Quality of Life Research (2019) 28:3117–3135
https://doi.org/10.1007/s11136-019-02250-5

Differential item functioning of the CAHPS® In-Center Hemodialysis Survey

Claude M. Setodji[1] · John D. Peipert[2] · Ron D. Hays[1,3]

Accepted: 16 July 2019 / Published online: 26 July 2019

Abstract

Purpose End-stage renal disease patients' experience of care is an integral part of the assessment of the quality of the care provided at hemodialysis centers and is needed to promote patient choice, quality improvement, and accountability. The purpose of this study is to evaluate the In-Center Hemodialysis Consumer Assessment of Healthcare Providers and Systems (ICH-CAHPS®) survey and its equivalence in different age, gender, race, and education subgroups.

Methods The ICH-CAHPS survey was administered to 1454 patients from 32 dialysis facilities. For the characteristics compared, the sample had 756 participants younger than 65 years old, 739 men, 516 Black, 567 White, and 970 with less than high school diploma. Three different patient experience constructs were studied including nephrologist's communication and caring, quality of care and operations, and providing information to patients. We used item response theory analysis to examine the possibility of differential item functioning (DIF) by patient age, gender, race, and education separately after controlling for the other DIF characteristics and additional confounding variables including survey mode, mental, and general health status as well as duration on dialysis.

Results The three constructs studied were unidimensional and no major DIF was observed on the composites. Some non-equivalences were observed when confounders were not controlled for, suggesting that such covariates can be important factors in understanding the possibility of disparity in patients' experience.

Conclusions The ICH-CAHPS is a promising survey to elicit hemodialysis patients' experience that has good psychometric properties and provides a standardized tool for assessing age, gender, race, or education disparity.

Keywords CAHPS In-Center Hemodialysis (ICH) Survey · Patient experiences of care · Health care disparities · Measurement equivalence · Differential item functioning · Item response theory

Introduction

Kidney failure, or end-stage renal disease (ESRD), is a complex health condition that arises when a patient's kidneys can no longer remove enough excess fluid and toxins from the body to sustain life. Though kidney transplantation is the medically optimal treatment for patients with ESRD, the vast majority of ESRD patients receive dialysis treatments to compensate for their lack of kidney function, and this trend is likely to continue into the future [1]. In the United States, there were 703,243 Americans with end-stage renal disease (ESRD) in 2015 (of whom 124,114 were new cases of ESRD) with nearly 500,000 of them receiving some form of dialysis treatments and approximately 200,000 living with kidney transplants [2]. The United States Renal Data System (USRDS) reported that 63% of all ESRD patients received hemodialysis therapy,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-019-02250-5) contains supplementary material, which is available to authorized users.

✉ Claude M. Setodji
setodji@rand.org

[1] RAND Corporation, 4570, 5th Avenue, Suite 600, Pittsburgh, PA 15213-2665, USA

[2] Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

[3] Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

while only 7% were treated with peritoneal dialysis in 2015. Further, among the hemodialysis patients, 98% used in-center hemodialysis instead of receiving dialysis treatments in their home [1].

Women have been found to be more likely than men to receive dialysis for less than 12 h per week [3] and to have less access to kidney transplantation than men [4]. Frail, elderly patients with multimorbidity tend to benefit from dialysis less than younger patients [5, 6]. In addition, younger age, greater education attainment, and higher income are associated with physical health [7, 8] and some researchers such as Khattak and colleagues have shown that higher education level is associated with improved survival of patients on dialysis [9]. Also, Blacks are about 4 times more likely than Whites to develop ESRD [1]. Black dialysis patients have reported report better health-related quality of life (HRQOL) than White patients and Peipert and colleagues found differential item functioning between these groups in one of the HRQOL survey questions [10].

Patient experience of care surveys have been used to publicly report providers' performances and for "pay-for-performance" programs to promote patient choice, quality improvement, and accountability [11–13]. In the context of ESRD in the United States, the Centers for Medicare & Medicaid Services (CMS) in concert with the Agency for Healthcare Research and Quality (AHRQ) developed a Consumer Assessment of Healthcare Providers and Systems (CAHPS®) survey to assess the experiences of care provided to patients at in-center hemodialysis facilities: the CAHPS In-Center Hemodialysis Survey (ICH-CAHPS).

Measurement equivalence of the ICH-CAHPS survey is needed for comparisons of patient experiences from different ICH subgroups, especially if they are used to assess disparities associated with gender, age, race, and education. Differential item functioning (DIF) examines whether or not the likelihood of a survey question category endorsement is equal across different subgroups, controlling for the effect of the underlying construct being measured. In this study, we assessed equivalence of survey responses to the ICH-CAHPS survey for different age, gender, education, and race as isolated sources of DIF, characteristics that have been reported to show disparities in hemodialysis care. We used classical test theory and item response theory (IRT) modeling to test whether ICH-CAHPS survey questions performed differently for men compared to women, younger (18–64) compared to older (65 + years of age), white compared to Black and less educated (High school degree or less) compared to those with more than a high school diploma. The focus in this paper on Black versus White race differences reflects the importance of this comparison in the United States. We discuss the implications of the results of the study for survey users and quality improvement.

Methods

We analyzed data from 1454 patients sampled from 32 dialysis facilities from different geographical locations (Northeast, South, Midwest, West, and rural vs urban) and facility size and type. Patients were eligible to participate if they had received hemodialysis for 3 or more months at one of the selected facilities. A random, systematic sample of 200 patients was drawn from large facilities, and for smaller facilities with up to 200 patients, a census of all patients was included in the sample. Details of the development of the survey including piloting, interviewing, sampling, and detailed response rates were reported elsewhere [14]. All data collection were approved by the RAND institutional review boards (FWA00003425, effective until June 22, 2023). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

ICH-CAHPS: patient experience measures

The ICH-CAHPS survey measures patient experience in the last 3 months at in-center hemodialysis facilities and features three composites: Nephrologist Communication and Caring (6 items, e.g., "In the past 3 months, how often did your kidney doctors explain things in a way that was easy for you to understand?"), Quality of Dialysis Facility Care and Operations (17 items, e.g., "In the last 3 months, how often did you feel the dialysis center staff really cared about you as a person?"), and Providing Information to Patients (9 items, e.g., "In the last 12 months, has a doctor or dialysis center staff explained to you why you are not eligible for a kidney transplant?"). Items were administered with either four response options ("never," "sometimes," "usually," "always") or two options ("Yes," "No"). The full list of items (questions) administered is provided in Online Appendix 1. The survey also asked about age, gender, race, education, mental health status, overall health status, and duration on dialysis. The survey was collected in two modes, by telephone or by mail.

Analysis plan

We used χ^2 and t tests to compare characteristics by gender, age, education, and race subgroups. Because observed differences in survey items can be confounded by other variables beyond the DIF variables examined [15] and with our goal to assess these characteristics as isolated sources of DIF, we used propensity score matching [16] to create one-to-one matches for the DIF variables examined. The goal of the propensity score matching is to make the groups to be compared (e.g., male vs female) similar on all other characteristics as if they were randomly selected to participate in the study and representative of their groups. The characteristics matched on included participants age, gender, race/ethnicity, education, health status, the length of time on dialysis as well as the mode of the survey, and the matching algorithm used a caliper of 0.2, as recommended by Wang et al. [17]. A sensitivity analysis was also conducted with the full sample of participants (matched and unmatched) to assess the generalizability of the inferences beyond the matched sample. To account for the different level of missingness (most because of non-eligibility), a full-information maximum-likelihood method was used for all the DIF analyses, a method that uses all the available data [18].

We evaluated whether the three patient experience composites (nephrologist communication and caring; quality of dialysis facility care and operations; providing information to patients) were sufficiently unidimensional for IRT analyses [19]. A categorical one-factor confirmatory factor analysis (CFA) was fitted, and the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) were used to assess model fit. For CFA, missingness is allowed to be a function of the observed items, analogous to a pairwise analysis. A bi-factor CFA with three local factors, each representing an ICH-CAHPS composite was also conducted as a sensitivity of the unidimensionality of the composites. CFI value > 0.95 generally indicates an acceptable fit to the data, while RMSEA values < 0.06 are considered good fit [20].

An IRT DIF analysis comparing groups of respondents (men vs women; age 65 and younger vs older than 65; up to high school degree vs more than high school) was conducted using Samejima's graded response model [21] and DIF evaluated using a two-step improved Wald test procedure [22, 23]. Since no prespecified anchor items were designed with the survey, the two-step calibration process first evaluated the ICH-CAHPS items for evidence of DIF using the Wald-2 anchor-all-test procedure. This process classifies items into two categories, the anchor items (the ones that display no DIF) and the candidate DIF items. Then in a second step, for items labeled as candidate DIF, separate parameters were estimated across the characteristic being tested conditioned on the estimated group score using the anchor set with the Wald-1 procedure. For the improved Wald tests used, the Wald-2 test does not require designated anchor items and has been recommended for the selection of anchor items while the Wald-1 test, like other DIF procedures, requires specified anchor items [23]. For the anchor items, the estimates for the first step were reported and for the candidate items, the parameter estimates for the second step were provided. Because multiple DIF tests were conducted (one for each item) and to avoid type I error inflation, the Benjamini–Hochberg multiple comparison adjustment was applied to all the DIF tests [24, 25]. Goodness-of-fit statistics were used to assess the adequacy of the IRT models.

