Title

The Use of External Anchors When Examining Differences in Scale Performance in Patient Experience Surveys

Permalink

https://escholarship.org/uc/item/281909dg

Journal

Medical Care, 63(4)

ISSN

0025-7079

Authors

Abel, Gary A
Hays, Ron D
Campbell, John L
et al.

Publication Date

2025-04-01

DOI

10.1097/mlr.0000000000002135

Peer reviewed

The Use of External Anchors When Examining Differences in Scale Performance in Patient Experience Surveys

Gary A. Abel, PhD, Ron D. Hays, PhD,† John L. Campbell, MB MD,* and Marc N. Elliott, PhD‡*

Objectives: To present an example of using vignettes as an external anchor to assess measurement equivalence for patient experience measures.

Background: Evaluating measurement equivalence and differences in scale use is helpful for identifying disparities in patient experience based on patient surveys. External anchors, often in the form of scored vignettes, provide an attractive approach to examining differences in scale use but are not commonly used.

Methods: We analyzed a UK dataset based on the General Practice Patient Survey and a U.S. dataset based on the Consumer Assessment of Healthcare Providers and Systems Clinician and Group survey. A total of 560 White British and 560 Pakistani adults were recruited from various locations across England; 575 Asian American and 505 non-Hispanic White patients were recruited from an internet panel in the United States. Patient encounters and rated the quality of communication using 5 General Practice Patient Survey questions and 3 Consumer Assessment of Healthcare Providers and Systems Clinician and Group questions.

Results: Using an external anchor in both United States and UK data produced substantial evidence of differential item functioning (DIF). However, an "internal" DIF analysis (without an external anchor) produced little evidence of DIF.

Conclusions: Using an external anchor does not require the assumption made by internal methods that some items do not display between-group DIF. These assumptions may not hold for patient experience items if a single factor, such as extreme or negative response tendency, governs all items equally.

Key Words: patient experience, scale use, item-response theory, extreme response tendency

(*Med Care* 2025;63:311–316)

Patient experience surveys are widely used in health care. Examples include the Consumer Assessment of Healthcare Providers and Systems (CAHPS) family of surveys (eg, HCAHPS, MA/PDP CAHPS, clinician and group-CAHPS) in the United States, and the General Practice Patient Survey (GPPS) and the adult in-patient surveys in the UK.[1–3] The results from these surveys are often publicly reported and serve multiple uses, including to inform public/service users, quality inspections,[4] and pay-for-performance. Patient experience is weakly positively correlated with technical quality of care,[5–7] and both vary between different demographic subgroups.[5,7–9] Given the high-profile nature of these surveys, substantial research efforts have been put into understanding their psychometric properties.

Measurement equivalence across ethnic and racial groups is required to understand the extent to which observed differences in patient experience scores can be interpreted as health disparities.[10,11] In the UK, South Asian patients have been a focus of inquiry as a high-deprivation group with below-average clinical quality of care and the lowest average patient experience scores.[12–14] In the United States, Asian American patients have average income and socio-economic status levels that exceed the U.S. average,[15] and above-average clinical quality, but typically report the lowest average patient experience scores.[6,16]

The commonly used approach to assessing measurement equivalence requires identifying a subset of anchor items free of differential item functioning (DIF) to identify those with DIF. This has been shown to work for knowledge assessments,[17,18] but may not be plausible for evaluative items sharing a standard ordinal response scale.[19] This internal anchor approach cannot detect DIF when differential scale use between patient groups is consistent across all items.

An alternative approach to examining measurement equivalence uses vignettes as an external anchor of care.[10,20] Participants are presented with scenarios and asked to assess them using the same survey instruments used in patient experience studies. The quality of experience being rated by different patient groups is standardized. Thus, any differences in reported experience must be related to the patient's translation of their experience to responses on the instrument. Often, the vignettes have been designed to cover a range of patient experiences.

From the *Department of Health and Community Sciences, University of Exeter, St Luke's Campus Exeter, UK; †Department of Medicine, University of California, Los Angeles, Los Angeles, CA; and ‡RAND Health, Santa Monica, CA.

G.A.A. was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research. The remaining authors declare no conflict of interest.

Correspondence to: Marc N. Elliott, PhD, RAND, 1776 Main Street, Santa Monica, CA 90401. E-mail: elliott@rand.org.

