Assessment of Psychometric Properties of an Oral Health Care Measure of Cultural Competence Among Dental Students Using Rasch Partial Credit Model

Yu Su, Linda S. Behar-Horenstein

Abstract: Reliability, validity, and feasibility of the only validated oral health care measure of cultural competence, the Knowledge, Efficacy, and Practices Instrument (KEPI), have been confirmed. However, the instrument's psychometric properties including item and person reliability, category response functioning, and scale targeting, as well as differential scale functioning for subgroups, have not yet been examined. The aim of this study was to test the psychometric properties of KEPI among dental students using Rasch Partial Credit Model to determine if this model provided broader valid information that cannot be demonstrated according to Classical Test Theory. A total 1,290 dental students in the first or final semester at four U.S. dental schools were invited to participate in the study in 2016. Of those, 1,231 individuals completed the survey, for a 95.4% response rate. The participants were 613 males and 618 females and 889 non-underrepresented minority (URM) and 342 URM students. The Rasch Partial Credit Model assessed the psychometric properties of KEPI's 20 items/three subscales. Differential scale functioning was found in the Culture-Centered Practice and Efficacy of Assessment subscales. Four items were endorsed differentially by gender; four items were endorsed differentially by URM/non-URM students. This study examined the psychometric properties of the KEPI using Rasch analysis to assess differential item functioning by dental student gender and race. The results provided valid evidence for the high internal reliability, measurement properties, and unidimensionality for the KEPI domains, ideal targeting, and well response category functioning, showing that the KEPI is a reliable instrument for measuring the subscales Knowledge of Diversity, Culture-Centered Practice skills, and Efficacy of Assessment for health care providers.

Yu Su, PhD, is a graduate of the School of Human Development and Organizational Studies in Education, University of Florida; Linda S. Behar-Horenstein, PhD, is Distinguished Teaching Scholar and Professor, Colleges of Dentistry and Education, Director of CTSI Educational Development & Evaluation, and Co-Director of HRSA Faculty Development in Dentistry, College of Dentistry, University of Florida. Direct correspondence to Dr. Linda S. Behar-Horenstein, Clinical Translational Science Institute, University of Florida, P.O. Box 100208, Room CG-72-6, Gainesville, FL 32610-0208; Lsbhoren@ufl.edu.

Keywords: dental education, cultural competence, cultural diversity, differential scale functioning, educational measurement, item response theory, KEPI, Rasch Partial Credit Model

Submitted for publication 12/20/17; accepted 2/27/18
doi: 10.21815/JDE.018.107

As numbers of ethnic and racial minorities in the U.S. continue to increase, the demand for culturally responsive health care has necessitated an examination and revision of what and how dental students are taught. To assess the impact of curricular changes, Like recommended measuring dental students' knowledge, efficacy, and practice beliefs at the beginning, interim points, and near the conclusion of their training.[1] Several psychological instruments measuring cultural competence among health professions students and providers have been used in applied research across health care disciplines, such as the Cultural Competence Health Practitioner Assessment (CCHPA),[2] the Tucker-Culturally Sensitive Health Care Inventory (TCSHCI),[3] the Defining Issues Test2 (DIT2),[4] and the Inventory of Assessing the Process of Cultural Competence among Health Professionals (IAPCC-R).[5] However, no studies have supported the reliability and validity of these instruments.

In 2012, Behar-Horenstein et al. conducted an initial validation study of the Knowledge, Efficacy, and Practices Instrument for Oral Health Providers (KEPI-OHP).[6] This scale consists of 20 items measuring three latent domains: Knowledge of Diversity, Culture-Centered Practice, and Efficacy of Assessment. The internal reliability of KEPI-OHP was assessed by Cronbach's alpha, and its factor structure was explored using exploratory and confirmatory factor analysis. The results were that the correlated three-factor model demonstrated acceptable internal consistency reliability and showed a good model fit

with the data. Later, the convergent and divergent validity of KEPI-OHP was assessed by Pearson correlation,[7] and the results ensured adequate external validity. Recently, a short form of KEPI with 17 items was developed by Garvan et al.[8] The revised form omitted items with weak factor loadings and confirmed construct validity across the three subscales.

Although both the original and revised versions of KEPI showed strong reliability and validity evidence to support their use for measuring cultural competence among oral health care providers, analysis of psychometric properties was limited to Classical Test Theory (CTT). In general, CTT is a traditional psychometric technique in which the participant performance and the properties of items are estimated in terms of summated scores across multiple items. Despite CTT's wide use over past decades, this technique is unable to examine specific item properties and a person's latent traits and characteristics. In addition, CTT cannot satisfactorily identify participant use of response categories in line with the intention of test developers.