Uniform DIF assesses whether one group is consistently more likely than another to endorse a survey item at each level of the patient experience trait, e.g., care quality. Non-uniform DIF is observed when there is cross-over, so that at certain levels of the trait, one group is more likely to endorse the item, while at other levels, the other group is more likely to endorse the same item. The DIF tests reported were supported by magnitude measures [26] assessed using the non-compensatory DIF (NCDIF) index [27, 28]. An NCDIF threshold of 0.054 and 0.006 has been recommended as reasonable cutoff for small differences for items with four and two response options, respectively [29]. For each DIF analysis, an item characteristics curve (ICC) that presents the relationship between a latent trait and the likelihood of endorsing a specific level of an item, superimposed by groups being compared, is used to represent the differences between the groups. We also report the differential test functioning (DTF) index [27], a summary of the compensatory DIF (CDIF) across items in a construct that vary by the number of response options. The cutoff for DTF, on the other hand, is computed by summing the NCDIF cutoff values for all the candidate items in a construct. For the number of items and response options in the different constructs, if they were all candidate items, a threshold of 0.324, 0.726, and 0.054 is recommended for the communication and caring, the quality of care and operation, and the providing information to patient constructs, respectively. We use test characteristic curves (TCC) depicting the expected scale scores of the items as a function of patient experience on the IRT scale for two groups.

Confirmatory factor analyses were conducted using Mplus [30]; the propensity score and sample matching was produced in SAS version 9.4 (SAS Institute Inc., Cary, NC, USA); and the IRT DIF analyses were conducted in flexMIRT [31]. For the analyses, we created a SAS program that generates the flexMIRT codes and produces the outputs and plots from SAS commands.

Results

Patient characteristics and differences across groups

The sample of 1454 patients in the study was 51% male, 39% White, 35% Black, 13% Hispanic, and 12% other races; 67% with less than a high school diploma (see Table 1). Most of them reported good (34%) or very good (24%) mental health, good (34%) or fair/poor (45%) general health, and on average they were on dialysis for three years. Men (n = 739) and women (n = 708) in the sample have similar race/ethnicity, age, education, and years on dialysis distributions but men were more likely than women to report excellent and very good mental and general health. Different age subgroups were similar on gender and time on dialysis. The younger participants (age less

Table 1 Characteristics of patient population overall and differences between groups on non-matched and matched sample

Variable	Overall % or mean	Gender				Age				Education				Race: White vs Black			
		Overall ($n = 1447$)		Matched ($n = 1246$)		Overall ($n = 1449$)		Matched ($n = 990$)		Overall ($n = 1443$)		Matched ($n = 836$)		Overall ($n = 1083$)		Matched ($n = 696$)	
		Male–female	p	Male– female	p	Young–older	p	Young–older	p	Less–more than HS	p	Less–more than HS	p	Black–white	p	Black–white	p
Gender %																	
Male	51.07					2.0	0.45	1.2	0.70	− 3.0	0.28	2.2	0.53	− 6.49	**0.03**	1.44	0.70
Female	48.93					− 2.0		− 1.2		3.0		− 2.2		6.49		− 1.44	
Race %																	
White	39	3.5	0.10	1.1	0.97	− 18.3	**0.00**	1.0	0.93	− 13.9	**0.00**	− 0.7	0.99				
Black	35.49	− 6.1		− 0.2		14.3		0.6		4.5		1.0					
Hispanics	13.07	0.6		− 0.3		3.3		− 0.4		10.6		− 0.2					
Other race	12.45	2.1		− 0.6		0.7		− 1.2		− 1.2		0.0					
Age %																	
< 65	52.17	2.0	0.45	1.3	0.65					− 16.2	**0.00**	− 0.2	0.94	21.78	**0.00**	0.57	0.88
65 + old	47.83	− 2.0		− 1.3						16.2		0.2		− 21.78		− 0.57	
Education %																	
< HS	67.22	− 2.7	0.28	− 1.0	0.72	− 14.3	**0.00**	1.0	0.73					10.58	**0.00**	0.29	0.94
HS +	32.78	2.7		1.0		14.3		− 1.0						− 10.58		− 0.29	
Mental health status %																	
Excellent	18.24	4.1	**0.01**	− 0.2	0.99	6.4	**0.01**	− 0.6	0.98	− 9.0	**0.00**	− 3.4	0.65	0.7	0.82	0.29	0.70
Very good	25.66	3.9		0.0		− 0.8		− 0.4		− 4.7		1.9		− 0.89		− 2.88	
Good	34.19	− 2.5		− 0.5		− 5.4		1.2		7.8		1.9		− 1.85		3.74	
Fair or poor	21.91	− 5.5		0.6		− 0.3		− 0.2		5.9		− 0.5		2.04		− 1.15	
General health status %																	
Excellent	5.74	1.8	**0.05**	0.0	0.98	3.5	**0.02**	0.4	0.86	1.0	0.67	0.0	0.99	2.09	**0.03**	0	0.96
Very good	15.71	3.6		− 0.8		1.2		− 1.4		− 1.2		− 0.7		3.67		− 1.15	
Good	33.98	0.0		0.6		− 2.3		2.0		− 2.0		0.2		2.35		1.73	
Fair or poor	44.84	− 5.5		0.2		− 2.4		− 1.0		2.3		0.5		− 8.1		− 0.57	
Survey mode %																	
Mail	43.47	− 0.6	0.81	1.5	0.61	− 10.0	**0.00**	0.4	0.90	3.0	0.28	− 2.9	0.40	− 7.31	**0.02**	0.57	0.88
Phone	56.53	0.6		− 1.5		10.0		− 0.4		3.0		2.9		7.31		− 0.57	
Years on dialysis (mean)	2.94	− 0.03	0.35	− 0.00	0.82	0.00	0.95	− 0.01	0.79	0.06	0.06	0.01	0.78	0.25	**0.00**	− 0.03	0.56

The matched samples are one-to-one matches and have the same number of participants in each group. For the overall sample gender has 739 men and 708 women, age has 756 less than 65 years old and 693 more than 65 years old, education has 970 with less than high school and 473 less than high school, and race has 567 white and 516 black

Bold numbers are for statistical significance at 0.05 significnace level

HS high school, *Young* less than 65 years old, *older* more than 65 years old than 65, $n=756$) were more likely to be minorities, to be more educated, to have excellent mental health, and to have excellent or very good general health than older participants (age more than 65, $n=693$). On the other hand, less educated (less than high school diploma, $n=970$) and more educated ($n=473$) patients were similar on gender, general health, and years on dialysis. The less educated were also less likely to be White, to be younger, and to have excellent or very good mental health. After propensity score matching, all these observed group differences became negligible with 624 men matched to 624 women, 501 younger patients matched to 501 older patients, and 419 less educated matched to 419 more educated patients.

Table 2 Summary DIF analysis for the ICH-CAHPS items: gender, age, race and education groups

Construct	Item description	Type of DIF, if present				DIF magnitude (NCDIF)			
		Gender	Age	Educ.	Race	Gender	Age	Educ.	Race
Nephrologist's Communication and caring (6 items)	Q3: Doctor listens carefully		A		A				
	Q4: Doctor explains things		U		U		0.015		0.005
	Q5: Doctor shows respect		A		A				
	Q6: Doctor spends enough time		A		U				0.004
	Q7: Doctor cared about you		NU		A		0.017		
	Q10: Doctor seemed informed		A		A				
Quality of care and operations (17 items)	Q11: Staff listens carefully			U	A			0.020	
	Q12: Staff explains in a way that is easy to understand	A		U	A			0.023	
	Q13: Staff shows respect	A		A	A				
	Q14 Staff spent enough time	A		A	A				
	Q15: Staff cared about you	NU		A	A	0.005			
	Q16: Staff makes you comfortable	A		A	A				
	Q17: Staff keep information private	A		A	A				
	Q20: Comfortable asking staff	A		A	A				
	Q24: Staff inserts needle w/o pain	U		A	NU	0.047			0.094
	Q25: Staff checks you closely	A		A	A				
	Q27: Staff manages problems	A		A	A				
	Q29: Staff behaves professionally	A		A	A				
	Q31 rev: Staff explains test result	A		A	A				
	Q32: Staff discuss diet enough	A		A	A				
	Q40: On machine within 15 min	A		U	NU			0.002	0.028
	Q42: Center is clean	A		A	A				
	Q51: Satisfied with ways problems handled	A		A	A				
Providing information to patients (9 items)	Q22: Knows to take care of dialysis connection	A	U		A		0.005		
	Q33: Staff gives information on patient rights	A	A		A				
	Q34: Staff reviews patient rights	A	A		A				
	Q35: Staff told you what to do if health problem at home	A	A		U				0.018
	Q36: Staff told you how to get off machine if emergency	NU	U		A	0.004	0.006		
	Q44: They talked about Which treatment is right for you	A	A		A				
	Q46: They explain why you are not eligible for transplant	A	A		U				0.025
	Q47: Talk about peritoneal dialysis	A	A		A				
	Q48: Involved in choosing treatment	A	A		A				

Many of the observed DIFs were non-significant at 0.05 level after multiple test adjustment

U uniform DIF, *NU* non-uniform DIF, *A* used as anchor item in the estimation

Table 3 Estimated item parameters for graded response model: comparison of gender (male vs female)

