Comparison of Two Rating Scales With the Orofacial Esthetic Scale and Practical Recommendations for Its Application.

Swaha Pattanaik (patta025@umn.edu)
University of Minnesota Twin Cities Campus: University of Minnesota Twin Cities
https://orcid.org/0000-0002-2352-7310
Mike John
University of Minnesota School of Dentistry
Seungwon Chung
University of Minnesota
San Keller
American Institutes for Research

Research

Keywords: Orofacial esthetic scale, scaling formats, 5-point numerical rating scale, 11-numerical rating scale oral health, item response theory, psychometric properties, dental patient-reported outcome measure, patient-centred care, standardization, reliability, validity, oral health impact profile

Posted Date: May 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-473429/v1

Abstract

Purpose

We compared measurement properties of 5-point and 11-point response formats for Orofacial Esthetic Scale (OES) items to determine whether collapsing the format would degrade OES score precision.

Methods

Data were collected from a consecutive sample of adult dental patients from HealthPartners dental clinics in Minnesota (N=2,078). We fitted an Item Response Theory (IRT) model to the 11-point scale and six, derived 5-point scales. We compared all response formats using test (or scale) information, correlation between the IRT scores, Cronbach's alpha estimates for each scaling format, correlations based on the observed scores for the seven OES items and the eighth global item, and the relationship of observed and IRT scores to an external criterion using orofacial appearance (OA) indicators from the Oral Health Impact Profile (OHIP).

Results

The correlations among scores based on the different response formats were uniformly high for observed (0.97-0.99) and IRT scores (0.96-0.99); as were correlations of both observed and IRT scores and the OHIP measure of OA (0.65-0.69). Cronbach's alpha based on any of the 5-point formats (α = 0.95) was nearly the same as that based on the 11-point format (α = 0.96). The weighted total information area for five of six, 5-point derived formats was 98% of that for the 11-point scale

Conclusions

Our results support the use of scores based on a 5-point response format for OES items. The measurement properties of scores based on a 5-point response format are comparable to those of scores based on the 11-point format.

Introduction

A major reason for dental patients to seek treatment is to enhance their orofacial appearance (OA) [1], which influences their self-esteem and social interactions as dental arrangement plays an important role in determining perceived personal beauty and success [2-5]. Research shows, faces with crowding and spacing of teeth appeared less intelligent, beautiful, and sexually attractive, and even socioeconomically disadvantaged to others than the same faces with ideal teeth arrangement [5]. OA or esthetics is thus an important dental patient-reported outcome (dPRO); and one of the four dimensions, or elemental building blocks of the dental patients' oral health related quality of life (OHRQoL) [6-7]. Patient-reported OA data would help dental patients and providers in shared treatment decisions[8], consequently, improving dental treatments effectiveness [6,9] and value-based oral health care [10]. The Orofacial Esthetic Scale (OES)

and the Oral Health Impact Profile (OHIP) are the dental patient-reported outcome measures (dPROM) or instruments commonly used to measure OA [6,9].

The OES was developed in a Swedish prosthodontic patient population [3]. Initially, it was measured on an 11-point numeric rating scale (0 = very dissatisfied, 10 = very satisfied) [3]. Since then, the OES has been translated and adapted for different countries [4, 11-17]. While some of these versions have used the original 11-point response scale [4, 14], others have used a more concise 5-point response scale (1 = unsatisfactory, 5 = excellent) [11-13]. The 5-point adjectival rating scale is the most widely used response format for dPROMs; in line with medical Patient-Reported Outcome Measures, or PROMs [6].

Application of PROMs with a 5-point adjectival format has conceptual and technical advantages. Compared to an 11-point scale, a 5-point scale is more comprehensible and easier to use, [11] and its conciseness can improve response rate and quality [18]. A technical advantage of the 5-point format is presence of fewer parameters when the response-scale data are modeled. However, no studies have been conducted that compare 5-point and 11-point formats in dPROMs.

Currently there is no consensus on the ideal response format for OES and other dPROMs assessing OA, which hinders efforts toward standardization of OA assessment. With regard to PROMs in general, a recent review of the evidence concluded that the issue required further empirical study within the context of particular therapeutic areas as results might vary according to disease and therapeutic specialty [19].

The purpose of our study was to compare measurement properties of the 5-point and 11-point response formats for the OES, to determine whether collapsing the 11-point format to a 5-point format would degrade OES score precision.

Methods

Study population, recruitment, and data collection

We recruited adult dental patients from HealthPartners dental clinics in Minnesota (N=2,115). Removing individuals who did not respond to OES items leaves N=2,078. Details about data collection and recruitment have also been provided in previous research papers [4,16].

Measure: Orofacial Esthetic Scale

Details of the OES development have been published elsewhere [3] and are briefly summarized here. OES consists of seven items addressing specific esthetic components (face, facial profile, mouth, rows of teeth, tooth shape/form, tooth color, gums) and one item assessing the overall impression [Table 1]. Originally, the response format was a 0 to 10 numeric rating scale, anchored only with “very dissatisfied” and “very satisfied” (with appearance) at the extremes of 0 and 10, respectively. Scores of items 1 through 7 can be summed up to form an OES summary score that can range from 0 through 70, with higher scores representing less impaired esthetics [3,16]. The eighth item represents an overall impression of OA and no specific esthetic component, so it is not included in any of the subscale scores. The OES

was initially tested among Swedish prosthodontic patients [3]. Since then, the validity of OES scores has been assessed for other dental patients [4,16], and general populations [20] in several other countries.