Unlike CTT, item response theory (IRT) allows examination of item property difficulty, item discrimination, and individual ability based on the amount of information test takers provide about the latent trait. The IRT models for ordered polytomous items (e.g., items rated on a Likert or other scale) generally include the partial credit model (PCM), rating scale model (RSM), and graded response model (GRM). These models assume a latent construct is underlying the observed data and that both the respondents and items can be arrayed along a continuum.[9] By examining the relationship between an individual's performance on each item and the respondents' levels of performance on an overall latent trait that the item was designed to measure, the models can provide information about the difficulty of endorsing an item, the ability to discriminate respondents with different levels of latent traits, and the category threshold unique to each item. In recent years, IRT models have been increasingly applied in assessing the knowledge, culture competence, and practices of health care professionals. For example, Haywood et al. conducted a comprehensive psychometric evaluation of Cultural Competence Health Practitioner Assessment (CCHPA-129) using CTT and IRT model.[10] In addition, differential item functioning analysis (DIF) was performed to assess if scores of CCHPA-129 had equivalent meaning across race, ethnicity, gender, and profession. Court et al. demonstrated the value of using Rasch-derived RSM to assess the psychometric properties of Spielberger State Anxiety Scale (SSAS) to measure patient anxiety.[11]

Although the reliability, validity, and feasibility of KEPI have been confirmed among dental faculty members and students, its psychometric properties including item and person reliability, category response functioning, and scale targeting, as well as differential scale functioning (DSF) for subgroups, have not yet been examined. The aim of this study was to test the psychometric properties of KEPI among dental students using PCM to determine if this model provided broader valid information that cannot be demonstrated according to CTT. In addition, we were interested in conducting DSF analysis to test the measurement equivalent across gender and ethnicity subgroups.

Methods

This study was approved by the University of Florida's Institutional Review Board (#2013-0989). A total 1,290 dental students in the first or final semester at four U.S. dental schools were invited to participate in the study in 2016.

The KEPI, a 20-item instrument, measures self-reported levels of cultural competence among predoctoral dental students and other oral health care professionals. Items on its three latent constructs related to cultural competence (Knowledge of Diversity, Culture-Centered Practice, and Efficacy of Assessment) are rated using a four-point scale, with response options 1=very limited to 4=very aware for items 1, 2, 9-17, and 19-27 and response options 1=strongly disagree to 4=strongly agree for items 3-8 and 18. Scores on Knowledge of Diversity provide a measure that reflects an individual's understanding of sociocultural and linguistically diverse groups. Efficacy of Assessment scores measure the extent to which individuals are capable of responding satisfactorily to culturally diverse patient oral health needs. Culture-Centered Practice scores measure an individual's awareness of sociocultural and linguistically diverse patients' needs for oral health care.

The internal consistency of each domain of KEPI was assessed by Cronbach's alpha. Higher test reliability indicates that the test measures what it intends to measure in a consistent manner. The value of that is equal to or larger than 0.7 suggests acceptable reliability. A confirmatory factor analysis was

conducted in Mplus Version 7[12] to check the assumptions of IRT models including unidimensionality and local independence. Weighted least squares estimation with adjusted means and variances (WLSMV) was applied as it is appropriate for categorical data and does not assume a linear relationship between the items and latent construct.[13] Unidimensionality indicates that the items are measuring a single construct. In this study, if the unidimensionality assumption was met, a single factor model should well fit the items for each subscale of KEPI. The goodness of model fit was assessed in terms of the root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis Index (TLI). Following the previous rule of thumb,[14,15] a RMSEA value less than 0.05 indicates acceptable degree of model fit, and CFI and TLI values that are larger than 0.95 show a good model fit.

The assumption of local independence indicates the latent trait is the only reason that caused the correlation between all pairs of the items. Local independence was assessed with a bivariate model fit information that checks for high correlations between pairs of items after conditioning on the latent construct.[16]

The Rasch Partial Credit Model (PCM) was applied to assess the psychometric properties of the 20-item and three-subscale KEPI structure. The reason we used Rasch PCM rather than Rating Scale Model was because examiners may interpret scales differently in terms of the items. The eRm package[17] in the R software was used to implement the analysis including item/person fit, scale precision, response category functioning, scale targeting, and differential item functioning across subpopulations of interest. The formula used for constructing measures through item response to the PCM is shown as follows:

$$P_{ix}(\theta_j) = \frac{exp\sum_{k=0}^{x}(\theta_j - \delta_{ik})}{\sum_{r=0}^{m_j}[exp\sum_{k=0}^{r}(\theta - \delta_{ik})]}$$

Where item i is scored $x = 0, \ldots, m_i$ for an item with $K_i = m_i + 1$ response categories, $P_{ix}(\theta_j)$ indicates the probability that an examinee j answers item i at category x correctly, conditioning on latent level θ_j; r=0,…,k,…,m and k refer to a specific response category that is modeled; and δ_{ik} ($k = 1, \ldots, m_i$) refers to the item step difficulty that measures the difficulty of item i being advanced from category $k - 1$ to k.

The scale precision was determined by item separation reliability (IR), indicating how well items can be discriminated from one another in terms of item difficulty. In addition, person separation reliability (PR) was estimated to identify the extent to which participants can be discriminated based on their estimated person ability.[18] When the values of PR and IR are larger than 0.8, it indicates an acceptable level of discrimination. If this condition is satisfied, then the Rasch analysis continues with an examination across each latent construct to determine the item and person fit in each domain. Survey items and participants that did not adequately fit the latent construct were identified using the infit mean-square statistics (Infit MSQ) and outfit mean-square statistics (Outfit MSQ), which are preferred indices of fit of the Rasch model.[19] Infit/outfit values close to 1 are considered ideal fit, while Infit/outfit values that are either below 0.5 or above 1.5 are considered misfitting. Poor fitting items indicate the need to remove or revise the item on the survey, while a large number of flagged persons indicates that the instrument is not working well for the sample of dental students who took the survey or that the participants were not providing thoughtful answers.