ICH-CAHPS items		Slope	Thresholds			DIF test: χ^2 (*p* value)	
		a	b_1	b_2	b_3	*a* DIF	*b* DIF
Nephrologist communication and caring (6 items), no DIF candidate item found, DTF = 0.0057							
Q3: Doctor listens carefully	Female	3.62	− 2.06	− 0.84	− 0.20	NS	NS
	Male	3.81	− 2.08	− 0.85	− 0.16		
Q4: Doctor explains things	Female	2.55	− 1.78	− 0.67	− 0.02	NS	NS
	Male	2.92	− 1.57	− 0.71	− 0.06		
Q5: Doctor shows respect	Female	3.30	− 1.96	− 0.93	− 0.38	NS	NS
	Male	3.23	− 1.99	− 1.00	− 0.30		
Q6: Doctor spends enough time	Female	3.42	− 1.59	− 0.42	0.26	NS	NS
	Male	2.98	− 1.69	− 0.50	0.20		
Q7: Doctor cared about you	Female	3.69	− 1.92	− 0.81	− 0.29	NS	NS
	Male	3.26	− 1.84	− 0.89	− 0.24		
Q10: Doctor seemed informed	Female	2.02	− 1.31			NS	NS
	Male	1.93	− 1.26				
Quality of dialysis facility care and operations (17 items), DTF = 0.0456							
Items candidate for DIF (Wald-1 estimates)							
Q11: Staff listens carefully	Female	3.33	− 2.45	− 1.02	− 0.25	NS	NS
	Male	4.06	− 1.95	− 0.93	− 0.23		
Q15: Staff cared about you	Female	4.53	− 1.89	− 0.86	− 0.25	5.4 (0.02)	NS
	Male	3.47	− 1.91	− 0.99	− 0.28		
Q24: Staff inserts needle w/o pain	Female	0.92	− 3.46	− 0.90	0.98	NS	15.9 (0.001)
	Male	1.03	− 3.73	− 1.37	0.41		
Items identified as anchor (Wald-2 anchor selection estimates)							
Q12: Staff explains in a way that is easy to understand	Female	2.76	− 2.10	− 1.05	− 0.23	NS	NS
	Male	3.02	− 1.93	− 1.01	− 0.16		
Q13: Staff shows respect	Female	3.87	− 1.93	− 0.97	− 0.28	NS	NS
	Male	3.66	− 2.11	− 1.00	− 0.26		
Q14 Staff spent enough time	Female	3.02	− 2.03	− 0.90	− 0.05	NS	NS
	Male	3.25	− 1.94	− 0.83	0.00		
Q16: Staff makes you comfortable	Female	3.32	− 2.27	− 1.18	− 0.48	NS	NS
	Male	3.25	− 2.27	− 1.16	− 0.37		
Q17: Staff keep information private	Female	1.47	− 2.18			NS	NS
	Male	1.69	− 2.26				
Q20: Comfortable asking staff	Female	2.04	− 1.76			NS	NS
	Male	1.99	− 1.85				
Q25: Staff checks you closely	Female	2.23	− 2.82	− 1.13	− 0.20	NS	NS
	Male	2.29	− 2.62	− 1.22	− 0.16		
Q27: Staff manages problems	Female	2.63	− 2.60	− 1.32	− 0.50	NS	NS
	Male	2.73	− 2.70	− 1.35	− 0.63		
Q29: Staff behaves professionally	Female	2.58	− 2.54	− 1.32	− 0.47	NS	NS
	Male	2.51	− 2.52	− 1.51	− 0.52		
Q31 rev: Staff explains test result	Female	1.44	− 0.41			NS	NS
	Male	1.79	− 0.29				
Q32: Staff discuss diet enough	Female	1.28	− 1.91			NS	NS
	Male	1.14	− 2.21				
Q40: On machine within 15 min	Female	1.11	− 3.19	− 1.04	0.54	NS	NS
	Male	0.90	− 3.98	− 1.49	0.48		
Q42: Center is clean	Female	1.37	− 3.68	− 2.02	− 0.60	NS	NS
	Male	1.19	− 3.99	− 2.31	− 0.51		

Table 3 (continued)

ICH-CAHPS items		Slope	Thresholds			DIF test: χ^2 (*p* value)	
		a	b_1	b_2	b_3	*a* DIF	*b* DIF
Q51: Satisfied with ways problems handled	Female	1.81	− 1.89	− 0.50	0.57	NS	NS
	Male	1.54	− 1.87	− 0.60	0.63		
Providing information to patients (9 items), DTF = 0.0035							
Items candidate for DIF (Wald-1 estimates)							
Q36: Staff told you how to get off machine if emergency	Female	0.93	− 1.66			3.9 (0.049)	NS
	Male	0.53	− 2.70				
Items identified as Anchor (Wald-2 anchor selection estimates)							
Q22: Knows to take care of dialysis connection	Female	1.31	− 1.69			NS	NS
	Male	1.12	− 1.95				
Q33: Staff gives information on patient rights	Female	2.04	− 1.49			NS	NS
	Male	1.58	− 1.63				
Q34: Staff reviews patient rights	Female	1.89	− 0.82			NS	NS
	Male	1.65	− 0.83				
Q35: Staff told you what to do if health problem at home	Female	1.82	− 1.04			NS	NS
	Male	1.53	− 1.19				
Q44: They talked about Which treatment is right for you	Female	2.25	− 0.81			NS	NS
	Male	2.52	− 0.85				
Q46: They explain why you are not eligible for transplant	Female	0.96	− 0.45			NS	NS
	Male	1.01	− 0.29				
Q47: Talk about peritoneal dialysis	Female	0.70	0.31			NS	NS
	Male	1.09	0.12				
Q48: Involved in choosing treatment	Female	1.66	− 1.50			NS	NS
	Male	2.31	− 1.17				

All the observed DIFs were non-significant at 0.05 level after multiple test adjustment

DTF construct-level differential test functioning

Assessment of factor structure

Adequate internal consistency and center level reliability of the ICH-CAHPS in the three survey domains were reported previously [14, 32]. In the current analyses, we found support for the unidimensionality of the three ICH-CAHPS scales with both the CFI and RMSEA indices from confirmatory factor analyses (using diagonally weighted least squares—WLSMV—in Mplus) indicating reasonably good fit. The CFIs for communication and caring, quality of care and operations, and providing information to patients were 0.988, 0.983, and 0.900 while the RMSEAs were 0.040, 0.044, and 0.075, respectively. Although the CFI and RMSEA for the composite on providing information were close but did not meet the threshold for good fit, the results of the bi-factor CFA model ($\chi^2 = 1230.09$; $df = 432$; $p < 0.001$) supported unidimensionality across the constructs with a CFI of 0.984 and RMSEA of 0.038.

Across the studied items, a small number of missing data were observed on eligible participants and 2% or less of eligible patients were dropped from each item analysis using full-information maximum-likelihood estimation. Out of the 32 items, only four that required eligibility have more than 3% missing data. Without accounting for eligibility to the specific questions, those four items have 25% (for Q25, "Staff checks you closely") to 76% (for Q51, "Satisfied with ways problems were handled") missing data but conditional on a participant being eligible to a specific question, the missing rate on those items was less than 2%. Results (not shown) from IRT model fit and frequencies of the different answer categories (never, sometimes, usually, always) for the item (Q31) "in the last 3 months, how often did dialysis center staff explain blood test results in a way that was easy to understand" revealed low levels of endorsement of all response options except for the "always" option the item characteristic curve supported collapsing the categories never, sometimes and usually together. Hence, for the remainder of the analysis, Item Q31 was coded as "not always" (1) and always (2).

Gender DIF

None of the nephrologists' communication and caring items showed DIF between men and women (See Table 2) and the DTF index was 0.0057, smaller than recommended cutoff of

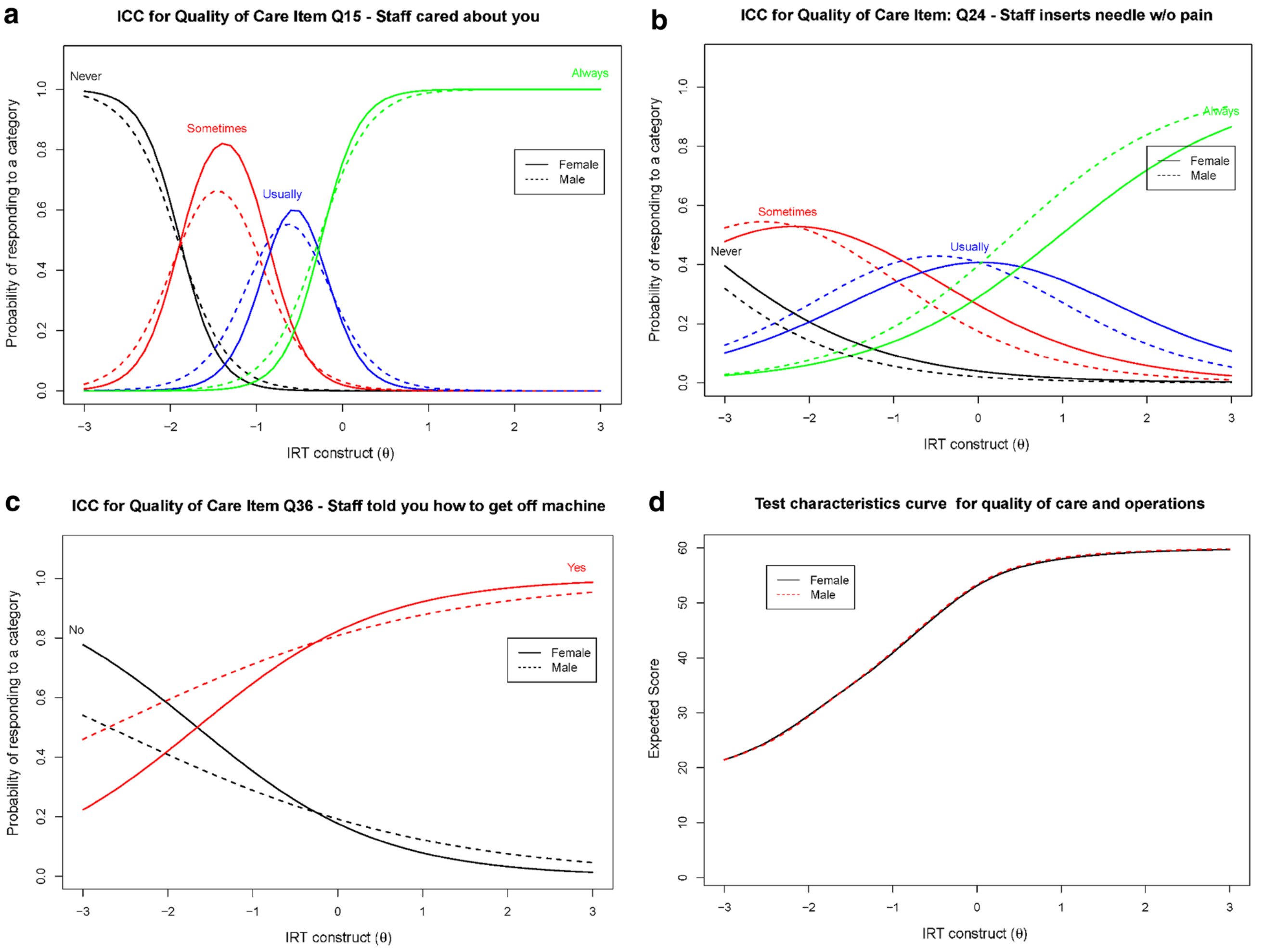

Fig. 1 Response curve for gender on selected candidate items and TCC curve for quality of care and operations