Additional measure: Oral Health Impact Profile

Details of OHIP development have been published elsewhere [21] and are briefly summarized here. The Oral Health Impact Profile (OHIP) is the most widely used instrument to measure OHRQoL in adults with oral disease conditions[6]. OHIP is a more comprehensive instrument than OES. While OES only measures OA, the OHIP measures seven conceptual dimensions of impact corresponding to Locker's model of Oral Health based on the World Health Organization's (WHO) classifications of disease impacts[22]. The dimensions of impact are functional limitation, physical pain, psychological discomfort, physical disability, psychological disability, social disability, and handicap. Originally, the OHIP questionnaire had 49 items organized into the seven dimensions, later researchers developed 14- and 5- item versions. Among them, questions 3, 4, 19, 20, 22, 31 capture OA (see Table 1). For each question, respondents are asked to indicate on a 5-point Likert scale (0- never, 1- hardly ever, 2-occassionally, 3-fairly often, and 4-very often) according to how frequently they experienced each problem within the past twelve months. Respondents may also be offered a "don't know" option for each question. All impacts in the OHIP are conceptualized as adverse outcomes, thus, a higher score indicates more negative impacts of oral health problems. Overall OHIP scores are computed in two ways. The simpler scoring method is to sum all 49 unweighted items. The second method is to standardize the seven subscale scores and then sum those standard scores.

Table 1 OES and OHIP Items

OES
How do you feel about the appearance of your face, your mouth, your teeth and your replacements (prostheses, crowns, bridges and implants)?
0: Very dissatisfied — 10: Very satisfied
1. Your facial appearance
2. Appearance of your facial profile
3. Your mouth's appearance (smile, lips, and visible teeth)
4. Appearance of your rows of teeth
5. Shape/form of your teeth
6. Color of your teeth
7. Your gum's appearance
8. Overall, how do you feel about your face, your mouth and your teeth?
OHIP
0:Never — 10: Very Often
3. Have you noticed a tooth which doesn't look right?
4. Have you felt that your appearance has been affected because of problems with your teeth, mouth, or dentures?
19. Have you been worried by dental problems?
20. Have you been self-conscious because of your teeth, mouth, or dentures?
22. Have you felt uncomfortable about the appearance of your teeth, mouth, or dentures?
31. Have you avoided smiling because of problems with your teeth, mouth, or dentures?

*OHIP items are numbered in the same way as in the original questionnaire

Statistical analysis

The hypothesis of our study was – when a 11-point format is collapsed to a 5-point format, psychometric properties of OES scores will not be compromised. Multiple options for the 5-point format exist if the study is designed to compare the 11-point format with a "derived" 5-point format. Thus, as the first step, we defined several “plausible” 5-point formats to be investigated in the study, each created by a different method of collapsing the 11-point scale. A challenge was that the 11 points be assigned relatively evenly among five categories. Hence, we set up two simple principles for grouping categories within the 11-point scale: ***Rule 1*** was to disallow 4-category grouping, and ***Rule 2*** was to disallow 1-category grouping with *Exception* (1-category is allowed) at the beginning and the end of the format. ***Rules 1*** and ***2*** yielded

balanced response groups, meaning that only groupings of 2- and 3-categories existed. Note that *Exception* corresponds to Patient-Reported Outcomes Measurement Information System (PROMIS) guidelines [23]. Following ***Rules 1*** and ***2*** coupled with *Exception*, we obtained six "derived" 5-point formats (see Figure. 1).

Descriptive analysis

For the 11-point scale, we plotted histograms for Item 1-8 to examine the frequencies in each response option.

Classical test theory (CTT)

Reliability analysis (Internal consistency)

We computed Cronbach's alpha [24] for the 11-point format and six derived 5-point formats to assess any changes in OES reliability. Also, we used a Bootstrap confidence interval for Cronbach's alpha because the distribution of item scores could not be well approximated by a normal distribution.

Validity analysis (Correlation analysis based on sum scores)

We computed correlations between the 11-point format and the six derived 5-point formats based on the observed scores (raw scores) for the seven items addressing specific esthetic components as well as the summary score. If the correlation is high ($r>0.95$), then we can infer that there is a close similarity in the scores between the two formats. Also, within each format, we computed the correlation between the aggregated seven items and a global item assessing overall impression. If these correlations were similar in size, indicating a similar relationship to overall OA, we could assume the scores based on the different response scale formats have a similar interpretation and so are measuring the same "construct." Furthermore, we computed correlations between summed scores of the 11-point and 5-point formats of OES, and that of the OA indicators from OHIP to determine whether the relationship of the scores to the external criterion was invariant across the two response formats. We performed two correlation analyses, one with all six indicators of the OA dimension (item 3, 4, 19, 20, 22, and 31) and another with only 4 indicators (item 3, 4, 22, and 31).