Response category functioning was examined by step measures within each category, advancing average measures as well as category probability curves.[20,21] Step measure parameters describe the boundaries between each of two adjacent categories for score x. They are expected to monotonically advance with increasing participant ability. Category probability curves were presented as a visual tool to help assess if the category threshold impeded the expected order. By inspecting person-item maps, we examined the scale targeting and compared the location of items and the distribution of person ability on the same logit scale. Ideally, targeting occurs when a set of items within a construct are located along the full range of KEPI score in the population and are well centered in terms of the person measure distribution.

In the final step, the differential step functioning (DSF) analysis in eRm package was applied to assess if each step of polytomous items was measuring the latent constructs differently for dental students across two subpopulations (male vs. female; URM vs. non-URM). Pervasive DSF indicates that all steps display a substantial DSF effect, while non-pervasive DSF means that only one or a few steps display a substantial DSF effect. A Lord's Wald-test[22] was applied in each of the three steps of KEPI items to locate the source of the biasing factor when polytomous DIF is present. The Type I error rates for DIF detection

using Lord's Wald tended to be within the expected nominal alpha value 0.05. A p-value of certain steps of polytomous items of <0.05 indicated that the item was flagged for DIF across subgroups. This study used an existing database for analysis.

Results

Of the students invited to participate in the study, 1,250 students agreed to participate, but 19 did not complete the survey and were excluded from the data analysis. The final number of participants was 1,231, for a response rate of 95.4%. The participants were 613 males, 618 females, 889 non-URM students, and 342 URM students. Mean scores, standard deviation, and internal reliability for the three subscales of KEPI are shown in Table 1. The overall reliability was 0.81, and the Cronbach's alpha for all three subscales was greater than 0.7, indicating that the instrument has adequate internal consistency.

Psychometric Properties

The CFA model fit generally supported the unidimensional assumption for the Knowledge of Diversity domain (RMSEA=0.01, CFI/TLI=0.987/0.978) and Culture-Centered Practice domain (RMSEA=0.019, CFI/TLI=0.970/0.939); however, for the Efficacy of Assessment domain (RMSEA=0.157, CFI/TLI=0.987/0.983), the value of RMSEA >0.1 indicated that the unidimensionality of efficacy of assessment may be tentative.[14] The assumption of local independence was assessed with a non-parametric test that checks for high correlations between pairs of items conditioning on the latent construct.[16] The test indicated that most pairs of items showed independence at the p<0.01 level, except items 1 and 5 and items 11 and 12. Given that the independent assumptions were not violated with respect to quite a few pairs of items, we concluded the local independence assumption was confirmed to an appropriate degree for interpreting results of the PCM model.

Table 1. Mean, standard deviation (SD), and internal reliability by domain

Domain/Item	Mean	SD	Reliability
Knowledge of diversity	2.26	0.59	0.85
1. How would you rate yourself in terms of understanding how your ethnicity/culture has influenced the way you think and act?			
10. At the present time, how would you rate your own understanding of the following terms: Culture?			
11. Ethnicity?			
12. Racism?			
14. Prejudice?			
15. Culturally diverse patients?			
Culture-centered practice	1.33	0.53	0.80
9. How would you rate your understanding of "patient management" for treating patients from ethnically/culturally diverse groups?			
13. At the present time, how would you rate your own understanding of the following terms: culturally diverse oral health care practices?			
16. Pluralism?			
17. Cultural encapsulation?			
Efficacy of assessment	1.80	0.57	0.91
18. In dentistry, patients from different ethnic/cultural groups should be given the same treatment.			
19. How would you rate your ability to: Accurately identify your own culturally biased assumptions as they relate to your professional practice?			
20. Effectively secure information and resources to better serve patients of different ethnic/cultural groups?			
21. Accurately assess the oral health care needs of women?			
22. Accurately assess the oral health care needs of men?			
23. Accurately assess the oral health care needs of older adults?			
24. Accurately assess the oral health care needs of gay, lesbian, bisexual, and transgendered individuals?			
25. Accurately assess the oral health care needs of patients with disabilities?			
26. Accurately assess the oral health care needs of persons who come from low socioeconomic environments?			
27. Identify your own strengths and weakness of oral health care treatment planning for persons from different ethnically/culturally diverse groups?			
Total	1.80	0.51	0.81

Next, the PCM-derived reliability for person and item of each domain was determined (Table 2). Item separation reliability for the three subscales ranged between 0.85 to 0.89, suggesting acceptable discrimination among items. Person separation reliability by domains ranged between 0.80 and 0.91, indicating the participants were well discriminated from each other on their estimated ability.