DOI: 10.1097/MLR.0000000000002135

As such, there is an independent assessment of quality that is external to the ratings made by participants. However, vignettes are not typically included in patient experience surveys, and, as noted previously, DIF assessment relies on identified "anchor" items to represent underlying patient experience.[21]

Two pairs of studies have reached contradictory findings about measurement equivalence in patient experience surveys when using different approaches. In the UK setting, a paper using an internal DIF analysis found no evidence of DIF in South Asian patients compared with White British patients.[19] In contrast, the vignette-based approach suggested that South Asian patients were more likely to evaluate the same quality of care more positively than their White British counterparts.[20] In the U.S. setting, a study used structural equation models to evaluate relationships among Medicare CAHPS reports and ratings of care. Its authors concluded that there was measurement equivalence between different ethnic groups.[22] In contrast, the U.S.-based vignette study concluded that Asian American patients, whose distribution of national origins within Asia may differ substantially from the corresponding distribution in the UK, rated the same care more negatively than White people.[10] Thus, as different techniques have been used on different datasets, it is hard to draw firm conclusions, and we do not know which conclusions are valid.

METHODS

We estimated DIF using data collected in the 2 vignette-based studies discussed in the introduction and compared the results to an internal anchor-based assessment of DIF. In this manuscript, we use the term "ethnicity" in the UK sense and the official (2011 census-based) UK response categories[23] when referring to UK data.

Sample and Measures

Study 1 was conducted in the UK and investigated how South Asian and White British patients described the quality of doctor communication on the GPPS in a series of video vignette consultations. Five hundred sixty White British and 560 Pakistani participants were recruited from various locations across England. Each participant viewed 4 simulated consultation videos from a pool of 16 and answered 5 questions from the GPPS about the quality of communication in the videos. Each question had a 5-point ordinal response scale ranging from "very poor to very good." Video consultations were also rated by trained raters using the Global Consultation Rating Scale.[10,24]

Study 2 investigated whether 575 Asian patients exhibited less "extreme response tendency" than 505 U.S. White patients recruited from an internet panel. Each participant was presented with the same 5 written vignettes describing a doctor-patient encounter and rated the quality of communication using 3 questions based upon the CAHPS Clinician and Group survey: (1) To what extent did this doctor listen carefully to the patient? (2) To what extent did this doctor show respect for what the patient had to say? (3) To what extent did this doctor spend enough time talking to the patient about his headaches?[9] The vignettes differed in the degree to which the physician was responsive to the patient's concerns and were ranked from 1 (least responsive) to 5 (most responsive). Full details of the study designs and data collection for the 2 studies are given elsewhere.[10,20]

Analytic Methods

Analyses were conducted separately for data from the 2 studies but followed the same methods. Firstly, we wanted to ensure the same age distribution in the ethnic groups within the study, so that any DIF associated with age would not appear to provide evidence of DIF associated with ethnicity. All participants in the smallest ethnic by age subgroup with the fewest participants were retained. At the same time, a random sample was taken of participants in the larger ethnic group in the same age category to obtain the same number of participants (thus reducing the overall sample size).

External Anchor-Based Differential Item Functioning Assessment

We evaluated DIF using either the trained rater score of communication quality (UK study) or the rank of responsiveness of the physician to the patient's concerns (U.S. study) as an external anchor; this is the external comparison. Importantly, these scores/ranks were decided before participants evaluated the vignettes and were independent of the participant's ethnicity. In the case of the U.S.-based study, a cross-over interaction was seen in the original study such that Asian patients rated the lowest-ranked vignettes more positively than White participants and the highest-ranked vignettes less positively than White participants.[10] For this reason, the analysis of the U.S.-based data was conducted separately for vignettes ranked 1 and 2 (low responsiveness) and vignettes ranked 4 and 5 (high responsiveness). Sensitivity analyses used all vignettes rated by each participant. To account for each participant having assessed multiple vignettes, bootstrapping clustered by participants was performed with 1000 replications.