Item Response Theory (IRT)

Item Response Theory (IRT) is a psychometric theory that refers to a family of associated statistical models that predict responses to a given set of items based on each item's properties and the respondent's position on continuum of latent trait of interest (OA) measured by the scale (OES) [25]. Unidimensionality of the scale is required to progress with IRT based analysis. Previous studies have supported unidimensionality of OES [3,16]. Samejima's graded response model (GRM) was used for calibration of our items [26]. This model is suitable for ordered scoring categories, which is the case for OES scale. GRM specifies the probability of responding to a particular category or higher, versus responding to lower categories for each value of latent variable (trait) θ, which is (perceived) OA in our

study. In GRM, each item is characterized by one slope parameter, and category threshold or location parameters at which the probability of responding to a particular category or higher is 0.5. Note that the number of category threshold parameters for an item equals one less than the number of categories. With the GRM parameters, we can derive category response curves (CRCs). A CRC represents the probability of responding in a particular category as a function of trait level θ. We fitted a GRM to the 11-point numeric rating scale (0 = very dissatisfied, 10 = very satisfied) and the six plausible options of the derived 5-point rating scale.

Reliability analysis (Item/Test information)

Information is analogous to reliability of measurement, and it is provided both at item and test (scale) level. An item information function or curve shows the amount of (Fisher) information an item contains along the continuum of a latent trait, i.e., OA [27]. CRCs from GRM can be transformed into an item information function. Multiple factors contribute to item information for polytomous models. For GRM, magnitude of the slope parameter, and the distance between the category thresholds or location parameters determine the amount of information. The test (or scale) information curve is obtained by simply summing the item information curves. Also note that the information function is related to measurement precision. Specifically, (conditional) information is inversely related to standard error of measurement (SEM) [28].

Furthermore, we computed the total information area (TIA), which represents the area under the test (or scale) information. To account for differential contribution due to unequal number of respondents along the latent trait continuum, we weighted the TIA with the proportion of respondents in each interval of the latent trait. We will term this index “weighted total information area (TIA).”

Validity analysis (Correlation analysis based on IRT scale scores)

We estimated the IRT scores using the GRM for each response format. The IRT scores refer to person location estimates from an IRT model. In IRT scoring, a respondent’s location on the OA continuum is obtained by utilizing the respondent’s item response pattern coupled with estimated item parameters [27]. Specifically, we obtained the *expected a posteriori* (EAP) scores [29]. EAP uses the mean of the posterior distribution as the latent traits. Then, we calculated the correlation between the IRT scores based on the 11-point scale and each of the six 5-point scales. Furthermore, we computed correlations between the EAP scores from the 11-point format and the derived 5-point formats of the OES and those from the OA indicators of the OHIP. Note that the analysis is identical to what we described above for the CTT framework, but now the correlation analysis was performed using the scores from IRT analysis instead of sum scores. All analyses were performed using the *mirt* package in R [30].

Results

Descriptive analysis

Figure 2 shows histograms of the 11-point scale for Items 1-8. Generally, they show a left-skewed distribution. Category 10 shows the highest frequency, suggesting that a majority of respondents were “very satisfied” with each component of OA (Items 1- 7), and were “very satisfied” overall with their OA (Item 8). Interestingly, patients’ responses to Item 6 (“Color of your teeth”) was relatively evenly spread.

CTT

Reliability analysis (Internal consistency analysis)

Cronbach alpha estimates for the 11-point format and six derived 5-point formats with their 95% confidence intervals (CIs) are presented in Table 2. We observe that the alpha estimates of the 5-point formats barely decreased. The alpha estimate from the 11-point format was 0.95, and the estimates from the 5-point format were 0.94 in all six possible formats.

Table 2 Cronbach alpha estimates for the 11-point format and the six derived 5-point formats

Response Format	Alpha
11-point	0.95 (0.94, 0.95)
5-point (Option 1)	0.94 (0.94, 0.95)
5-point (Option 2)	0.94 (0.93, 0.95)
5-point (Option 3)	0.94 (0.94, 0.95)
5-point (Option 4)	0.94 (0.94, 0.95)
5-point (Option 5)	0.94 (0.93, 0.95)
5-point (Option 6)	0.94 (0.93, 0.94)

Validity analysis (Correlation analysis)

Correlations between the 11-point format and the six, 5-point formats based on the raw scores are presented in Table 3. The first seven columns show the item correlation between the response formats for Items 1 – 7, and the last column is the correlation based on the summary score of the 7 items. The summary score correlation was well above 0.98 across all the 5-point formats (Options 1-6), suggesting that there is a very strong relationship between the two response formats. The correlation examined by each item also indicates that the 5-point formats are highly correlated with the 11-point format. Specifically, it ranges from 0.97 to 0.99. Correlation of 0.97, 0.98, and 0.99 were measured with a standard error of 0.005, 0.004, and 0.003, respectively, providing all correlations in a high precision.

Table 3 Correlation estimates between the 11-point format and the six derived 5-point formats based on item scores and summary scores.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Summary Score
Option 1	0.97	0.97	0.97	0.98	0.97	0.97	0.97	0.98
Option 2	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99
Option 3	0.98	0.98	0.99	0.99	0.98	0.98	0.98	0.99
Option 4	0.98	0.98	0.99	0.99	0.98	0.98	0.98	0.99
Option 5	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99
Option 6	0.97	0.97	0.98	0.98	0.97	0.98	0.97	0.98

In addition, correlations between the aggregated items (Item 1 - 7) and the global item (Item 8) and their 95% CIs are presented in Table 4. The correlation was 0.91 for the 11-point scale, and it ranged from 0.88 to 0.90 for the six 5-point scale options. Overall, the difference was minimal between the 11-point scale and any of the 5-point scales. Given that the correlations were similar in magnitude, we determine that the relationship between the global item score and the seven-item composite scores remain largely the same even after collapsing 11 categories to 5. In other words, the six, 5-point formats are measuring the same “construct”.