0.276. Nevertheless, prior to the Benjamini–Hochberg correction, in the care and operations construct, a uniform DIF was observed for the item Q24 "how often did dialysis center staff insert your needles with as little pain as possible" where men had a lower difficulty estimates (See exact estimates in Table 3). For the same construct, non-uniform DIF was observed on the item Q15 "how often the dialysis center staff really cared about you as a person" while the item Q11 "how often did the dialysis center staff listen carefully to you" did not show any DIF after the Wald-1 test. On the NCDIF, the DIF magnitude for those 2 items Q15 and Q24 was of small size (0.005 and 0.047, respectively); values smaller than the 0.054 recommended cutoff for items with four response options [29]. Also, at the trait level, the DTF index was 0.0456 and smaller than the recommended cutoff of 0.163 for 3 candidate items. For the final trait of providing information to patients, only one non-uniform DIF was observed in the item Q36, "has any dialysis center staff ever told you how to get off the machine if there is an emergency at the center," where women have slightly larger discrimination. The NCDIF magnitude was very small (0.004) for this dichotomous item. The ICC for the selected items which is reported in Fig. 1a–c and Fig. 1d presented the TCC for the quality of care and operations. The TCC showed no practical difference between the two groups, and a similar TCC plot was also observed in the providing information to patients construct. Most of these observed differences remained significant after adjustment for multiple comparisons, but all the DIF magnitudes were below the recommended cutoff for practical DIF.

Age DIF

Table 4 reported the exact estimates of the age DIF analysis. For the comparison between patients younger than 65 years old and the ones 65 year or older, two DIF items were observed in the communication and caring trait prior to multiple comparison adjustments. A non-uniform DIF was observed for the item Q7 "how often did you feel your kidney doctors really cared about you as a person," with the

Table 4 Estimated item parameters for graded response model: comparison of age (less than 65 vs 65 or more)

ICH-CAHPS items		Slope	Thresholds			DIF test: χ^2 (p value)	
		a	b_1	b_2	b_3	a DIF	b DIF
Nephrologist communication and caring (6 items), DTF = 0.0111							
Items candidate for DIF (Wald-1 estimates)							
Q4: Doctor explains things	Young	3.09	− 1.81	− 0.78	− 0.20	NS	12 (0.007)
	Older	2.79	− 1.56	− 0.74	− 0.03		
Q7: Doctor cared about you	Young	4.25	− 1.78	− 0.82	− 0.32	4.4 (0.036)	NS
	Older	3.25	− 1.97	− 1.02	− 0.41		
Items identified as Anchor (Wald-2 anchor selection estimates)							
Q3: Doctor listens carefully	Young	3.75	− 2.05	− 0.92	− 0.29	NS	NS
	Older	3.68	− 2.16	− 0.96	− 0.30		
Q5: Doctor shows respect	Young	3.21	− 2.10	− 1.04	− 0.41	NS	NS
	Older	4.00	− 1.79	− 0.98	− 0.44		
Q6: Doctor spends enough time	Female	2.92	− 1.73	− 0.54	0.13	NS	NS
	Male	3.44	− 1.67	− 0.56	0.12		
Q10: Doctor seemed informed	Young	2.07	− 1.23			NS	NS
	Older	2.11	− 1.34				
Quality of dialysis facility care and operations (17 items), No DIF candidate item found, DTF = 0.0514							
Q11: Staff listens carefully	Young	3.53	− 2.27	− 1.01	− 0.27	NS	NS
	Older	3.79	− 2.04	− 1.03	− 0.32		
Q12: Staff explains in a way that is easy to understand	Young	3.13	− 1.92	− 1.09	− 0.31	NS	NS
	Older	2.89	− 1.94	− 0.95	− 0.17		
Q13: Staff shows respect	Young	4.00	− 2.10	− 0.95	− 0.34	NS	NS
	Older	4.03	− 1.93	− 1.05	− 0.30		
Q14 Staff spent enough time	Young	3.17	− 2.03	− 0.91	− 0.09	NS	NS
	Older	3.51	− 1.88	− 0.84	− 0.07		
Q15: Staff cared about you	Young	4.02	− 1.89	− 0.96	− 0.31	NS	NS
	Older	3.89	− 1.93	− 0.98	− 0.33		
Q16: Staff makes you comfortable	Young	3.54	− 2.37	− 1.21	− 0.43	NS	NS
	Older	3.40	− 2.27	− 1.25	− 0.54		
Q17: Staff keep information private	Young	1.31	− 2.49			NS	NS
	Older	1.40	− 2.50				
Q20: Comfortable asking staff	Young	1.90	− 1.96			NS	NS
	Older	2.07	− 1.83				
Q24: Staff inserts needle w/o pain	Young	0.99	− 3.77	− 1.10	0.62	NS	NS
	Older	1.08	− 3.34	− 1.14	0.43		
Q25: Staff checks you closely	Young	2.46	− 2.64	− 1.21	− 0.20	NS	NS
	Older	2.19	− 2.70	− 1.26	− 0.30		
Q27: Staff manages problems	Young	2.76	− 2.60	− 1.29	− 0.58	NS	NS
	Older	2.75	− 2.66	− 1.31	− 0.66		
Q29: Staff behaves professionally	Young	2.57	− 2.43	− 1.41	− 0.52	NS	NS
	Older	2.29	− 2.59	− 1.53	− 0.64		
Q31 rev: Staff explains test result	Young	1.63	− 0.42			NS	NS
	Older	1.68	− 0.27				
Q32: Staff discuss diet enough	Young	1.05	− 2.39			NS	NS
	Older	1.40	− 1.96				
Q40: On machine within 15 min	Young	1.01	− 3.82	− 1.42	0.37	NS	NS
	Older	0.95	− 3.74	− 1.21	0.54		
Q42: Center is clean	Young	1.39	− 3.82	− 2.01	− 0.55	NS	NS
	Older	1.17	− 4.46	− 2.67	− 0.88		

Table 4 (continued)

ICH-CAHPS items		Slope	Thresholds			DIF test: χ^2 (p value)	
		a	b_1	b_2	b_3	a DIF	b DIF
Q51: Satisfied with ways problems handled	Young	1.59	− 1.94	− 0.57	0.66	NS	NS
	Older	1.27	− 2.34	− 0.36	0.69		
Providing information to patients (9 items), DTF = 0.0211							
Items candidate for DIF (Wald-1 estimates)							
Q22: Knows to take care of dialysis connection	Young	1.09	− 2.07			NS	8.6 (0.003)
	Older	1.04	− 1.62				
Q36: Staff told you how to get off machine if emergency	Young	0.50	− 2.94			NS	4.6 (0.032)
	Older	0.74	− 1.62				
Items identified as anchor (Wald-2 anchor selection estimates)							
Q33: Staff gives information on patient rights	Young	2.28	− 1.04			NS	NS
	Older	1.40	− 1.69				
Q34: Staff reviews patient rights	Young	1.90	− 0.54			NS	NS
	Older	1.29	− 0.77				
Q35: Staff told you what to do if health problem at home	Young	1.22	− 1.10			NS	NS
	Older	1.85	− 0.84				
Q44: They talked about Which treatment is right for you	Young	1.80	− 0.70			NS	NS
	Older	3.01	− 0.59				
Q46: They explain why you are not eligible for transplant	Young	1.26	0.24			NS	NS
	Older	1.18	− 0.09				
Q47: Talk about peritoneal dialysis	Young	0.85	0.35			NS	NS
	Older	1.06	0.58				
Q48: Involved in choosing treatment	Young	2.07	− 1.00			NS	NS
	Older	1.65	− 1.14				

All the observed DIFs were non-significant at 0.05 level after multiple test adjustment

Young less than 65 years old, *Older* more than 65 years old, *DTF* construct-level differential test functioning

younger patients having higher discrimination and a uniform DIF was observed in the item Q4 “how often did your kidney doctors explain things in a way that was easy to understand.” The NCDIF magnitude for the differences were 0.015 for Q4 and 0.017 for Q7 and the DTF index for the trait was 0.0111 (smaller than the cutoff). The quality of dialysis care and operations items did not exhibit any detectable DIF (with the DTF index of 0.0514) but two DIF items were observed in providing information to patients composite prior to multiple comparison adjustments. Uniform DIF was observed in Q22 “do you know how to take care of your graft, fistula, or catheter” with NCDIF magnitude of 0.005 as well as in Q36 “has any dialysis center staff ever told you how to get off the machine if there is an emergency at the center” with a DIF magnitude of 0.006. The DTF index for this construct was also estimated to be 0.0211. Figure 2a–c presented the ICC curves for selected DIF assessment and Fig. 2d presented the TCC for the providing information to patients construct and showed no practical difference between the two. Again, for all the observed DIF, the magnitudes were below the recommended cutoff for practical DIF.