Item and Person Statistics

Table 3 shows the item level results from the Rasch analysis. Specifically, item difficulty estimates, step measure, item statistics, and category frequency for each item are presented. All response options were intended to form an ordered categorical scale with 0 as low and 3 as high. The range of difficulty in KEPI was from -0.06 to 2.05. Item 9 ("How would you rate your understanding of patient management for treating patients from ethnically/culturally diverse groups?") was the easiest one with 46.3% of the students reporting "Aware." Item 25 ("How well would you rate your ability to accurately assess the oral health care needs of patients with disabilities?") was the most difficult one, with 39.4% students reporting "Aware." It appears from the category frequency of Table 3 that the participants' responses did not fully cover the range of the four-point scale. For all three subscales of KEPI, the Category 3 "Aware" or "Agree" was selected most frequently, followed by Category 2 "Limited" or "Disagree," Category 4 "Strongly Agree" or "Very Aware," and Category 1 "Very Limited" or "Strongly Disagree." The responses were more well distributed across the four-

Table 2. Person/item reliability estimates by domain

Person/Item	Knowledge of Diversity	Culture-Centered Practice	Efficacy of Assessment
Person separation	0.81	0.80	0.91
Item separation	0.85	0.88	0.89

Table 3. Item difficulty, step measure, category frequency, and outfit/infit statistics for each item by domain

	Item Difficulty	Step Measure			Category Frequency (%)				Item Statistics	
Scale/Item	Location	T1	T2	T3	1	2	3	4	Outfit	Infit
Scale 1 KOD										
Item 12	0.64	-2.52	0.12	4.57	0.24	3.82	54.03	41.90	0.60	0.65
Item 11	0.97	-3.79	1.16	5.55	0.16	10.25	60.37	29.21	0.59	0.63
Item 15	1.28	-2.52	1.01	5.36	0.33	9.11	58.99	31.57	0.74	0.79
Item 10	1.34	-2.00	0.61	5.42	0.41	6.83	61.59	31.16	0.59	0.65
Item 14	1.41	-1.76	0.72	5.27	0.49	7.24	59.48	32.79	0.69	0.73
Item 1	1.71	-0.54	0.77	4.91	1.14	6.75	54.68	37.43	1.55	1.41
Scale 2 CCP										
Item 9	-0.06	-3.45	0.08	3.19	3.90	37.89	46.26	11.95	0.92	0.92
Item 13	0.27	-2.80	0.43	3.19	7.15	40.65	40.57	11.63	0.73	0.74
Item 16	1.59	-0.96	1.88	3.85	24.55	47.24	22.03	6.18	0.75	0.74
Item 17	1.91	-0.55	1.99	4.28	29.67	44.88	20.49	4.96	0.70	0.68
Scale 3 EOA										
Item 18	0.30	-2.75	0.34	3.30	2.43	22.05	48.96	26.56	2.77	2.37
Item 19	0.38	-3.76	0.11	4.80	1.04	21.70	64.67	12.59	1.29	1.30
Item 20	0.77	-2.77	1.01	4.07	2.69	31.34	48.44	17.53	1.24	1.24
Item 21	0.87	-1.53	0.71	3.43	6.08	23.87	45.57	24.48	0.50	0.51
Item 22	0.92	-1.59	0.65	3.70	5.82	23.78	48.78	21.61	0.46	0.46
Item 26	0.98	-1.59	0.57	3.98	5.73	23.18	52.17	18.92	0.57	0.57
Item 27	1.00	-1.92	0.65	4.28	4.69	25.26	53.91	16.15	0.77	0.77
Item 23	1.30	-1.55	1.07	4.39	6.42	29.60	49.13	14.84	0.50	0.50
Item 24	1.47	-1.00	1.27	4.14	9.29	29.95	44.27	16.49	0.57	0.57
Item 25	2.05	-0.77	1.94	4.99	11.72	38.80	39.41	10.07	0.72	0.71

Note: Scale 1 refers to Knowledge of Diversity (KOD), scale 2 refers to Culture-Centered Practice (CCP), and scale 3 refers to Efficacy of Assessment (EOA).

point rating scale in the Efficacy of Assessment and Culture-Centered Practice domains. An extremely low number of students choose Category 1 "Very Limited" in the domain Knowledge of Diversity.

Inspection of the item statistics (Outfit/Infit MSQ) for the 27 items indicated that the Culture-Centered Practice domain displayed reasonable fit to the PCM model with no misfit for any items. That domain performed well in fitting the model with only one item (Item 1, "How would you rate yourself in terms of understanding how your ethnicity/culture has influenced the way you think and act?") showing Outfit/Infit values larger than 1.5. However, two items (Item 18, "In dentistry, patients from different ethnic/cultural groups should be given the same treatment," and Item 22, "How would you rate your ability to accurately assess the oral health care needs of men?") in the Efficacy of Assessment domain were identified beyond the cut-off. These items were flagged as poorly fitting by targeted subscale.

A high percentage of participants (36%) in the Knowledge of Diversity domain were identified as poorly fitting persons. This finding indicated that the subscale may not adequately measure respondents' knowledge of culture diversity. Both the Culture-Centered Practice and Efficacy of Assessment domains adequately measured participants' latent ability with less than ten persons being marked as poorly fitting.