Table 4 Correlation estimates between the aggregated items (Item 1-7) and the global item (Item 8) and their 95% CIs.

	r (95% CI)
11-point	0.91 (0.90, 0.91)
5-point (Option1)	0.88 (0.87, 0.89)
5-point (Option2)	0.89 (0.88, 0.90)
5-point (Option3)	0.89 (0.88, 0.90)
5-point (Option4)	0.89 (0.88, 0.90)
5-point (Option5)	0.89 (0.88, 0.90)
5-point (Option6)	0.88 (0.87, 0.89)

Table 5 shows estimated correlations and their 95% CIs between the summed scores from the 11-point format and the six derived 5-point formats of the OES and those of the OA indicators of OHIP. The sum scores from the two response formats (11-point and six options of 5-point) of the OES scale correlate similarly with the sum score of the OA dimension of OHIP scale for both six and four indicators. We observe negative correlation estimates as the scoring system of OES is inverse to that of OHIP. While for OHIP the higher the score means worse OA (‘bad’ OA), for OES, higher score means better OA (‘good’ OA).

Table 5 Correlations between the sum scores of the OES scales (11-point scale and the 5-point scales) and the external measure (OA from OHIP) and their 95% CI

	OA from OHIP (6 items)	OA from OHIP (4 items)
11-point	-0.69 (-0.72, -0.68)	-0.69 (-0.71, -0.67)
5-point (Option1)	-0.67 (-0.71, -0.66)	-0.67 (-0.70, -0.65)
5-point (Option2)	-0.69 (-0.72, -0.68)	-0.69 (-0.71, -0.66)
5-point (Option3)	-0.68 (-0.72, -0.67)	-0.68 (-0.70, -0.66)
5-point (Option4)	-0.68 (-0.72, -0.67)	-0.68 (-0.70, -0.66)
5-point (Option5)	-0.68 (-0.72, -0.67)	-0.68 (-0.71, -0.66)
5-point (Option6)	-0.68 (-0.72, -0.67)	-0.68 (-0.71, -0.66)

IRT

Reliability analysis (item/test information)

We compared the test (or scale) information functions of the 11-point scale and the six 5-point scale options to examine the loss of information when a 5-point scale is used at the scale level (Figure 3). We found that some loss of information occurred when going from the 11-point scale to the 5-point scale. The shapes of information functions for the six 5-point scales differed. Option 1 showed loss of more information in the middle range (level between -1.5 and .5), with the greatest information loss occurring at around 0. We observed that the information loss mostly occurs within locations of high information, but barely where information is low. The other 5-point scales (Options 2 – 6) showed relatively similar patterns in the way the information curves for the 5-point scales were shrunken compared to that of the 11-point scale. The loss of information was relatively even across the range of the latent trait (.

Examining the weighted TIA (Table 6), we found that the information of the 5-point scales resulted in above 98% of that of the 11-point scale for all of the collapsed options except Option 6. In Option 1, even though loss of information appeared substantial in the middle range of latent trait (θ) (see Figure 3), the TIA when weighted by the unequal distribution of respondents was nearly the same as that of the 11-point scale. On the other hand, Option 6 where the loss of information occurred for the high latent trait (θ) resulted in a relatively greater reduction in the proportion of the weighted TIA due to the left-skewed distribution, as shown in Figure 2. However, even for Option 6 where the information loss was the highest, we observed about 88%.

Table 6 Weighted total information area (TIA) for the OES with the 11-point item response format compared to the OES with alternative 5-point item response formats

	TIA	Ratio (5-point/11-point)
11-point	29.52	
5-point (Option1)	29.16	0.99
5-point (Option2)	28.93	0.98
5-point (Option3)	29.15	0.99
5-point (Option4)	29.15	0.99
5-point (Option5)	29.12	0.99
5-point (Option6)	26.07	0.88

Validity analysis (IRT Scoring)

IRT scores were estimated, and the scores of the six 5-point scale options were compared against those of the 11-point scale. The correlations between the EAP scores of the 11-point scale and those of the 5-point scales and their 95% CI are displayed in Table 7. For Option 1, the correlation is almost 1, and the other options also shows high correlations ranging from 0.93 to 0.96. As expected by the weighted TIA, Option 6 showed the lowest correlation. Nevertheless, correlation was greater than 0.90 in all the scenarios.

Table 7 Correlations between the EAP scores of the 11-point scale and the 5-point scales and their 95% CI

	r (95% CI)
Option 1	0.99 (.99,.99)
Option 2	0.96 (.95,.96)
Option 3	0.96 (.96,.96)
Option 4	0.96 (.96,.96)
Option 5	0.96 (.96,.96)
Option 6	0.93 (.92,.93)

Table 8 displays estimated correlations and their 95% CIs between the EAP scores from the 11-point format and the six derived 5-point formats of the OES along with those of the OA indicators of OHIP. It showed that both response formats have nearly identical correlations with the external measure.