Education DIF

Patients with less than high school education were also compared to those with at least a high school education. Prior to multiple comparison adjustments, none of the items in the communication and caring trait (DTF index = 0.0039) as well as the providing information to patients’ composite showed any significant DIF (DTF index = 0.0126). In the quality of dialysis care and operations trait composite, items Q11 “how often did the dialysis center staff listen carefully to you,” Q12 “how often did the dialysis center staff explain things in a way that was easy to understand,” and Q40 “how often did you get put on the dialysis machine within 15 min of your appointment or shift time” all showed uniform DIF with magnitudes 0.020, 0.023, and 0.002, respectively (see parameter estimates in Table 5). The DTF index for this trait was estimated at 0.1059 and smaller than the recommended cutoff for three candidate DIF analysis. Figure 3a–c reported the ICC plot for the DIF items and Fig. 3d the TCC of the quality of care and operations construct. Again, all the DIF

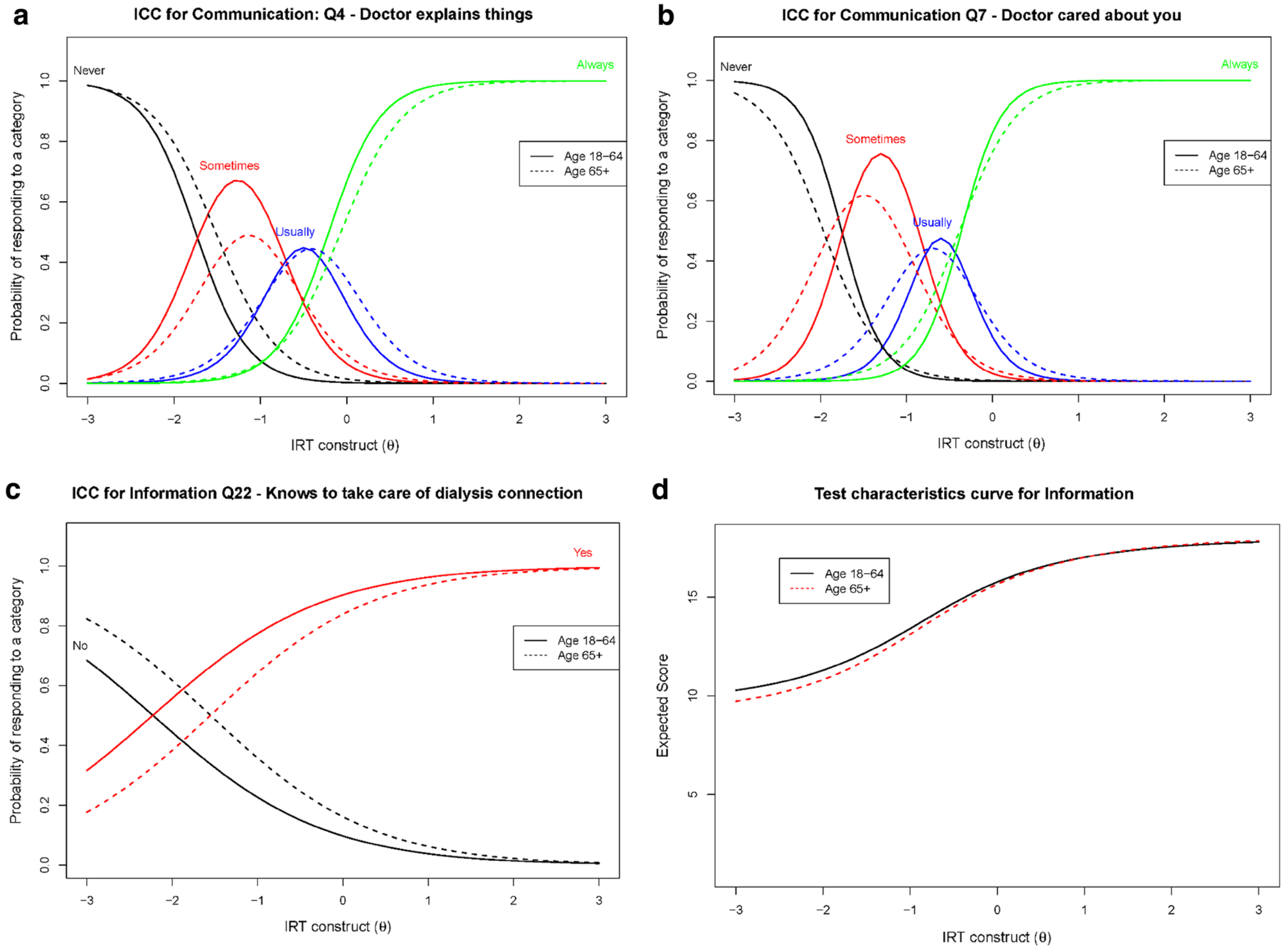

Fig. 2 Response curve for age on selected items and TCC curve for communication

magnitudes were smaller than the recommended cutoff for practical DIF.

Race DIF

Two items in each construct showed significant DIF between White and Black (see Table 6). The items Q4 "how often did your kidney doctors explain things in a way that was easy to understand" and Q6 "how often did your kidney doctors spend enough time with you" in the nephrologists' communication and caring construct showed uniform DIF. The DIF magnitudes of those items were small of size 0.005 and 0.004, respectively, and the DTF index for the trait was 0.0108 (small). For the quality of cares and operations construct, non-uniform DIFs were observed in the item Q24 "how often did dialysis center staff insert your needles with as little pain as possible" (DIF magnitude 0.0094) and in Q40 "how often did you get put on the dialysis machine within 15 min of your appointment or shift time" (DIF magnitude 0.028). The DTF index for the trait was 0.2224. Finally, for the information construct, two uniform DIF were observed, one in Q35 "has dialysis center staff ever told you what to do if you experience a health problem at home" (DIF magnitude 0.018) and Q46 "has either a doctor or dialysis center staff explained to you why you are not eligible for a kidney transplant" (DIF magnitude 0.025) and the DTF index for the trait was 0.0098. The ICC for selected items with observed DIF are reported in Fig. 4a–c, and Fig. 4d presents the TCC for the communication construct.

Sensitivity analysis

Because the patient characteristics comparisons in DIF groups revealed some confounding (Table 1), we examined matched samples for these DIF analyses to assure that any observed (or non-observed) DIF is not due to confounders. We also conducted a sensitivity analysis on the full sample to understand the impact of the confounders on any possible DIF. All the DIF analyses conducted were replicated on the full sample ignoring the confounders. In

Table 5 Estimated item parameters for graded response model: comparison of education level (less than high school vs high school or more)

ICH-CAHPS items		Slope	Thresholds			DIT test: χ^2 (p value)	
		a	b_1	b_2	b_3	a DIF	b DIF
Nephrologist communication and caring (6 items), no DIF candidate item found, DTF = 0.0039							
Q3: Doctor listens carefully	< HS	3.41	− 2.26	− 0.92	− 0.24	NS	NS
	HS +	4.28	− 2.15	− 0.99	− 0.14		
Q4: Doctor explains things	< HS	2.44	− 1.91	− 0.80	− 0.11	NS	NS
	HS +	2.94	− 1.73	− 0.82	− 0.03		
Q5: Doctor shows respect	< HS	3.05	− 2.11	− 0.97	− 0.36	NS	NS
	HS +	3.33	− 2.19	− 1.05	− 0.35		
Q6: Doctor spends enough time	< HS	2.85	− 1.69	− 0.49	0.18	NS	NS
	HS +	3.31	− 1.82	− 0.60	0.19		
Q7: Doctor cared about you	< HS	3.51	− 1.90	− 0.96	− 0.30	NS	NS
	HS +	3.06	− 2.00	− 0.92	− 0.20		
Q10: Doctor seemed informed	< HS	1.62	− 1.43			NS	NS
	HS +	2.03	− 1.33				
Quality of dialysis facility care and operations (17 items), DTF = 0.1059							
Items candidate for DIF (Wald-1 estimates)							
Q11: Staff listens carefully	< HS	3.26	− 2.21	− 0.82	− 0.11	NS	18.7 (0.000)
	HS +	3.54	− 2.32	− 1.17	− 0.13		
Q12: Staff explains in a way that is easy to understand	< HS	2.55	− 2.01	− 0.84	− 0.03	NS	16.5 (0.001)
	HS +	2.85	− 2.14	− 1.19	− 0.11		
Q40: On machine within 15 min	< HS	0.93	− 3.60	− 1.07	0.62	NS	14.8 (0.002)
	HS +	0.96	− 3.94	− 1.39	0.89		
Items identified as anchor (Wald-2 anchor selection estimates)							
Q13: Staff shows respect	< HS	3.26	− 2.28	− 1.00	− 0.22	NS	NS
	HS +	3.63	− 2.16	− 1.07	− 0.17		
Q14 Staff spent enough time	< HS	2.65	− 2.14	− 0.70	0.11	NS	NS
	HS +	3.04	− 2.17	− 0.94	0.13		
Q15: Staff cared about you	< HS	3.32	− 1.90	− 0.86	− 0.18	NS	NS
	HS +	3.57	− 2.03	− 0.91	− 0.10		
Q16: Staff makes you comfortable	< HS	2.97	− 2.54	− 1.10	− 0.33	NS	NS
	HS +	3.15	− 2.23	− 1.14	− 0.22		
Q17: Staff keep information private	< HS	1.49	− 2.48			NS	NS
	HS +	2.19	− 1.78				
Q20: Comfortable asking staff	< HS	2.32	− 1.76			NS	NS
	HS +	3.11	− 1.51				
Q24: Staff inserts needle w/o pain	< HS	0.77	− 3.97	− 1.18	0.94	NS	NS
	HS +	1.13	− 3.08	− 1.04	0.80		
Q25: Staff checks you closely	< HS	2.09	− 3.11	− 1.20	− 0.17	NS	NS
	HS +	2.47	− 2.56	− 1.12	0.07		
Q27: Staff manages problems	< HS	2.75	− 2.64	− 1.45	− 0.63	NS	NS
	HS +	2.12	− 2.94	− 1.38	− 0.41		
Q29: Staff behaves professionally	< HS	2.48	− 2.68	− 1.20	− 0.39	NS	NS
	HS +	2.86	− 2.47	− 1.45	− 0.37		
Q31 rev: Staff explains test result	< HS	1.60	− 0.19			NS	NS
	HS +	1.65	− 0.27				
Q32: Staff discuss diet enough	< HS	1.15	− 2.16			NS	NS
	HS +	1.18	− 1.89				
Q42: Center is clean	< HS	1.15	− 4.41	− 2.22	− 0.55	NS	NS
	HS +	1.18	− 3.50	− 2.00	− 0.27		