Response Category Functioning and Scale Targeting

The step measures for each item increased monotonically across rating scale categories (from threshold 1 to threshold 3) for all three subscales and demonstrated well the response category functioning across the domains. The examples of category probability curves are shown in Figure 1, visualizing well order threshold across rating scales.

Person-item maps were used to measure the extent to which survey items targeted the participants' ability of the latent construct. In Figure 2, the bars on the top part of the map represent the distribution of students' ability on the domain (lower number indicates less endorsement to the survey items). The dots in the lower part of the map indicate item difficulty. The targeting of the Knowledge of Diversity domain was poor. The items were relatively homogenous in difficulty of endorsement and are generally distributed in the left side of respondent' ability distribution, indicating that the items were, on average, too easy for the participants. For the Culture-Centered Practice and Efficacy of Assessment domains, the set of items in each construct were nearly located along the full range of KEPI score in the population and were well centered in terms of the person measure distribution, indicating that the survey items ideally targeted the participants' ability of the domains.

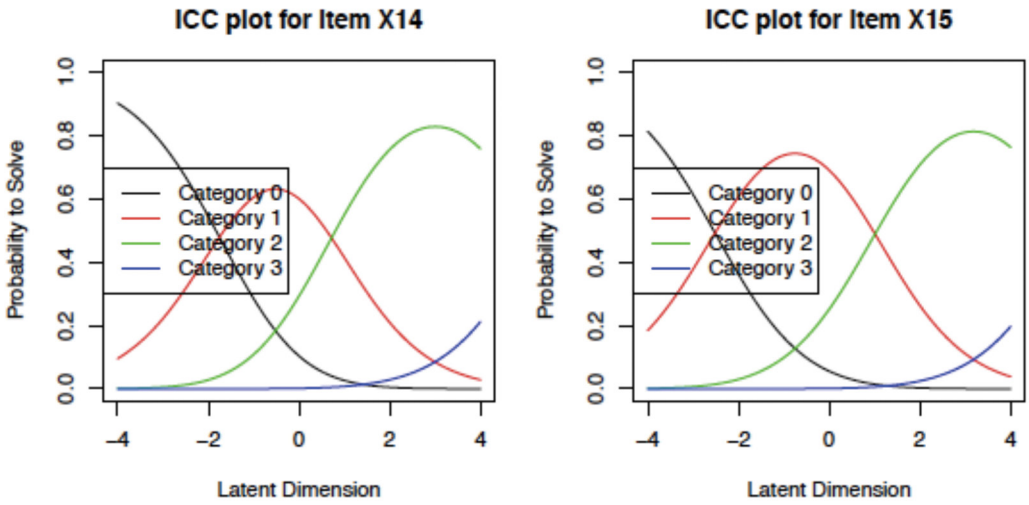

Figure 1. Category probability curves of items 14 and 15, showing "well-ordered threshold," with the step measures for each item increased monotonically across rating scale

Note: For items 14 and 15, 1=very limited was designated Category 0 (reference category); 2=limited was designated Category 1; 3=aware was designated Category 2; and 4=very aware was designated Category 3.

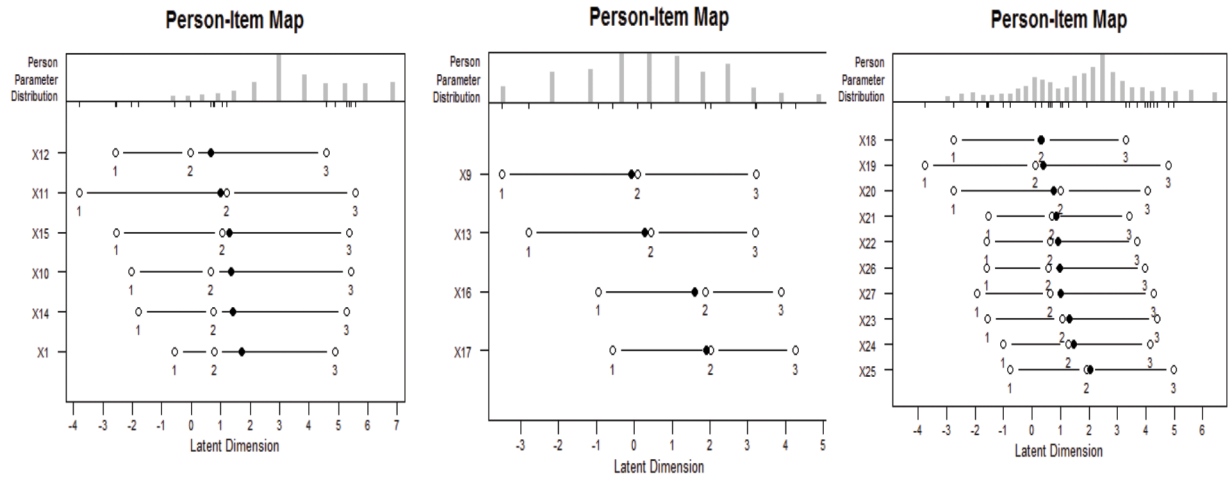

Figure 2. Person-item maps by domain: Knowledge of Diversity (left panel), Culture-Centered Practice (middle panel), Efficacy of Assessment (right panel)