Table 8 Correlations between the EAP scores of the OES scales (11-point scale and the 5-point scales) and the external measure (OA from OHIP) and their 95% CI

	OA from OHIP (6 items)	OA from OHIP (4 items)
11-point	-0.66 (-0.68, -0.63)	-0.65 (-0.67, -0.62)
5-point (Option1)	-0.66 (-0.68, -0.63)	-0.65 (-0.67, -0.62)
5-point (Option2)	-0.66 (-0.68, -0.63)	-0.65 (-0.67, -0.62)
5-point (Option3)	-0.66 (-0.68, -0.63)	-0.65 (-0.67, -0.62)
5-point (Option4)	-0.66 (-0.68, -0.63)	-0.65 (-0.67, -0.62)
5-point (Option5)	-0.66 (-0.68, -0.63)	-0.65 (-0.67, -0.62)
5-point (Option6)	-0.67 (-0.69, -0.65)	-0.66 (-0.68, -0.63)

Discussion

On rigorous testing of the research hypothesis using CTT- and IRT-based approaches; we found that the measurement properties of OES were not compromised when an 11-point format was collapsed to a 5-point format. The internal consistency analysis showed that scale reliability hardly decreased when the number of response categories was reduced. Also, the correlation analyses based on observed or raw scores showed that scale validity was not undermined. Specifically, we found a strong linear relationship between the summary scores of the 5- and 11-point scales. The item score correlation results also supported similarity between the two scales. Additionally, we observed high correlations between the seven OES items and the global assessment item across the 11- and 5-point scales, implying that both measured the same construct (OA).

We scrutinized item and test (or scale) information for both formats to assess IRT-based reliability and found some loss of information for the 5-point format. This was expected when reducing the number of response options, given that each response category provided information for polytomous items. Considering the relationship between information and SEM; loss of information meant decrease in precision of measurement, and in reliability. Importantly, the IRT analysis helped pinpoint where information loss occurred heavily, as information is given as a function of latent trait and pertinent to individual score [31]. We evaluated six 5-point response formats created by collapsing categories in different manners. While the location and the amount of information loss differed across the six 5-point response formats, the general trend was that scale reliability was sacrificed to a limited extent when using the 5-point format. However, examining the impact of loss of information on individual scores, we observed that it was overall not meaningful for the IRT-based scores, particularly the EAP scores. For all the 5-point response formats, the correlations between EAP-scores for the 11-point with any of the 5-point formats were greater than 0.9.

In general, the optimum number of response categories in rating scales has been widely debated, yet there is no consensus on the best scaling format [11, 32]. Coarser scales (with fewer categories) tend to lower the discriminating power that the respondents might be capable of, while, finer rating scales (with

several categories) may go beyond their discriminating ability [32]. Previous researchers investigating an optimal scaling format found that increasing the number of categories did not necessarily improve scale reliability and validity [32]. The specific number of scale points beyond which increases in scale reliability and discrimination become negligible, has also been a contentious issue [33-35]. Garner explained that this number beyond which there will be no improvement in scale discrimination, is a function of the amount of discriminability inherent in the items rated [34]. Maydeu-Olivares et al. concluded that the choice of psychometric framework also influences the effect of response format on the reliability and validity of scores [36]. For example, within the IRT framework, they suggested that applied researchers consider factors such as the number of items in an instrument, the items' discriminating ability, and the goodness of fit of the model in selecting the optimal response format [36].

Previous researchers have successfully applied a 5-point OES to clinical settings [11-13]; in fact, Persic and colleagues strongly recommended its use due to practical benefits for face-to-face and telephone interviews [11]. However, unlike our study, these previous researchers did not perform a comparative analysis of the 11-point to the 5-point response format. Ours is the first study to conduct an in-depth comparison of these two scaling formats commonly used for responses to OES and other dPROM items. Within the area of patient (medical) reported outcomes, researchers have compared different response formats for a given scale [36-38], using a methodological approach that differed from ours. For example, Hendriks et al. and Garratt et al. concluded that compared to the 10-point response scale, the 5-point scale produced better quality data with fewer missing data, more variance, distributions with less skew and kurtosis [37] and lower floor and ceiling effects [38].

Strengths and Limitations

We compared measurement properties of item- and total scores based on responses to the 11-point scale with scores based on responses to six plausible 5-point scales. The 5-point response scales were derived from collapsing the response categories on the 11-point scale. Our study may be limited due to this research design, as we did not administer both the scales separately to the patients. Maydeau-Olivares used a repeated measures design [36] where a group of students was divided into two samples. Each sample received a test battery consisting of four instruments, with a target questionnaire that was administered three to four times, each time with a different number of response alternatives. This design helped them capture variability in measurement properties due to respondent in addition to that due to number of response alternatives. Other researchers randomized the patients in their study to receive either a 5-point scale or a 10-point scale. This design helped them compare the quality of data yielded by the two response formats under conditions similar to the way the questionnaire would be administered clinically – that is, in a clinical setting, each patient would receive a single type of response format. By contrast, the limitations of working with "derived" 5-point format are that we cannot determine the variability in data quality due to respondent and the impact on the data quality during the actual administration of the collapsed response options [36-39]. On the other hand, an advantage of our study design was that we did not need to consider factors such as the influence of test–retest time on the results. Another strength specific to our own study is that we examined six, 5-point formats instead of

choosing just one, which is typically done when collapsing categories as a post-hoc analysis. The results from all possible “reasonable” scenarios increases the generalizability of our results.