Internal Method-Based Differential Item Functioning Assessment

We used a Mantel-Haenszel method to look for evidence of DIF. For each possible cut-point on the scale, we tested for DIF first by using the sum of the 5 or 3 items (depending on the survey) to estimate underlying communication quality; this is the internal comparison. Further details, along with the strengths and weaknesses of this method, have been described previously.[25]

RESULTS

The analysis of the UK study and U.S. study data included 848 and 778 participants, respectively. In each case, half of the included participants were non-Hispanic White. Due to age matching, the age distribution was identical for Asian and White participants, though there

TABLE 1. Age, Gender, and Ethnicity of the Participants in the Analyses of the UK Study and U.S. Study Data

	UK study; n (%)		U.S. study; n (%)	
Gender/Age	**White**	**Asian**	**White**	**Asian**
Male	245 (57.8)	185 (43.6)	193 (0.5)	165 (0.4)
Female	179 (42.2)	239 (56.4)	196 (0.5)	224 (0.6)
18–24	40 (9.4)	40 (9.4)	47 (0.1)	47 (0.1)
25–34	56 (13.2)	56 (13.2)	55 (0.1)	55 (0.1)
35–44	70 (16.5)	70 (16.5)	67 (0.2)	67 (0.2)
45–54	57 (13.4)	57 (13.4)	85 (0.2)	85 (0.2)
55–64	94 (22.2)	94 (22.2)	83 (0.2)	83 (0.2)
65–74	70 (16.5)	70 (16.5)	46 (0.1)	46 (0.1)
75–84	32 (7.5)	32 (7.5)	6 (0.0)*	6 (0.0)*
85 or over	5 (1.2)	5 (1.2)	—	—

*In the U.S. study, the oldest age group was 75 or over.

were more females among the Asian participants in both studies (Table 1).

Table 2 shows the results of the DIF analyses of the UK study data, and Table 3 shows the results of the U.S. study data.

External Anchor-Based Differential Item Functioning Assessment

UK Data

Based on the trained rater score of communication quality, we found evidence of DIF for all 5 items for both intermediate cut points and for 3 out of 5 items for the "very poor" cut-point versus all other cut points. The odds ratios (ORs) for all items at all 3 of these cut points exceeded 1, indicating that more positive ratings were more likely to be given by South Asian participants, consistent with the findings of the original work using these data.[19]

U.S. Data

For one item, only one White participant used the highest response option in rating vignettes 1 and 2; therefore, estimating DIF for the "to a great extent" versus all other cut points for that item was impossible. When considering vignettes 1 and 2 (representing the least responsive physicians), we found evidence of DIF for all 3 items considering the "to a great extent" versus all other cut points, and 2 out of 3 items for the other 2 cut points. The ORs for all cut points and all items are <1, indicating Asian participants rated these vignettes more positively than White participants, consistent with the original study using these data.[10] When considering vignettes 4 and 5 (representing the most responsive physicians), we only found evidence of DIF for the "not at all" versus all other cut points. While the original study did find that White participants rated care more positively than Asian participants for these vignettes (consistent with the ORs found here), the difference was only statistically significant for vignette 5.

Internal Method-Based Differential Item Functioning Assessment

UK Data

Using the sum of all rated items (as the estimate of underlying experience), we found limited evidence of DIF.