Differential Step Functioning

In the last step, a PCM derived differential step functioning analysis was used to identify whether the KEPI instrument measured respondents' cultural competence without discrimination across female and male students and URM and non-URM students (Table 4). The Knowledge of Diversity domain performed well in measuring participants' ability with no DSF for gender and race presented in any items. However, non-pervasive DSF for gender was detected for Culture-Centered Practice domain items 16 and 17. Item 16 also showed non-pervasive DSF for race, and item 9 showed pervasive DSF across race. In addition, two items (20, 26) on the Efficacy of Assessment domain showed non-pervasive DSF across gender, and three items (18, 25 and 26) showed pervasive DSF for race.

The z-statistic is positive when the item difficulty for female and URM dental students is larger than the estimate for their counterparts. Of the DSF items, items 16, 17, and 20 were more difficult to endorse for male students, while item 26 was more difficult to endorse for female students. In addition, items 16 and 18 were more likely to be endorsed by non-URM students, while items 9 and 26 tended to be much too easy for URM students.

Table 4. Items showing DSF for gender and race

	Threshold 1		Threshold 2		Threshold 3	
Item	statistics	p-value	statistics	p-value	statistics	p-value
Gender						
Item 16 CCP	-0.13	0.89	2.06	0.04	1.24	0.21
Item 17 CCP	1.37	0.17	2.20	0.03	1.54	0.13
Item 20 EOA	1.86	0.06	2.98	<0.001	2.87	<0.001
Item 26 EOA	-2.27	0.02	-1.93	0.05	-1.80	0.07
Race						
Item 9 CCP	-2.09	0.04	-3.42	<0.001	-3.86	<0.001
Item 18 EOA	4.01	<0.001	4.28	<0.001	3.36	<0.001
Item 25 EOA	2.23	0.03	2.31	0.02	2.31	0.02
Item 26 EOA	-1.12	0.26	-2.13	0.03	-2.67	0.01

CCP=Culture-Centered Practice domain; EOA=Efficacy of Assessment domain
Note: The significant level is at 0.05.

Discussion

We found no previous studies that have examined the psychometric properties of KEPI using IRT theory. However, KEPI has been increasingly applied in dental and other health care disciplines to assess and measure cultural competence for oral health care providers in an effort to guide changes in dental education programs and curriculum modifications in academic dental institutions. Thus, it is essential to identify whether the KEPI can produce reliable estimates of participant ability in population-based samples. The PCM model is one of the most widely used IRT models for polytomous items. Unlike the Rating Scale Model, the number of response categories in the PCM are allowed to vary across items, and item discrimination parameters are not fixed. Edelen and Reeve found that less constrained models produced more accurate estimations of data.[23] Thus, we used a Rasch-derived PCM model in this study to assess the psychometric properties of the KEPI, which have been well validated using CTT theory.

Our research findings demonstrated three domains in the KEPI: Knowledge of Diversity, Culture-Centered Practice, and Efficacy of Assessment were all producing reliable item and person estimates in a large population-based sample of U.S. dental students. The Rasch analysis showed both high item and person separation and reliability, indicating that the items posed high discrimination ability and that respondents could be appropriately discriminated between each other based on their estimated ability.

Also, each KEPI domain demonstrated acceptable internal consistency. Moreover, the item data in each domain conformed to the Rasch measurement model, indicating that the items measured a single construct in the manner that is expected under latent measurement model theory. In KEPI, for successive response categories across all subscales, the Rasch PCM found that all were located in the expected order. No KEPI items showed "disordered threshold." The average and step measures for each item advanced monotonically in the expected direction across the four-point rating scale; no misfitting category was identified. Also, the evidence demonstrated that it was difficult for respondents to discriminate more than five response categories, which could easily lead thresholds to be disordered.[24,25]

The Rasch PCM model indicated that the Culture-Centered Practice domain displayed reasonable fit with no misfit for any items and that most items fit the Culture-Centered Practice and Efficacy of Assessment domains well with the exception of items 1, 18, and 22. The acceptable item fit findings indicated that the assumption of monotonicity in IRT was met. It should be noted that 5% of the respondents in the Knowledge of Diversity domain were identified as poorly fitting persons. As reported by Wright and Masters, a large number of flagged respondents suggests that the instrument may not work well for measuring the ability of sample who took the survey or perhaps participants were not providing thoughtful answers.[19] Also, if moderate or small numbers of respondents were detected for misfit, as in our case, they should potentially be excluded from the final analysis. It should be noted that participants were well able to discriminate among less than five response categories, beyond which points the "disordered threshold" problem was likely to occur.