Table 5 (continued)

ICH-CAHPS items		Slope	Thresholds			DIT test: χ^2 (p value)	
		a	b_1	b_2	b_3	a DIF	b DIF
Q51: Satisfied with ways problems handled	< HS	1.48	− 1.64	− 0.50	0.75	NS	NS
	HS +	1.57	− 1.73	− 0.51	1.37		
Providing information to patients (9 items), DTF = 0.0126							
Q22: Knows to take care of dialysis connection	< HS	0.77	− 2.80			NS	NS
	HS +	1.09	− 2.35				
Q33: Staff gives information on patient rights	< HS	1.96	− 1.46			NS	NS
	HS +	1.83	− 1.73				
Q34: Staff reviews patient rights	< HS	2.14	− 0.83			NS	NS
	HS +	1.73	− 0.74				
Q35: Staff told you what to do if health problem at home	< HS	1.67	− 1.10			NS	NS
	HS +	1.57	− 1.10				
Q36: Staff told you how to get off machine if emergency	< HS	0.68	− 2.49			NS	NS
	HS +	0.59	− 2.47				
Q44: They talked about Which treatment is right for you	< HS	2.16	− 0.88			NS	NS
	HS +	2.37	− 0.78				
Q46: They explain why you are not eligible for transplant	< HS	0.60	− 0.36			NS	NS
	HS +	1.21	− 0.43				
Q47: Talk about peritoneal dialysis	< HS	0.77	0.20			NS	NS
	HS +	0.94	0.00				
Q48: Involved in choosing treatment	< HS	2.20	− 1.22			NS	NS
	HS +	1.91	− 1.37				

All the observed DIFs were non-significant at 0.05 level after multiple test adjustment

HS high school diploma, *DTF* construct-level differential test functioning

this setting (summary results reported in Online Appendix 2–6), many more items were observed with DIF. For gender comparison, three cases of uniform DIF and one case of non-uniform DIF were detected (only one, Q47, had an NCDIF larger than the recommended threshold) compared to two cases of non-uniform and one case of uniform DIF (all of them with NCDIF below the threshold) when samples were matched for confounder adjustment. For the age comparison, sic cases of non-uniform DIF and two cases of uniform DIF were observed (with two having NCDIF slightly larger than the recommended threshold) compared to three cases of uniform and one case of non-uniform when adjusting for confounders (all of them with NCDIF smaller than the threshold). For the education comparison, while only three cases of uniform DIF items were observed when we controlled for the confounders, using the full sample led to seven cases of uniform and four cases of non-uniform DIF (all but one of them with NCDF below the thresholds). For the race comparison, eight cases of uniform DIF and three cases of non-uniform DIF (three of them with NCDIF larger than the threshold) were observed compared to four cases of uniform and two cases of non-uniform when adjusting for confounders with a total of three of them with NCDIF larger than the threshold. More importantly, when taking into account the NCDIF, the confounder-free model resulted in no item with practically large DIF across the three constructs for gender, age, and education and only three DIF with magnitude larger than what is recommended as a threshold for practical DIF in race, but without such confounder adjustment, a larger number of items remained with practical DIF. This sensitivity analysis revealed that, because there are competing differences in the sample being analyzed, observed DIF in the full sample is potentially attributed to such additional differences, since a matched sample is less likely to produce biased estimates. The matching method is thus a useful tool to account for the confounding when the intent is to isolate specific sources of DIF.

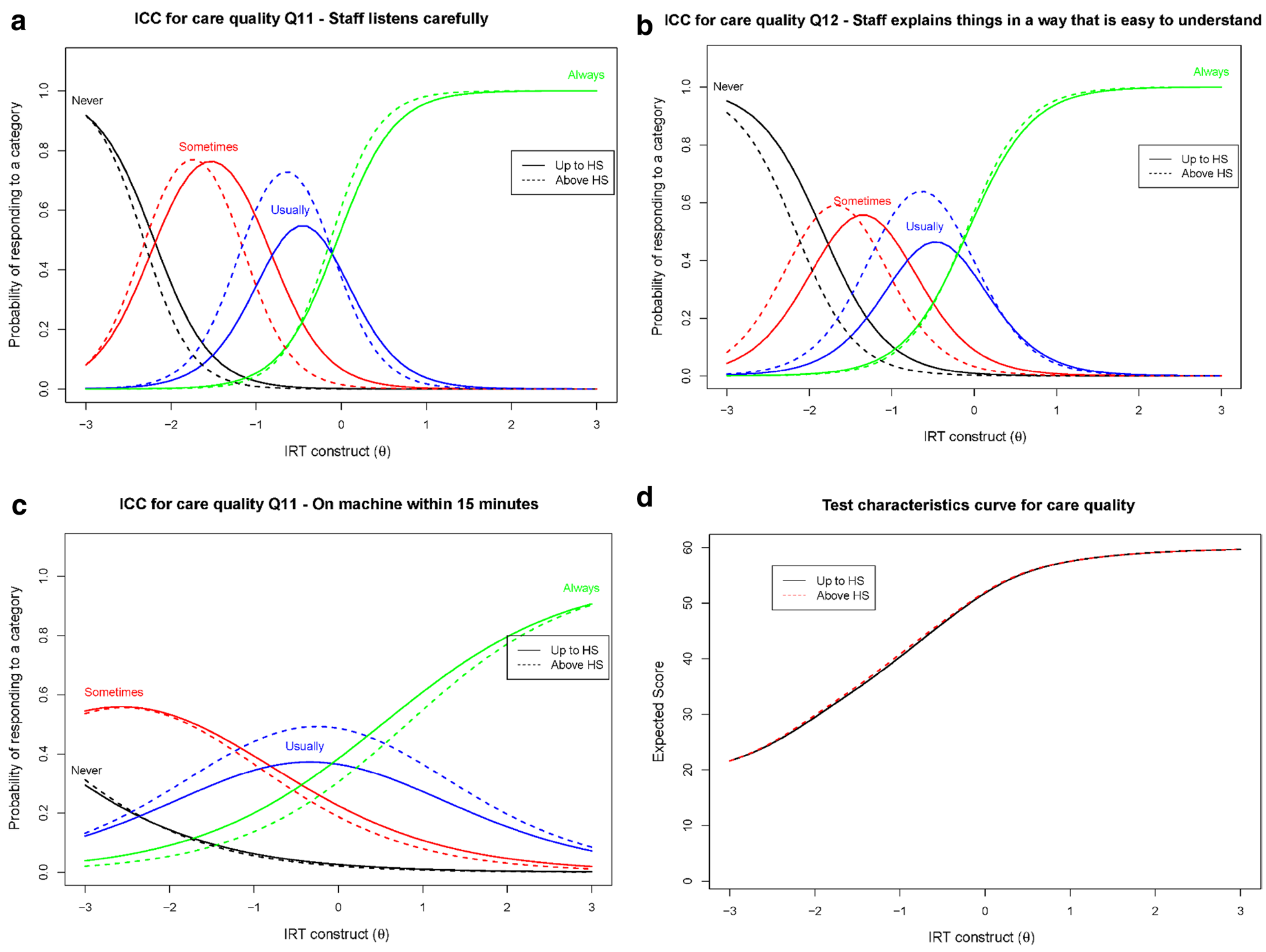

Fig. 3 Response curve for education on DIF items and TCC curve for quality of care and operations

Discussion

Demographic differences (e.g., in gender, age, education, race) may be associated with variation in care expectations among dialysis patients that can yield differences in reported experiences of care in hemodialysis patients. Differences by characteristics such as these have been reported. For example, frail, elderly patients have been found to benefit less than other [27], leading to hemodialysis rationing by age in the United States that was subsequently reversed under criticism of agism [33, 34]. Another study found significant associations of patient age and gender and scores on the Kidney Disease Quality of Life (KDQOL-36) measure [35]. A recent evaluation of DIF for the KDQOL-36 noted that despite some DIF, the magnitude was minimal [36].

Our study sheds a similar light on the possibility of survey questions used in ICH-CAHPS functioning differently for men vs women, those less than 65 vs 65 years or more, those with less than high school vs high school or more education and for White vs Black. The results suggested that the three ICH-CAHPS composites are sufficiently unidimensional for IRT analyses. In addition, no major DIF was observed across the three patient experience composites when confounders were controlled for, and the few differences observed have DIF magnitude smaller than what has been recommended as practical DIF, expect for race where some residual differences remained. The findings were different if covariates were not accounted for, indicating that subgroup characteristics impact survey item responses. Had any substantial DIF been present in age, gender, and education, the results would have suggested that ICH patients interpreted the survey questions differently. The observed lack of meaningful DIF implies that, any differences in care experience using ICH-CAHPS between gender, age, and education level patients may reflect disparities and inadequate care that have been provided on ICH patients based on different characteristics. With the DIF observed in three items (even though with small magnitude) between Whites and Blacks, caution

Table 6 Estimated item parameters for graded response model: comparison of race groups (Black vs White)