TABLE 2. DIF Analysis Results From the UK Data*

	Internal comparison		External score comparison	
Item	***p***	**OR (95% CI)**	***p***	**OR (95% CI)**
"Very good" vs "good," "neither good nor poor," "poor" and "very poor"				
Giving you enough time	0.587	0.90 (0.61, 1.33)	0.264	0.80 (0.54, 1.19)
Listening to you	0.352	0.81 (0.52, 1.26)	0.853	1.04 (0.69, 1.56)
Explaining tests and treatments	0.004	1.78 (1.20, 2.64)	0.218	1.31 (0.85, 2.00)
Involving you in decisions about your care	0.108	1.41 (0.93, 2.13)	0.825	0.95 (0.62, 1.47)
Treating you with care and concern	0.001	0.51 (0.34, 0.76)	0.154	0.74 (0.50, 1.12)
"Very good" and "good" vs "neither good nor poor," "poor" and "very poor"				
Giving you enough time	0.572	0.89 (0.59, 1.34)	<0.001	2.44 (1.60, 3.72)
Listening to you	0.435	1.18 (0.78, 1.80)	<0.001	2.45 (1.58, 3.78)
Explaining tests and treatments	0.599	0.91 (0.63, 1.31)	0.001	1.93 (1.33, 2.81)
Involving you in decisions about your care	0.088	1.41 (0.95, 2.10)	<0.001	2.08 (1.39, 3.11)
Treating you with care and concern	0.203	0.76 (0.50, 1.16)	<0.001	2.03 (1.36, 3.01)
"Very good," "good" and "neither good nor poor" vs "poor" and "very poor"				
Giving you enough time	0.049	0.69 (0.47, 1.00)	0.037	1.59 (1.03, 2.47)
Listening to you	0.087	1.39 (0.95, 2.02)	<0.001	2.31 (1.49, 3.59)
Explaining tests and treatments	0.997	1.00 (0.70, 1.43)	<0.001	2.46 (1.62, 3.73)
Involving you in decisions about your care	0.080	1.42 (0.96, 2.10)	0.001	2.14 (1.38, 3.30)
Treating you with care and concern	0.180	0.75 (0.49, 1.14)	0.007	1.78 (1.17, 2.69)
"Very good," "good," "neither good nor poor" and "poor" vs "very poor"				
Giving you enough time	0.002	0.46 (0.28, 0.76)	0.680	1.19 (0.52, 2.71)
Listening to you	0.799	0.94 (0.57, 1.55)	0.045	2.04 (1.02, 4.07)
Explaining tests and treatments	0.661	0.90 (0.55, 1.47)	0.369	1.31 (0.73, 2.35)
Involving you in decisions about your care	0.691	0.91 (0.57, 1.45)	0.016	1.93 (1.13, 3.30)
Treating you with care and concern	<0.001	2.57 (1.64, 4.03)	0.002	2.23 (1.34, 3.72)

*Odds ratios >1 indicate South Asian participants were more likely to endorse categories reflecting better patient experience than White participants.
DIF indicates differential item functioning; OR, odds ratio.

TABLE 3. DIF Analysis Results From the U.S. Data*

		Internal comparison		External score comparison	
Vignette	**Item**	***P***	**OR (95% CI)**	***p***	**OR (95% CI)**
"Not at all," "very little" and "to some extent," vs "to a great extent"					
Vignette 1 and 2	… listen carefully to patient	0.063	1.83 (0.97, 3.47)	0.024	0.74 (0.57, 0.96)
	… show respect for what patient had to say	0.192	1.43 (0.84, 2.43)	0.001	0.66 (0.52, 0.84)
	… spend enough time talking with patient	0.006	0.48 (0.28, 0.81)	< 0.001	0.52 (0.41, 0.68)
"Not at all" and "very little" vs "to some extent" and "to a great extent"					
Vignette 1 and 2	… listen carefully to patient	0.088	1.85 (0.91, 3.76)	0.090	0.79 (0.60, 1.04)
	… show respect for what patient had to say	0.508	0.80 (0.41, 1.55)	< 0.001	0.53 (0.38, 0.75)
	… spend enough time talking with patient	0.228	0.58 (0.24, 1.40)	< 0.001	0.45 (0.30, 0.67)
"Not at all" vs "very little," "to some extent" and "to a great extent"					
Vignette 1 and 2	… listen carefully to patient	—	—	0.144§	0.49 (0.19, 1.27)§
	… show respect for what patient had to say	—	—	0.004§	0.22 (0.08, 0.63)§
	… spend enough time talking with patient	—	—	0.002§	0.14 (0.04, 0.49)§
"Not at all," "very little" and "to some extent," vs "to a great extent"					
Vignette 4 and 5	… listen carefully to patient	0.667†	1.40 (0.30, 6.48)†	0.995	1.00 (0.40, 2.47)
	… show respect for what patient had to say	0.010†	6.50 (1.55, 27.21)†	0.636	1.20 (0.56, 2.59)
	… spend enough time talking with patient	0.029†	0.24 (0.06, 0.86)†	0.418	0.80 (0.47, 1.37)
"Not at all" and "very little" vs "to some extent" and "to a great extent"					
Vignette 4 and 5	… listen carefully to patient	0.721‡	0.84 (0.32, 2.19)‡	0.566	1.12 (0.75, 1.67)
	… show respect for what patient had to say	0.546‡	1.30 (0.55, 3.10)‡	0.258	1.25 (0.85, 1.83)
	… spend enough time talking with patient	0.781‡	0.87 (0.33, 2.28)‡	0.456	1.12 (0.83, 1.52)
"Not at all" vs "very little," "to some extent" and "to a great extent"					
Vignette 4 and 5	… listen carefully to patient	0.092	1.70 (0.92, 3.13)	0.008	1.42 (1.10, 1.84)
	… show respect for what patient had to say	0.966	1.01 (0.53, 1.94)	0.040	1.30 (1.01, 1.66)
	… spend enough time talking with patient	0.058	0.53 (0.28, 1.02)	0.035	1.30 (1.02, 1.66)