We also acknowledge that the study findings are limited by the instrument (or dPROM) we chose to examine. Although our findings evidence the reliability and validity of the 5-point scale, more methodological work is needed to establish its suitability for other dPROMs. Also, we specifically compared the 5- and 11- scaling points as they are commonly used in clinical settings [36-37]. Additional research will be needed if researchers are interested in fewer than five response alternatives. We could have taken an exploratory approach and determined if some of the categories in the 11-point format could have been collapsed by examining whether certain CRCs were subsumed by adjacent CRCs. Instead, we adopted a confirmatory approach to specifically address the increasing application of the 11-point scale over the 5-point scale in clinical and research settings [40-41]. Dental practitioners and researchers already recognize the practical benefits of using the 5-point scale [11,40]. Our findings further assure them that using a more concise 5-point format does not compromise the scale reliability and the loss of information is limited and not clinically relevant. The robustness of our study findings is supported by the use of a large (N=2,078) sample of dental patients. A large sample size is required to obtain stable item parameter estimates. We also used both IRT and CTT methods, as each have their advantages and disadvantages. In general, previous studies suggest that different psychometric frameworks (e.g. IRT versus CTT) can produce discrepant findings [36]. We believe these two frameworks provided complementary information thus, adding to the strength of our study.

Significance of the study; Recommendations for research and practice

The 5-point scale clearly has several practical and technical advantages over the 11-point scale, making it easier to implement dPROs necessary for pursuing evidence-based dentistry across dental disciplines [42-43]. Firstly, fully labeled scales are more reliable than partially labeled scales [44]. The current 11-point scale provides label on the first and last category only. Secondly, when researchers employ an IRT framework to evaluate the precision of question responses, they would have fewer parameters to estimate with the 5-point response format compared to the 11-point response format. This would reduce the number of items and responses required to derive stable parameter estimates. Maydeu-Olivares et al. recommended that applied researchers use fewer response alternatives if they are concerned with the goodness of fit of their model and want to be confident that their latent trait estimates are highly reliable [36]. Lastly, the 11-point scale may overestimate precision of patients’ responses. Clinically, a 5-point response format is less burdensome and time-consuming for respondents [18], considering that there are limits to respondents’ capacity to process or discern a large number of response categories [44]. It is also easier for clinicians to administer the 5-point scale, especially when they are reading aloud the response categories to their patients who might need assistance with filling out surveys such as the elderly and those with low literacy level [11,45]. Such verbal clarification becomes more impractical with increasing number of response categories such as in the 11-point format [45]. Researchers might be inclined to use more number of response categories to maximize reliability (precision) [36], however, evidence shows that

patients are often reluctant to use all the scale points [38] resulting in response biases such as going for extreme or neutral responses.

Currently, there is no consensus on the most appropriate number of response categories for OES and our study offers promising evidence to support broader application of the 5-point scale. Additional research using a randomized or a repeated measures design will help account for any issues that might occur during the administration phase. Separate investigations are needed when comparing the 11- and 5-point scales for other dPROMs. Our analytical procedures offer guidance for conducting similar investigations for other dPROMs. Although further methodological work is needed, our study findings pave the way for standardization efforts with OES and possibly other dPROMs as well. Greater adoption of the 5-point format will help alleviate discrepancies in OES item response formats amongst dental providers and researchers and will make it easier for them to communicate their results.

Conclusion

To conclude, our study findings are highly encouraging for clinicians and researchers in the dental community who would like to use a 5-point format for responses to OES items. Our results showed high correlations between OES scores based on the 5-point response format and OES scores based on the 11-point response format, and the latent scores of the majority of the respondents were recovered well with all of the 5-point scales. From a psychometric point of view, OES scores based on an 11-point response format were equivalent to those based on a 5-point response format, hence, using the 5-point response format instead of the 11-point response format would have a negligible impact on OES score reliability and validity. The evidence we provide along with the practical and technical advantages of using a more concise 5-point format, alleviates any concerns that the psychometric properties of OES scores would be compromised by collapsing the 11-point response scale categories into 5.

Declarations

Ethical approval and consent to participate

This research was conducted in accordance with accepted ethical standards for human-patient research practice, undergoing review and approval by the Institutional Review Board of the HealthPartners Institute in Minneapolis, MN (registration A11-136). All the participants completed an informed consent form before their enrollment.

Consent for publication

Not applicable

Availability of data and materials

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

Competing interests

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

The National Institute of Dental and Craniofacial Research of the National Institutes of Health, USA, under Award Numbers R01DE022331 and R01DE028059, supported the study.

Authors' contributions

SP, MJ, SC, and SK contributed to the study design, data analysis, interpretation, and drafting the manuscript.

Acknowledgments

Not applicable

References

[1] John MT, Sekulić S, Bekes K, et al. Why patients visit dentists – A study in all WHO regions. *J Evid Based Dent Pract*. 2020;20(3):1-12.

[2] Isiekwe GI, Sofola OO, Onigbogi OO, Utomi IL, Sanu OO, daCosta OO. Dental esthetics and oral health-related quality of life in young adults. *Am J Orthod Dentofac Orthop*.2016;150(4):627–36.

[3] Larsson P, John MT, Nilner K, Bondemark L, List T. Development of an orofacial esthetic scale in prosthodontic patients. *Int J Prosthodont*. 2010;23(3):249–56.

[4] Simancas-Pallares M, John MT, Prodduturu S, Rush WA, Enstad CJ, Lenton P. Development, validity and reliability of the Orofacial Esthetic Scale - Spanish version. *J Prosthodont R*. 2018;62(4):456–61.

[5] Kerosuo H, Hausen H, Laine T, Shaw WC. (1995). The influence of incisal malocclusion on the social attractiveness of young adults in Finland. *Eur J Orthod*.1995;17(6):505-12.