Targeting of the Knowledge of Diversity domain was less than desirable as can be seen in the person-item maps, in which the difference between item and person means was larger than 5.00. Items were relatively homogenous in difficulty of endorsement and were generally distributed on the left side of respondent ability distribution, indicating that the items were, on average, too easy for the participants. The targeting problem could be due to the significant floor effects in the subscale. Indeed, around 80% of the respondents selected "agree" and "strongly agree" response option, while only less than 1% selected "Very Limited" option. We can infer that the subscale provided a lot of information about dental students who demonstrated a high level of knowledge of racial diversity, but that it was not well designed to measure dental students who posed relatively low levels of Knowledge of Diversity. As suggested by Linacre,[26] collapsing categories may be a useful way to optimize the effectiveness of polytomous response categories and to improve the model fit indices of IRT as well as reducing the total time and burden to complete the survey.[27] In fact, collapsing or increasing the number of response categories is a trade-off between instrument effectiveness and the overall model fit.[28] If the item numbers are small and item discrimination is problematic, a researcher may consider increasing response categories to improve the model fit. However, when the items have high discrimination or the number of items is large, and the model fit is relatively poor, then it is feasible to consider less response categories.

Optimal targeting was observed in the Culture-Centered Practice and Efficacy of Assessment do-

mains, indicating survey items ideally measured the participants' ability in the domains. Our analyses revealed a different type of DSF: non-pervasive and pervasive, which showed a different type of bias. No DSF was found for the Knowledge of Diversity domain. However, non-pervasive DSF by gender and pervasive DSF by race were found on certain items across the Culture-Centered Practice and Efficacy of Assessment domains. As reported by Satcher, any significant DIF effects may be the result of ethnic and cultural diversity or gender expectations that lead to certain feelings or behavior more or less "strong/sensitive" across different subgroups.[29] For example, item 16 ("At the present time, how would you rate your own understanding of the following term: Pluralism?"), item 17 ("At the present time, how would you rate your own understanding of 'cultural encapsulation'?") and item 20 ("How would you rate your ability to effectively secure information and resources to better serve patients of different ethnic/cultural groups?") were more difficult to endorse for male students. Perhaps this is the reason that female and URM groups are considered more sensitive to racism than their counterparts.[30,31]

In addition, four items that exhibited pervasive DSF across race warrant concern. Non-URM students were more likely to agree that "In dentistry, patients from different ethnic/cultural groups should be given the same treatment," while URM students were more likely to be confident in rating their understanding of "patient management" for treating patients from culturally diverse groups and in their ability to accurately assess the oral health care needs of patients with disabilities and those from low socioeconomic situations. The reason for these group differences may indicate cultural taboos or may be simply because exposure to people from culturally diverse groups is more common among some racial groups than others. Previous research on color blind racial attitudes among health care providers found different levels of color blind racial attitudes among URM and non-URM groups.[31] For example, URM students had significantly lower scores on Racial Privilege, Institutional Discrimination, and Blatant Racial Issue. Ancis et al. reported that non-URM students in the U.S. usually experience less racial discrimination and thus have reported less interracial tension or racial conflict.[32] Moreover, non-URM students showed significantly better understanding of the term "pluralism" than their counterparts. Perhaps this could be due to the wording of the questions or that it was more difficult for minority groups to understand, leading to either an endorsement bias or an actual variation in awareness of cultural diversity.

These results are limited in generalizing across states and among larger subgroups as the findings may be more representative of a specific state or institutional culture. First, with regard to the PCM model applied in the study, it was not representative of all possible models that exist in realistic situation. For future research, a variety of IRT models such as graded response model and rating scale model could be used to fit KEPI scale, respectively, and to compare which model results in best model fit. In addition, we noticed that, for some items, there were few responses in the lowest or highest categories. In this case, researchers may wish to score the four-point scale into binary scale (0/1), and future research could examine if using dichotomous IRT (1PL, 2PL, and 3PL) to model these responses leads to more accurate parameter estimates. Further studies that utilize much larger sample sizes across wider areas with various minority groups are needed. Previous studies reported that a high level of color blind racial attitudes and pervasiveness of racial stereotypes existed among clinical faculty members.[31,33] Ancis et al. found that clinical faculty racism was a potential stressor for URM students, so we recommend examining DSF effects across students and faculty from different racial groups.[32]

The development of the KEPI is responsive to the Culturally and Linguistically Appropriate Health Care Services (CLAS) standards. Despite limitations reported in this study, KEPI was found to be a useful and reliable instrument for measuring Knowledge of Diversity, the skills in Culture-Centered Practice, and Efficacy of Assessment for oral health care providers.[34]

Conclusion

This study examined the psychometric properties of the KEPI using Rasch analysis to assess differential item functioning by dental student gender and race. The results provided valid evidence of high internal reliability, measurement properties, unidimensionality of the three subscales of KEPI, ideal targeting, and well response category functioning. Items that posited significant pervasive or non-pervasive DSF across gender and racial groups do not suggest that the questions should be deleted. It simply indicated that item bias makes group comparisons problematic because the items showed slightly

different properties across the subgroups that were examined. Special care should be taken when conducting gender/racial comparisons using KEPI scale.