ICH-CAHPS items		Slope	Thresholds			DIF test: χ^2 (*p* value)	
		a	b_1	b_2	b_3	*a* DIF	*b* DIF
Nephrologist communication and caring (6 items), DTF = 0.0108							
Items candidate for DIF (Wald-1 estimates)							
Q4: Doctor explains things	Black	2.57	− 2.06	− 0.80	− 0.15	NS	13.9 (0.003)
	White	3.06	− 1.70	-0.88	− 0.04		
Q6: Doctor spends enough time	Black	3.04	− 1.86	− 0.45	0.16	NS	16.5 (0.001)
	White	3.44	− 1.72	− 0.64	0.22		
Items identified as anchor (Wald-2 anchor selection estimates)							
Q3: Doctor listens carefully	Black	4.82	− 2.28	− 0.81	− 0.30	NS	NS
	White	3.77	− 1.96	− 0.94	− 0.20		
Q5: Doctor shows respect	Black	2.96	− 2.12	− 1.00	− 0.42	NS	NS
	White	3.73	− 1.93	− 1.07	− 0.39		
Q7: Doctor cared about you	Black	3.47	− 2.01	− 0.84	− 0.32	NS	NS
	White	3.96	− 1.95	− 1.01	− 0.37		
Q10: Doctor seemed informed	Black	2.10	− 1.20			NS	NS
	White	2.04	− 1.44				
Quality of dialysis facility care and operations (17 items), DTF = 0.2224							
Items candidate for DIF (Wald-1 estimates)							
Q24: Staff inserts needle w/o pain	Black	0.65	− 4.64	− 1.11	1.16	12 (0.001)	34.9 (0.000)
	White	1.20	− 3.94	− 1.66	0.36		
Q40: On machine within 15 min	Black	0.74	− 4.24	− 1.24	0.76	7.4 (0.006)	19.5 (0.000)
	White	1.14	− 3.55	− 1.53	0.47		
Items identified as anchor (Wald-2 anchor selection estimates)							
Q11: Staff listens carefully	Black	3.67	− 2.50	− 1.13	− 0.28	NS	NS
	White	3.38	− 2.33	− 1.13	− 0.21		
Q12: Staff explains in a way that is easy to understand	Black	2.83	− 2.10	− 1.19	− 0.35	NS	NS
	White	2.55	− 2.25	− 1.24	− 0.15		
Q13: Staff shows respect	Black	3.70	− 2.44	− 1.17	− 0.32	NS	NS
	White	3.68	− 2.29	− 1.12	− 0.28		
Q14 Staff spent enough time	Black	2.96	− 2.25	− 0.95	− 0.05	NS	NS
	White	2.65	− 2.36	− 1.12	0.04		
Q15: Staff cared about you	Black	3.68	− 2.23	− 1.03	− 0.22	NS	NS
	White	3.88	− 2.03	− 1.14	− 0.35		
Q16: Staff makes you comfortable	Black	3.13	− 2.53	− 1.28	− 0.47	NS	NS
	White	3.19	− 2.27	− 1.32	− 0.45		
Q17: Staff keep information private	Black	1.90	− 2.13			NS	NS
	White	1.73	− 2.26				
Q20: Comfortable asking staff	Black	2.19	− 1.92			NS	NS
	White	2.48	− 1.89				
Q25: Staff checks you closely	Black	2.53	− 2.95	− 1.24	− 0.36	NS	NS
	White	2.31	− 2.67	− 1.40	− 0.15		
Q27: Staff manages problems	Black	2.10	− 3.45	− 1.47	− 0.68	NS	NS
	White	2.27	− 3.31	− 1.73	− 0.74		
Q29: Staff behaves professionally	Black	2.79	− 2.49	− 1.46	− 0.58	NS	NS
	White	2.73	− 2.99	− 1.70	− 0.60		
Q31 rev: Staff explains test result	Black	1.48	− 0.39			NS	NS
	White	1.50	− 0.49				
Q32: Staff discuss diet enough	Black	0.94	− 2.77			NS	NS
	White	1.40	− 1.66				

Table 6 (continued)

ICH-CAHPS items		Slope	Thresholds			DIF test: χ^2 (p value)	
		a	b_1	b_2	b_3	a DIF	b DIF
Q42: Center is clean	Black	1.39	− 3.54	− 2.11	− 0.51	NS	NS
	White	0.97	− 5.75	− 2.75	− 0.65		
Q51: Satisfied with ways problems handled	Black	1.12	− 2.50	− 0.50	0.68	NS	NS
	White	1.52	− 1.95	− 0.63	0.82		
Providing information to patients (9 items), DTF = 0.0098							
Items candidate for DIF (Wald-1 estimates)							
Q35: Staff told you what to do if health problem at home	Black	1.49	− 1.53			NS	13.7 (0.000)
	White	1.83	− 0.90				
Q46: They explain why you are not eligible for transplant	Black	1.01	0.05			NS	10.7 (0.001)
	White	0.99	− 0.71				
Items identified as anchor (Wald-2 anchor selection estimates)							
Q22: Knows to take care of dialysis connection	Black	1.15	− 1.94			NS	NS
	White	0.69	− 3.35				
Q33: Staff gives information on patient rights	Black	1.44	− 1.86			NS	NS
	White	1.55	− 2.02				
Q34: Staff reviews patient rights	Black	1.86	− 0.96			NS	NS
	White	1.86	− 0.93				
Q36: Staff told you how to get off machine if emergency	Black	0.63	− 2.57			NS	NS
	White	0.58	− 2.46				
Q44: They talked about Which treatment is right for you	Black	2.57	− 0.92			NS	NS
	White	3.02	− 0.83				
Q47: Talk about peritoneal dialysis	Black	0.95	0.11			NS	NS
	White	0.85	0.14				
Q48: Involved in choosing treatment	Black	1.65	− 1.42			NS	NS
	White	2.52	− 1.47				

All the observed DIFs were non-significant at 0.05 level after multiple test adjustment

HS high school diploma, *DTF* construct-level differential test functioning

should be used when making inference on race disparity if those items are included in the comparisons.

As correlation is not causation and in DIF analysis, participants are not randomized into different groups (e.g., gender is not a random assignment), the traditional methods of conducting DIF analysis have the potential to show items that are functioning differentially, but such inference can be confounded by other factors observed or non-observed in the data. With the propensity score matching methods we used in this study, we were able to render the comparison between groups such as male vs female similar to a setting where the group assignment can be randomized. This method provides the advantage of eliminating potential of the observed or non-observed DIF being attributable to other confounders. As an add-on methodology to any method that can be subject to selection and confounder bias (including DIF), this approach warrants further consideration as an addition to existing psychometric toolbox for examining DIF when the purpose is to isolate particular sources of DIF.

This study has limitations. The 46% response rate is modest and may limit inferences, but higher response rates to such surveys do not generally yield significantly different estimates [37, 38]. Also, with participants being sampled at the facilities, there is a potential that facility clustering not been accounted for in our method, can lead to potential bias or a too small standard errors in the estimators. In addition, the relatively small number of participants in some demographic subgroups and the matched samples, allowing, for example, only 58% of the sample to be matched and used for analysis in the education level DIF, can tamper the generalizability of the results in some cases. The generalizability can also be weakened for the "providing information to patients" construct where the CFA fit was not optimal. Relatedly, structured missing data in some of the survey items can reduce the analytic sample size, which can bias the DIF estimates and inferences reported. Nevertheless, the sample size is large enough for meaningful inferences and the results of this study are relevant for policy makers.

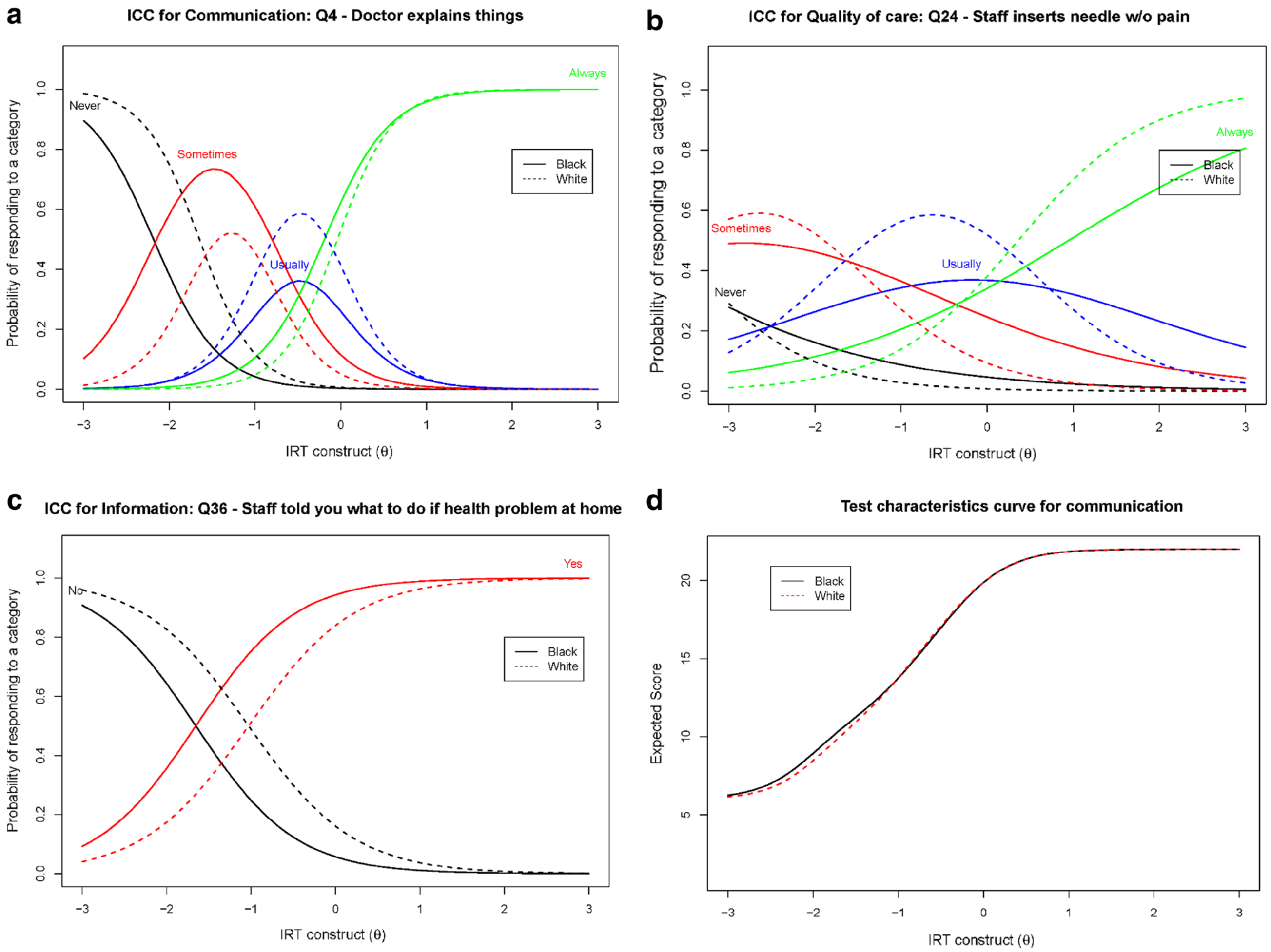

Fig. 4 Response curve for race (black vs white) on selected items and TCC curve for communication

Conclusion

The success and quality of treatments for renal failure via hemodialysis has been historically assessed using measures considered important by physicians, and although instruments measuring the patient's perspective have been available for decades for health care in general and more recently for kidney disease, their incorporation into routine clinical hemodialysis practice has only been recent. The fact that the Centers for Medicare and Medicaid Services (CMS) now has ICH-CAHPS playing a significant role in dialysis service evaluation in the United States by, for example, making it part of CMS's ratings of dialysis center performance [39] is a signal of the importance of these patient experience reports. This study supports continued use of ICH-CAHPS to evaluate the care delivered to dialysis facilities [40]. Decision makers using ICH-CAHPS can make substantive comparisons of hemodialysis care received between men and women, younger and older, between more and less educated patients, and between black and white patients because we found no impact of measurement non-invariance between these groups after controlling for other potentially confounding differences between the groups. The ability to make such comparison is also critical for patient choice of providers and facilities.