*Odds ratios > 1 indicate White participants were more likely to endorse categories reflecting better patient experience than Asian participants.
†Only 377 of 1000 bootstrap replications achieved convergence and resulting p-values and confidence intervals should be treated with caution.
‡Only 862 of 1000 bootstrap replications achieved convergence and resulting p-values and confidence intervals should be treated with caution.
§Only 877 of 1000 bootstrap replications achieved convergence and resulting p-values and confidence intervals should be treated with caution.
DIF indicates differential item functioning; OR, odds ratio.

No evidence of DIF was found for the intermediate cut points (ie, "very good" and "good" vs "neither good nor poor," "poor" and "very poor" or "very good," "good" and "neither good nor poor" vs "poor" and "very poor"). For the extreme cut points ("very good" vs all others or "very poor" vs all others), evidence of DIF was found for 2 out of the 5 items examined; however, the direction of this DIF was inconsistent.

U.S. Data

As with the UK study, limited evidence of DIF was found in the U.S. data, with only 2 of 15 comparisons examined reaching statistical significance.

DISCUSSION

Different conclusions are reached from the same data when a DIF analysis is applied using external anchors or internal-based methods. The analysis using an external anchor provides results consistent with previous analyses of these data for both the UK and U.S. studies. This suggests that differences in findings between the two methods may be a function of the methods used rather than how the data are collected.

Popular approaches to estimating DIF use internal methods[17,18] that assume differences unrelated to the measured construct are not the same across items. The underlying assumptions may be reasonable in educational settings where DIF was developed, and items assess knowledge.[17,18] In contrast, with evaluative items such as patient experience measures, rather than testing knowledge, the respondent is making an evaluative judgment, which might be governed by scale use, especially if all items use the same response scale, as is typical for patient experience survey items.

The assumption may be unwarranted if, for example, extreme response tendency or negative response tendency governs all items equally.[26] There is evidence that this may hold for patient experience surveys. If different scale use is primarily governed by the form of responses rather than differences in the construct being evaluated, assuming that differences between groups do not represent differences in scale use may be incorrect; this is what was done in effect in the study reported by Setodji et al.[19] As they note, "the lack of evidence of DIF is consistent with either (1) no differences in expectations or scale use between the 2 racial/ethnic groups or (2) differences in expectations or scale use that were the same across all items. Similar differences in scale use across all items are plausible, given that the response scale and verbal labels were the same for all items, and is consistent with the evidence of different use of verbal labels in categorical (eg, Likert) response scales across cultures.[27] In one study,[22] associations between ratings of specific aspects of care and global ratings were compared between different ethnic groups, and are less prone to this mechanism. However, if response tendencies have some consistency across different response scales and types of items, it may still be hard to establish a lack of measurement equivalence when it exists. For example, there is evidence that Asian Americans are less likely to use both extremes of response scales than other respondents. Combined with the skewed distributions typical of patient experience scores, it has an asymmetric effect, reducing mean scores.[10,26]

In contrast, external anchor-based methods assume that the vignettes (treated as a gold standard) sufficiently represent patient experiences in the real world. Consistent with the findings of the vignette studies reported here, there is evidence of an aversion to extremes of response scale in Asian participants in the United States,[28,29] indicating a generic effect that is not specific to health care.

In terms of what this means for patient experience research, if we were to accept the findings of studies using internal methods, we would take survey results at face value, that is, that patient experience is by far the poorest for Asian American patients in the United States and for South Asian patients in the UK. However, our analysis suggests that these interpretations bear further examination. The analysis presented here casts doubt on the claims of measurement equivalence, and it is more likely that differential scale use explains the difference seen. External anchor-based approaches suggest that the experiences of Asian Americans in the United States may be somewhat better and those of South Asians in the UK markedly worse than would be implied by accepting measurement equivalence conclusions based on the assumptions of internal comparisons.