[6] Mittal H, John MT, Sekulić S, Theis-Mahon N, Rener-Sitar K. Dental patient-reported outcome measures for adults: A systematic review. *J Evid Based Dent Pr*. 2019;1(19):53-70.

[7] Rener-Sitar K, John MT, Truong V, Tambe S, Theis-Mahon N. Nonmalignant oral disease-specific dental patient-reported outcome measures for adult patients: A systematic review. *J Evid Based Dent Pr*. 2021;1(21):1-21.

[8] Palaiologou A, Kotsakis GA. Dentist-patient communication of treatment outcomes in periodontal practice: A need for dental patient-reported outcomes. *J Evid Based Dent Pr*. 2020;20(2):101443.

[9] John MT. Health outcomes reported by dental patients. *J Evid Based Dent Pr*. 2018;18(4): 332–35.

[10] Listl S. Value-based oral health care: Moving forward with dental patient-reported outcomes. *J Evid Based Dent Pr*. 2019;19(3):255-59.

[11] Persic S, Milardovic S, Mehulic K, Celebic A. Psychometric properties of the Croatian version of the Orofacial Esthetic Scale and suggestions for modification. *Int J Prosthodont*. 2011;24(6):523–33.

[12] Zhao Y, He SL. Development of the Chinese version of the Orofacial Esthetic Scale. *J Oral Rehabil*.2013;40(9):670-77.

[13] Bimbashi V, Čelebić A, Staka G, Hoxha F, Peršić S, Petričević N. Psychometric properties of the Albanian version of the Orofacial Esthetic Scale: OES-ALB.*BMC Oral Health*. 2015;15(1):1-8.

[14] Reissmann DR, Benecke AW, Aarabi G, Sierwald I. Development and validation of the German version of the Orofacial Esthetic Scale. *Clinical Oral Investigations*.2015;19(6): 1443–50.

[15] Wetselaar P, Koutris M, Visscher CM, Larsson P, John MT, Lobbezoo F. Psychometric properties of the Dutch version of the Orofacial Esthetic Scale (OES-NL) in dental patients with and without self-reported tooth wear. *J Oral Rehabil*.2015;42(11):803–9.

[16] Reissmann DR, John MT, Enstad CJ, Lenton PA, Sierwald I. (2019). Measuring patients' orofacial appearance: Validity and reliability of the English-language Orofacial Esthetic Scale. *J Am Dent Assoc*.2019;150(4):278–86.

[17] Campos LA, Marôco J, John MT, Santos-Pinto A, Campos JADB. Development and psychometric properties of the Portuguese version of the Orofacial Esthetic Scale: OES-Pt. *PeerJ*.2020;8:e8814.

[18] Babakus E, Mangold WG. Adapting the SERVQUAL scale to hospital services: an empirical investigation. *Health Serv Res*. 1992;26(6):767–86.

[19] Gries K, Berry P, Harrington M, Crescioni M, Patel M, Rudell K, Safikhani S, Pease S, Vernon M. Literature review to assemble the evidence for response scales used in patient-reported outcome measures. *Journal of patient-reported outcomes*. 2018;2(41):1-14.

[20] John MT, Larsson P, Nilner K, Bandyopadhyay D, List T. Validation of the Orofacial Esthetic Scale in the general population. *Health Qual Life Outcomes*. 2012;10(135):1-7.

[21] Slade GD, Spencer AJ. Development and evaluation of the Oral Health Impact Profile. *Community Dent Health*. 1994;11(1):3-11.

[22] Locker D. Measuring oral health: A conceptual framework. *Community Dent Health*. 1988;5:3–18

[23] PROMIS Cooperative Group. (2013). *PROMIS® instrument development and validation scientific standards version 2.0*. *0*(May), 1–72. https://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf Accessed 28 Feb 2021.

[24] Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*.1951;16: 297-334.

[25] Yang FM, Kao ST. Item response theory for measurement validity. *Shanghai Arch Psychiatry*.2014;26(3):171–77.

[26] Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement.*1969;34:1-97.

[27] Embretson SE. *Item response theory for psychologists*. 1st ed. Psychology Press;2000.

[28] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st Century. *Medical Care*. 2000;38(9):28-42.

[29] Bock RD, Mislevy R J. Adaptive EAP estimation of ability in a microcomputer environment. *Appl Psychol Meas*. 1982;6(4):431–44.

[30] Chalmers RP. MIRT: A multidimensional item response theory package for the R environment. *J Stat Softw*. 2012;48(6):1–29.

[31] DeMars C. *Item response theory*. 1st ed. New York:Oxford University Press;2010.

[32] Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert-scale items? *Educational and Psychological Measurement*. 1972;31(3):657–74.

[33] Bendig AW. Reliability and the number of rating-scale categories. *J Appl Psychol*. 1954;38(1):38–40.

[34] Garner WR. Rating scales, discriminability, and information transmission. *Psychol Rev*. 1960;67(6):343–52.

[35] Symonds PM. On the loss of reliability in ratings due. *J Exp Psychol*.1924;7:456–61.

[36] Maydeu-Olivares A, Kramp U, García-Forero C, Gallardo-Pujol D, Coffman D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behav Res Methods*. 2009;41(2):295–308.