Funding This research was supported through cooperative agreements from the Agency for Healthcare Research and Quality: U18 HS016980 and U18 HS016978.

Compliance with ethical standards

Conflict of interest None of the authors have conflicts of interest or any financial relationships relevant to this article to disclose.

Informed consent Informed consent was obtained from all individual participants included in the study.

Ethical approval All data collection were approved by the RAND institutional review boards (FWA00003425, effective until June 22, 2023). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

References

1. (USRDS), U. S. R. D. S. (2017). 2017 USRDS annual data report: Epidemiology of kidney disease in the United States. http://www.usrds.org. Accessed October 18, 2018.
2. Saran, R., Robinson, B., Abbott, K. C., Agodoa, L. Y. C., Bragg-Gresham, J., Balkrishnan, R., et al. (2018). US Renal Data System 2017 annual data report epidemiology of kidney disease in the United States. *American Journal of Kidney Diseases, 71*(3), S1–S676. https://doi.org/10.1053/j.ajkd.2018.01.002.
3. Couchoud, C., Kooman, J., Finne, P., Leivestad, T., Stojceva-Taneva, O., Ponikvar, J. B., et al. (2009). From registry data collection to international comparisons: Examples of haemodialysis duration and frequency. *Nephrology, Dialysis, Transplantation, 24*(1), 217–224. https://doi.org/10.1093/ndt/gfn442.
4. Kjellstrand, C. M. (1988). Age, sex, and race inequality in renal-transplantation. *Archives of Internal Medicine, 148*(6), 1305–1309.
5. Rosansky, S. J. (2012). The sad truth about early initiation of dialysis in elderly patients. *Jama-Journal of the American Medical Association, 307*(18), 1919–1920. https://doi.org/10.1001/jama.2012.3522.
6. Thorsteinsdottir, B., Swetz, K. M., Feely, M. A., Mueller, P. S., & Williams, A. W. (2012). Are there alternatives to hemodialysis for the elderly patient with end-stage renal failure? *Mayo Clinic Proceedings, 87*(6), 514–516. https://doi.org/10.1016/j.mayocp.2012.02.016.
7. Adler, N. E., Boyce, W. T., Chesney, M. A., Folkman, S., & Syme, S. L. (1993). Socioeconomic inequalities in health. No easy solution. *Journal of the American Medical Association, 269,* 3140–3145.
8. Nelson, E. C., Hays, R. D., Arnold, S., Kwoh, K., & Sherbourne, C. (1989). *Age and functional health status*. Santa Monica, CA: The RAND Corporation.
9. Khattak, M., Sandhu, G. S., Desilva, R., & Goldfarb-Rumyantzev, A. S. (2012). Association of education level with dialysis outcome. *Hemodialysis International, 16*(1), 82–88. https://doi.org/10.1111/j.1542-4758.2011.00615.x.
10. Peipert, J. D., Bentler, P., Klicko, K., & Hays, R. D. (2018). Negligible impact of differential item functioning between Black and White dialysis patients on the Kidney Disease Quality of Life 36-item short form survey (KDQOL((TM))-36). *Quality of Life Research, 27*(10), 2699–2707. https://doi.org/10.1007/s11136-018-1879-3.
11. Salisbury, C. (2009). Using patient experience within pay for performance programmes. *Bmj-British Medical Journal*. https://doi.org/10.1136/bmj.b4224.
12. Al-Abri, R., & Al-Balushi, A. (2014). Patient satisfaction survey as a tool towards quality improvement. *Oman Medical Journal, 29*(1), 3–7. https://doi.org/10.5001/omj.2014.02.
13. Gleeson, H., Calderon, A., Swami, V., Deighton, J., Wolpert, M., & Edbrooke-Childs, J. (2016). Systematic review of approaches to using patient experience data for quality improvement in healthcare settings. *BMJ Open*. https://doi.org/10.1136/bmjopen-2016-011907.
14. Weidmer, B. A., Cleary, P. D., Keller, S., Evensen, C., Hurtado, M. P., Kosiak, B., et al. (2014). Development and evaluation of the CAHPS (Consumer Assessment of Healthcare Providers and Systems) survey for in-center hemodialysis patients. *American Journal of Kidney Diseases, 64*(5), 753–760. https://doi.org/10.1053/j.ajkd.2014.04.021.
15. Setodji, C. M., Reise, S. P., Morales, L. S., Fongwa, M. N., & Hays, R. D. (2011). Differential item functioning by survey language among older hispanics enrolled in medicare managed care a new method for anchor item selection. *Medical Care, 49*(5), 461–468. https://doi.org/10.1097/MLR.0b013e318207edb5.
16. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41.
17. Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, J., et al. (2013). Optimal caliper width for propensity score matching of three treatment groups: A Monte Carlo study. *PLoS ONE, 8*(12), e81045. https://doi.org/10.1371/journal.pone.0081045.
18. Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330.
19. Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Multivariate applications series. Handbook of item response theory modeling: Applications to typical performance assessment*. Taylor & Francis Group: New York, NY.
20. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling-A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.
21. Samejima, F. (1996). Evaluation of mathematical models for ordered polychrotomous responses. *Behaviormetrika, 23,* 17–35. https://doi.org/10.2333/bhmk.23.17.
22. Wang, M., & Woods, C. M. (2017). Anchor selection using the wald test anchor-all-test-all procedure. *Applied Psychological Measurement, 41*(1), 17–29. https://doi.org/10.1177/0146621616668014.
23. Woods, C. M., Cai, L., & Wang, M. A. (2013). The langer-improved wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532–547. https://doi.org/10.1177/0013164412464875.
24. Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458–470. https://doi.org/10.1177/0013164412467033.
25. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological, 57*(1), 289–300.
26. Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*(11–12), 1651–1683. https://doi.org/10.1002/(SICI)1097-0258(20000615/30)19:11/12%3c1651:AID-SIM453%3e3.0.CO;2-H.
27. Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353–368.
28. Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*(4), 309–326. https://doi.org/10.1177/01466219922031437.
29. Raju, N. S. (1999). *DFITP5: A Fortran program for calculating dichotomous DIF/DTF*. Chicago: Illinois Institute of Technology.
30. Muthén, L., & Muthén, B. (2001). *Mplus User's Guide*. CA: Los Angeles.

31. Cai, L. (2012). *Vector Psychometric Group*. WA: Seattle.
32. Wood, R., Paoli, C. J., Hays, R. D., Taylor-Stokes, G., Piercy, J., & Gitlin, M. (2014). Evaluation of the consumer assessment of healthcare providers and systems in-center hemodialysis survey. *Clinical Journal of the American Society of Nephrology, 9*(6), 1099–1108. https://doi.org/10.2215/Cjn.10121013.
33. Kjellstrand, C. (1997). All elderly patients should be offered dialysis. *Geriatric Nephrology and Urology, 6*, 129–136. https://doi.org/10.1007/BF00249628.
34. Avorn, J. (1984). Benefit and cost analysis in geriatric care. Turning age discrimination into health policy. *New England Journal of Medicine, 310*(20), 1294–1301. https://doi.org/10.1056/nejm198405173102005.
35. Veerappan, I., Arvind, R. M., & Ilayabharthi, V. (2012). Predictors of quality of life of hemodialysis patients in India. *Indian Journal of Nephrology*, *22*(1), 18–25. https://doi.org/10.4103/0971-4065.91185.
36. Peipert, J. D., Bentler, P. M., Klicko, K., & Hays, R. D. (2018). Psychometric properties of the kidney disease quality of life 36-item short-form survey (KDQOL-36) in the United States. *American Journal of Kidney Diseases, 71*(4), 461–468. https://doi.org/10.1053/j.ajkd.2017.07.020.
37. Groves, R., Floyd, J., & Fowler, J. M. (2013). *Survey methodology*. Hoboken: Wiley.
38. Halbesleben, J. R. B., & Whitman, M. V. (2013). Evaluating survey quality in health services research: A decision framework for assessing nonresponse bias. *Health Services Research, 48*(3), 913–930. https://doi.org/10.1111/1475-6773.12002.
39. Richardson, M. M., & Grobert, M. E. (2014). ICH-CAHPS: What signal on the chadburn? *American Journal of Kidney Diseases, 64*(5), 670–672. https://doi.org/10.1053/j.ajkd.2014.09.002.
40. Peipert, J. D., & Hays, R. D. (2017). Methodological considerations in using patient reported measures in dialysis clinics. *Journal of Patient-Reported Outcomes, 1*(1), 11. https://doi.org/10.1186/s41687-017-0010-9.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.