.212 | -0.180 | -0.184 |
| CORR | c1 | -0.021 | -0.111 | 1.000 | 0.585 | 0.601 | 0.389 | 0.514 | 0.553 | 0.312 | 0.271 | 0.199 |
| CORR | c2 | -0.094 | -0.130 | 0.585 | 1.000 | 0.592 | 0.521 | 0.701 | 0.592 | 0.355 | 0.292 | 0.219 |
| CORR | c3 | -0.100 | -0.153 | 0.601 | 0.592 | 1.000 | 0.491 | 0.575 | 0.582 | 0.341 | 0.250 | 0.204 |
| CORR | c4 | -0.003 | -0.069 | 0.389 | 0.521 | 0.491 | 1.000 | 0.469 | 0.539 | 0.476 | 0.325 | 0.195 |
| CORR | c5 | -0.129 | -0.110 | 0.514 | 0.701 | 0.575 | 0.469 | 1.000 | 0.588 | 0.283 | 0.206 | 0.132 |
| CORR | c6 | -0.050 | -0.135 | 0.553 | 0.592 | 0.582 | 0.539 | 0.588 | 1.000 | 0.400 | 0.266 | 0.228 |
| CORR | doctor_scale | 0.165 | -0.212 | 0.312 | 0.355 | 0.341 | 0.476 | 0.283 | 0.400 | 1.000 | 0.263 | 0.316 |
| CORR | z2 | 0.096 | -0.180 | 0.271 | 0.292 | 0.250 | 0.325 | 0.206 | 0.266 | 0.263 | 1.000 | 0.460 |
| CORR | z3 | 0.131 | -0.184 | 0.199 | 0.219 | 0.204 | 0.195 | 0.132 | 0.228 | 0.316 | 0.460 | 1.000 |

Table 4: Rank Correlation Matrices and Summary Statistics for the Three Study Groups Used in the Testing Sample

| | | Propensity Scores | | Communication | | | | | | Overall | Shared Decision | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Variable | Physician | Patients | Explains | Listens | Clear | Knows | Shows | Enough | Overall | Rx Pros & | Asks for |
| _TYPE_ | Names | Score | Score | things | Carefully | Instructions | Pt. Hx | Respect | Time | Rating | Cons | choice |
| **Spanish Cases: Testing Sample** | | | | | | | | | | | | |
| MEAN | | -0.646 | -2.243 | 4.736 | 4.868 | 3.585 | 4.679 | 4.906 | 4.717 | 9.453 | 0.868 | 0.849 |
| STD | | 1.108 | 0.498 | 0.593 | 0.394 | 1.658 | 0.827 | 0.295 | 0.662 | 1.153 | 0.342 | 0.361 |
| N | | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 |
| CORR | xbeta_doc | 1.000 | 0.394 | -0.240 | 0.006 | -0.002 | 0.018 | -0.081 | -0.024 | -0.216 | -0.075 | -0.049 |
| CORR | xbeta_pts | 0.394 | 1.000 | -0.073 | 0.056 | -0.158 | 0.158 | 0.183 | 0.112 | -0.136 | 0.127 | 0.087 |
| CORR | c1 | -0.240 | -0.073 | 1.000 | 0.453 | 0.253 | 0.162 | 0.331 | 0.156 | 0.411 | 0.102 | 0.096 |
| CORR | c2 | 0.006 | 0.056 | 0.453 | 1.000 | 0.198 | 0.394 | 0.284 | 0.266 | 0.304 | 0.206 | 0.176 |
| CORR | c3 | -0.002 | -0.158 | 0.253 | 0.198 | 1.000 | 0.215 | 0.031 | -0.070 | 0.230 | 0.155 | 0.256 |
| CORR | c4 | 0.018 | 0.158 | 0.162 | 0.394 | 0.215 | 1.000 | 0.314 | 0.425 | 0.449 | 0.369 | 0.476 |
| CORR | c5 | -0.081 | 0.183 | 0.331 | 0.284 | 0.031 | 0.314 | 1.000 | 0.672 | 0.410 | 0.446 | 0.225 |
| CORR | c6 | -0.024 | 0.112 | 0.156 | 0.266 | -0.070 | 0.425 | 0.672 | 1.000 | 0.254 | 0.396 | 0.202 |
| CORR | doctor_scale | -0.216 | -0.136 | 0.411 | 0.304 | 0.230 | 0.449 | 0.410 | 0.254 | 1.000 | 0.393 | 0.237 |
| CORR | z2 | -0.075 | 0.127 | 0.102 | 0.206 | 0.155 | 0.369 | 0.446 | 0.396 | 0.393 | 1.000 | 0.770 |
| CORR | z3 | -0.049 | 0.087 | 0.096 | 0.176 | 0.256 | 0.476 | 0.225 | 0.202 | 0.237 | 0.770 | 1.000 |
| | | | | | | | | | | | | |
| **English Cases: Testing Sample** | | | | | | | | | | | | |