This study has both strengths and limitations. The main strength is applying the DIF methodology to a dataset from a vignette-based study to elucidate whether differences in findings derive from different analytic methods or experimental setups. A limitation of the study is that we used simple DIF methods. Our conclusions from the internal DIF analyses used a sum score rather than anchor items identified by purification, and we had low power for DIF detection because of the small number of items. Other DIF methods may have reached different conclusions.[30–33] External anchors are often difficult to obtain or unavailable for a given application, given they usually require a bespoke study to be conducted. Our findings suggest that when other methods are used, results should be interpreted with caution, especially in areas, such as patient experience surveys, where there is evidence that the assumptions of these methods may be unwarranted.

CONCLUSION

Patient experience research, including research on health equity, should rigorously test for measurement equivalence when comparing groups who may use evaluative response scales differently. Our results indicate the importance of comparing findings of internal DIF assessment with an external comparison approach when all items in a scale may have DIF.

ACKNOWLEDGMENTS

The authors thank Katherine Osby, BA, and Biayna Darabidian, MPP, for their help with the preparation of the manuscript.

REFERENCES

1. Department of Health and Human Services Consumer Assessment of Healthcare Providers and Systems. Agency For Healthcare Research and Quality. 2024. Accessed August 15, 2024. https://datatools.ahrq.gov/cahps
2. Health Services Advisory Group. 2024. hcahpsonline.org. hcahpsonline.org.
3. Orr N, Zaslavsky AM, Hays RD, et al. Development, methodology, and adaptation of the Medicare Consumer Assessment of Healthcare Providers and Systems (CAHPS) patient experience survey, 2007–2019. *Health Serv Outcomes Res Methodol*. 2023;23:1–20.
4. Commission CQ 2023. Accessed September 18, 2024. https://www.cqc.org.uk/guidance-providers/gps/how-we-monitor-gp-practices
5. Anhang Price R, Elliott MN, Zaslavsky AM, et al. Examining the role of patient experience surveys in measuring health care quality. *Med Care Res Rev*. 2014;71:522–554.
6. Martino S, Elliott M, Dembosky Jet al. *Disparities in health care in Medicare Advantage by race, ethnicity, and sex*. Baltimore, MD. 2022. Accessed on August 12, 2024. https://www.cms.gov/About-CMS/Agency-Information/OMH/research-and-data/statistics-and-data/stratified-reporting
7. Weech-Maldonado R, Elliott MN, Adams JL, et al. Do racial/ethnic disparities in quality and patient experience within Medicare plans generalize across measures and racial/ethnic groups? *Health Serv Res*. 2015;50:1829–1849.
8. Llanwarne NR, Abel GA, Elliott MN, et al. Relationship between clinical quality and patient experience: analysis of data from the English quality and outcomes framework and the National GP Patient Survey. *Ann Fam Med*. 2013;11:467–472.
9. Centers for Medicare & Medicaid Services. Stratified Reporting. Baltimore, MD. 2023. Accessed September 20, 2024. https://www.cms.gov/priorities/health-equity/minority-health/research-data/stratified-reporting
10. Mayer LA, Elliott MN, Haas A, et al. Less use of extreme response options by Asians to standardized care scenarios may explain some racial/ethnic differences in CAHPS scores. *Med Care*. 2016;54:38–44.
11. Weech-Maldonado R, Elliott MN, Oluwole A, et al. Survey response style and differential use of CAHPS rating scales by Hispanics. *Med Care*. 2008;46:963.
12. Burt J, Lloyd C, Campbell J, et al. Variations in GP–patient communication by ethnicity, age, and gender: evidence from a national primary care patient survey. *Br J Gen Pract*. 2016;66:e47–e52.
13. Government of the United Kingdom. People living in deprived neighbourhoods. 2023. Accessed August 28, 2024. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/people-living-in-deprived-neighbourhoods/latest
14. Lyratzopoulos G, Elliott M, Barbiere J, et al. Understanding ethnic and other socio-demographic differences in patient experience of primary care: evidence from the English General Practice Patient Survey. *BMJ Qual Safe*. 2012;21:21–29.
15. Gebeloff R, Lu D, Jordan M. Inside the diverse and growing Asian population in the US. *NY Times (Print)*. 2021;21.
16. Stratified Reporting. Centers for Medicare & Medicaid Services. 2023. Accessed April 12, 2024.
17. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*, 2. Sage. 1991.
18. Lord FM. *Applications of item response theory to practical testing problems*. 1st ed. Routledge. 2012.
19. Setodji CM, Elliott MN, Abel G, et al. Evaluating differential item functioning in the english general practice patient survey. *Med Care*. 2015;53:809–817.
20. Burt J, Abel G, Elmore N, et al. Understanding negative feedback from South Asian patients: an experimental vignette study. *BMJ Open*. 2016;6:e011256.