REFERENCES

1. Like R. A failure to communicate caring for patients with limited English proficiency. Focus Multicult Healthcare 2007;3(4):6-9.
2. Agency for Healthcare Research and Quality. Cultural competence health practitioner assessment. At: innovations.ahrq.gov/qualitytools/cultural-competence-health-practitioner-assessment. Accessed 13 Dec. 2017.
3. Mirsu-Paun A, Tucker CM, Herman K, Hernandez CA. Validation of a provider self-report inventory for measuring patient-centered cultural sensitivity in health care using a sample of medical students. J Community Health 2010;35:198-207.
4. Rest JR, Narvaez D, Thoma SJ, Bebeau MJ. DIT2: devising and testing a revised instrument of moral judgment. J Educ Psychol 1999;91(4):644-59.
5. Camphina-Bacote J. The inventory for assessing the process of cultural competence among health care professionals. At: www.transculturalcare.net/iapcc-r.htm. Accessed 14 Dec. 2017.
6. Behar-Horenstein LS, Garvan CW, Moore TE, Catalanotto FA. The knowledge, efficacy, and practices instrument for oral health providers: a validity study with dental students. J Dent Educ 2013;77(8):998-1005.
7. Behar-Horenstein LS, Garvan CW. Relationships among the knowledge, efficacy, and practices instrument, color-blind racial attitudes scale, Deamonte Driver survey, and defining issues test 2. J Dent Educ 2016;80(3):355-64.
8. Garvan GJ, Garvan CW, Behar-Horenstein LS. Developing and testing the short-form knowledge, efficacy, and practices instrument for assessing cultural competence. J Dent Educ 2016;80(10):1245-52.
9. Kan CC, Breteler MH, Timmermans EA, et al. Scalability, reliability, and validity of the benzodiazepine dependence self-report questionnaire in outpatient benzodiazepine users. Compr Psychiatry 1999;40(4):283-91.
10. Haywood SH, Goode T, Gao Y, et al. Psychometric evaluation of a cultural competency assessment instrument for health professionals Med Care 2014;52(2):e7.
11. Court H, Greenland K, Margrain TH. Measuring patient anxiety in primary care: Rasch analysis of the 6-item Spielberger state anxiety scale. Value Health 2010;13(6):813-9.
12. Muthén LK, Muthén BO. Mplus Editor (version 7.0). Los Angeles: Muthen & Muthen, 2012.
13. Muthén BO. Goodness of fit with categorical and other nonnormal variables. SAGE Focus Editions 1993;154:205.
14. Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociol Methods Res 1992;21(2):230-58.
15. Bentler PM. Comparative fit indexes in structural models. Psych Bull 1990;107(2):238.
16. Ponocny I. Nonparametric goodness-of-fit tests for the Rasch model. Psychometrika 2001;66(3):437-59.
17. Mair P, Hatzinger R, Maier MJ, et al. Package eRm. Version 0.14-0. 2016.
18. Bond T, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. New York: Routledge, 2001.
19. Wright BD, Masters GN. Rating scale analysis: Rasch measurement. Chicago: MESA Press, 1982.
20. Linacre JM. Investigating rating scale category utility. J Outcome Meas 1999;3:103-22.
21. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002;3(1):85-106.
22. Wright BD, Stone MH. Best test design. Portland, OR: Rasch Measurement, 1979.
23. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16(1):5.
24. Khadka J, Gothwal VK, McAlinden C, et al. The importance of rating scales in measuring patient-reported outcomes. Health Qual Life Outcomes 2012;10(1):1.
25. Fenwick E, Rees G, Pesudovs K, et al. Social and emotional impact of diabetic retinopathy: a review. Clin Exp Ophthalmol 2012;40(1):27-38.
26. Linacre JM. Rasch model estimation: further topics. J Appl Meas 2004;5(1):95-110.
27. Gothwal VK, Wright TA, Lamoureux EL, Pesudovs K. Rasch analysis of the quality of life and vision function questionnaire. Optom Vis Sci 2009;86(7):E836-44.
28. Olivares J, Sánchez-García R, López-Pina JA, Rosa-Alcázar AI. Psychometric properties of the social phobia and anxiety inventory for children in a Spanish sample. Span J Psychol 2010;13(2):961-9.
29. Satcher D. The surgeon general's call to action to promote sexual health and responsible sexual behavior. Am J Health Educ 2001;32(6):356-68.
30. Su Y, Behar-Horenstein LS. Color-blind racial beliefs among dental students and faculty. J Dent Educ 2017;81(9):1098-107.
31. Neville H, Spanierman L, Doan BT. Exploring the association between color-blind racial ideology and multicultural counseling competencies. Cultur Divers Ethnic Minor Psychol 2006;12(2):275.
32. Ancis JR, Sedlacek WE, Mohr JJ. Student perceptions of campus cultural climate by race. J Couns Devel 2000;78(2):180-5.
33. McCann AL, Lacy ES, Miller BH. Underrepresented minority students' experiences at Baylor College of Dentistry: perceptions of cultural climate and reasons for choosing to attend. J Dent Educ 2014;78(3):411-22.
34. U.S. Department of Health and Human Services. National culturally and linguistically appropriate health care services (CLAS) standards. At: minorityhealth.hhs.gov/omh/browse.aspx?lvl=2&lvlid=53. Accessed 3 Feb. 2018.