| MEAN | | -0.976 | -2.636 | 4.811 | 4.842 | 4.814 | 4.796 | 4.886 | 4.788 | 9.587 | 0.894 | 0.840 |
| STD | | 1.480 | 0.431 | 0.583 | 0.523 | 0.540 | 0.526 | 0.417 | 0.554 | 0.930 | 0.308 | 0.367 |
| N | | 387 | 387 | 387 | 387 | 387 | 387 | 387 | 387 | 387 | 387 | 387 |
| CORR | xbeta_doc | 1.000 | 0.411 | 0.042 | -0.042 | -0.026 | -0.032 | 0.061 | 0.036 | -0.004 | -0.033 | -0.021 |
| CORR | xbeta_pts | 0.411 | 1.000 | -0.052 | -0.092 | -0.118 | -0.087 | -0.048 | -0.135 | -0.139 | -0.084 | -0.049 |
| CORR | c1 | 0.042 | -0.052 | 1.000 | 0.512 | 0.521 | 0.456 | 0.464 | 0.406 | 0.337 | 0.345 | 0.277 |
| CORR | c2 | -0.042 | -0.092 | 0.512 | 1.000 | 0.631 | 0.574 | 0.601 | 0.395 | 0.470 | 0.381 | 0.404 |
| CORR | c3 | -0.026 | -0.118 | 0.521 | 0.631 | 1.000 | 0.529 | 0.523 | 0.410 | 0.475 | 0.352 | 0.301 |
| CORR | c4 | -0.032 | -0.087 | 0.456 | 0.574 | 0.529 | 1.000 | 0.527 | 0.444 | 0.415 | 0.286 | 0.271 |
| CORR | c5 | 0.061 | -0.048 | 0.464 | 0.601 | 0.523 | 0.527 | 1.000 | 0.527 | 0.448 | 0.459 | 0.365 |
| CORR | c6 | 0.036 | -0.135 | 0.406 | 0.395 | 0.410 | 0.444 | 0.527 | 1.000 | 0.481 | 0.419 | 0.212 |
| CORR | doctor_scale | -0.004 | -0.139 | 0.337 | 0.470 | 0.475 | 0.415 | 0.448 | 0.481 | 1.000 | 0.378 | 0.310 |
| CORR | z2 | -0.033 | -0.084 | 0.345 | 0.381 | 0.352 | 0.286 | 0.459 | 0.419 | 0.378 | 1.000 | 0.513 |
| CORR | z3 | -0.021 | -0.049 | 0.277 | 0.404 | 0.301 | 0.271 | 0.365 | 0.212 | 0.310 | 0.513 | 1.000 |
| | | | | | | | | | | | | |
| **English Controls: Testing Sample** | | | | | | | | | | | | |
| MEAN | | -1.458 | -2.936 | 4.713 | 4.747 | 4.721 | 4.665 | 4.808 | 4.703 | 9.410 | 0.841 | 0.778 |
| STD | | 0.927 | 0.338 | 0.624 | 0.564 | 0.609 | 0.618 | 0.505 | 0.604 | 1.005 | 0.366 | 0.416 |
| N | | 947 | 947 | 947 | 947 | 947 | 947 | 947 | 947 | 947 | 947 | 947 |
| CORR | xbeta_doc | 1.000 | -0.327 | 0.031 | 0.016 | -0.032 | 0.009 | 0.047 | -0.010 | -0.030 | -0.063 | 0.009 |
| CORR | xbeta_pts | -0.327 | 1.000 | -0.260 | -0.289 | -0.271 | -0.205 | -0.251 | -0.234 | -0.215 | -0.128 | -0.144 |
| CORR | c1 | 0.031 | -0.260 | 1.000 | 0.646 | 0.554 | 0.595 | 0.599 | 0.556 | 0.460 | 0.307 | 0.275 |
| CORR | c2 | 0.016 | -0.289 | 0.646 | 1.000 | 0.635 | 0.565 | 0.626 | 0.616 | 0.477 | 0.341 | 0.317 |
| CORR | c3 | -0.032 | -0.271 | 0.554 | 0.635 | 1.000 | 0.549 | 0.576 | 0.582 | 0.501 | 0.369 | 0.288 |
| CORR | c4 | 0.009 | -0.205 | 0.595 | 0.565 | 0.549 | 1.000 | 0.545 | 0.579 | 0.528 | 0.355 | 0.327 |
| CORR | c5 | 0.047 | -0.251 | 0.599 | 0.626 | 0.576 | 0.545 | 1.000 | 0.594 | 0.458 | 0.404 | 0.305 |
| CORR | c6 | -0.010 | -0.234 | 0.556 | 0.616 | 0.582 | 0.579 | 0.594 | 1.000 | 0.497 | 0.338 | 0.350 |
| CORR | doctor_scale | -0.030 | -0.215 | 0.460 | 0.477 | 0.501 | 0.528 | 0.458 | 0.497 | 1.000 | 0.351 | 0.321 |
| CORR | z2 | -0.063 | -0.128 | 0.307 | 0.341 | 0.369 | 0.355 | 0.404 | 0.338 | 0.351 | 1.000 | 0.503 |
| CORR | z3 | 0.009 | -0.144 | 0.275 | 0.317 | 0.288 | 0.327 | 0.305 | 0.350 | 0.321 | 0.503 | 1.000 |