21. Bower P, Roland M, Campbell J, et al. Setting standards based on patients' views on access and continuity: secondary analysis of data from the general practice assessment survey. *Brit Med J*. 2003;326:258.
22. Hays RD, Chawla N, Kent EE, et al. Measurement equivalence of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) Medicare survey items between Whites and Asians. *Qual Life Res*. 2017;26:311–318.
23. Government of the United Kingdom. List of ethnic groups. 2022. Accessed August 14, 2024. https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups/
24. Burt J, Abel G, Elmore N, et al. Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview. *BMJ Open*. 2014;4:e004339.
25. Dorans NJ, Holland PW. DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series*. 1992;1992:i–40.
26. Elliott MN, Haviland AM, Kanouse DE, et al. Adjusting for subgroup differences in extreme response tendency in ratings of health care: impact on disparity estimates. *Health Serv Res*. 2009;44 (2p1):542–561.
27. Weech-Maldonado R, Elliott M, Pradhan R, et al. Can hospital cultural competency reduce disparities in patient experiences with care? *Med Care*. 2012;50:S48–S55.
28. Greenleaf EA. Improving rating scale measures by detecting and correcting bias components in some response styles. *J Mark Res*. 1992;29:176–188.
29. Greenleaf EA. Measuring extreme response style. *Public Opin Q*. 1992;56:328–351.
30. Teresi JA, Wang C, Kleinman M, et al. Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS) measures: methods, challenges, advances, and future directions. *Psychometrika*. 2021;86:674–711.
31. Kleinman M, Teresi JA. Differential item functioning magnitude and impact measures from item response theory models. *Psychol Test Assess Model*. 2016;58:79.
32. Doebler A. Looking at DIF from a new perspective: a structure-based approach acknowledging inherent indefinability. *Appl Psychol Meas*. 2019;43:303–321.
33. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Med Care*. 2006;44:S152–S170.

Erratum

Philadelphia Latine Immigrant Birthing People's Perspectives on Mitigating the Chilling Effect on Prenatal Care Utilization: Erratum

Diana Montoya-Williams, Alejandra Barreto, Alicia Laguna-Torres, Diana Worsley, Kate Wallis, Michelle-Marie Peña, Robin Ortiz, Lauren Palladino, Nicole Salva, Lisa Levine, Angelique Rivera, Rosalinda Hernandez, Elena Fuentes-Afflick, Katherine Yun, Scott Lorch, Senbagam Virudachalam

In the above-mentioned article that appeared on pages 404–415 of the June 2024 issue of *Medical Care*, a coauthor was inadvertently left off the author byline. Dr. Robin Ortiz, MD, MSHP, completed all the requirements for co-authorship and approved of this paper prior to publication. Here is the correct byline: Diana Montoya-Williams, Alejandra Barreto, Alicia Laguna-Torres, Diana Worsley, Kate Wallis, Michelle-Marie Peña, Robin Ortiz, Lauren Palladino, Nicole Salva, Lisa Levine, Angelique Rivera, Rosalinda Hernandez, Elena Fuentes-Afflick, Katherine Yun, Scott Lorch, Senbagam Virudachalam

REFERENCE

Montoya-Williams D, Barreto A, Laguna-Torres A, et al. Philadelphia Latine immigrant birthing people's perspectives on mitigating the chilling effect on prenatal care utilization. *Med Care*. 2024;62:404–415. doi: 10.1097/MLR.0000000000002002

DOI: 10.1097/MLR.0000000000002136