[37] Hendriks AAJ, Vrielink MR, van Es S, De Haes HJ, Smets EM. Assessing inpatients' satisfaction with hospital care: Should we prefer evaluation or satisfaction ratings? *Patient Educ Couns*. 2004;55(1):142–46.

[38] Garratt AM, Helgeland J, Gulbrandsen P. Five-point scales outperform 10-point scales in a randomized comparison of item scaling for the Patient Experiences Questionnaire. *J Clin Epidemiol*. 2011;64(2):200–207.

[39] Peršič S, Palac A, Bunjevac T, Čelebič A. Development of a new chewing function questionnaire for assessment of a self-perceived chewing function. *Community Dent Oral Epidemiol*. 2013;41(6):565–73.

[40] Leao A, Sheiham A. The development of a socio-dental measure of dental impacts on daily living. *Community Dent Health*. 1996;13(1):22–26.

[41] Krosnick JA, Berent MK. Comparisons of party identification and policy preferences: The impact of survey question format. *Am J Pol Sci*. 1993;37(3):941-64.

[42] Hua F. Increasing the value of orthodontic research through the use of dental patient-reported outcomes. *J Evid Based Dent Pract*. 2019;19(2):99-105.

[43] Reissmann DR. Dental patient-reported outcome measures are essential for evidence-based prosthetic dentistry. *J Evid Based Dent Pract*. 2019;19(1):1-6.

[44] Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol Rev*. 1956;63(2):81-97.

[45] Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *Int J Mark Res*. 2008;50(1):61–77.

Figures

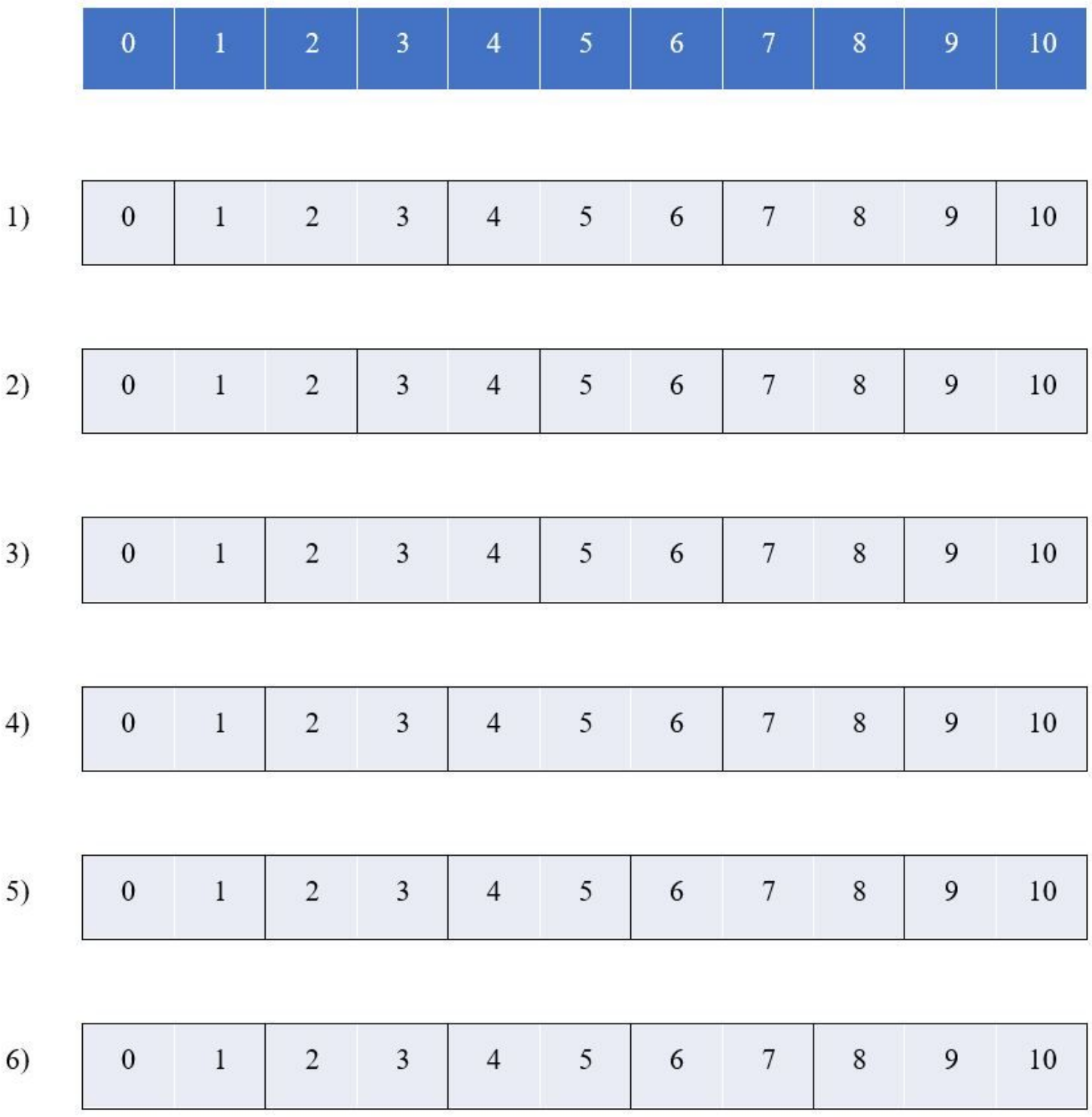

Figure 1

Six 5-point scale formats derived from the 11-point scale

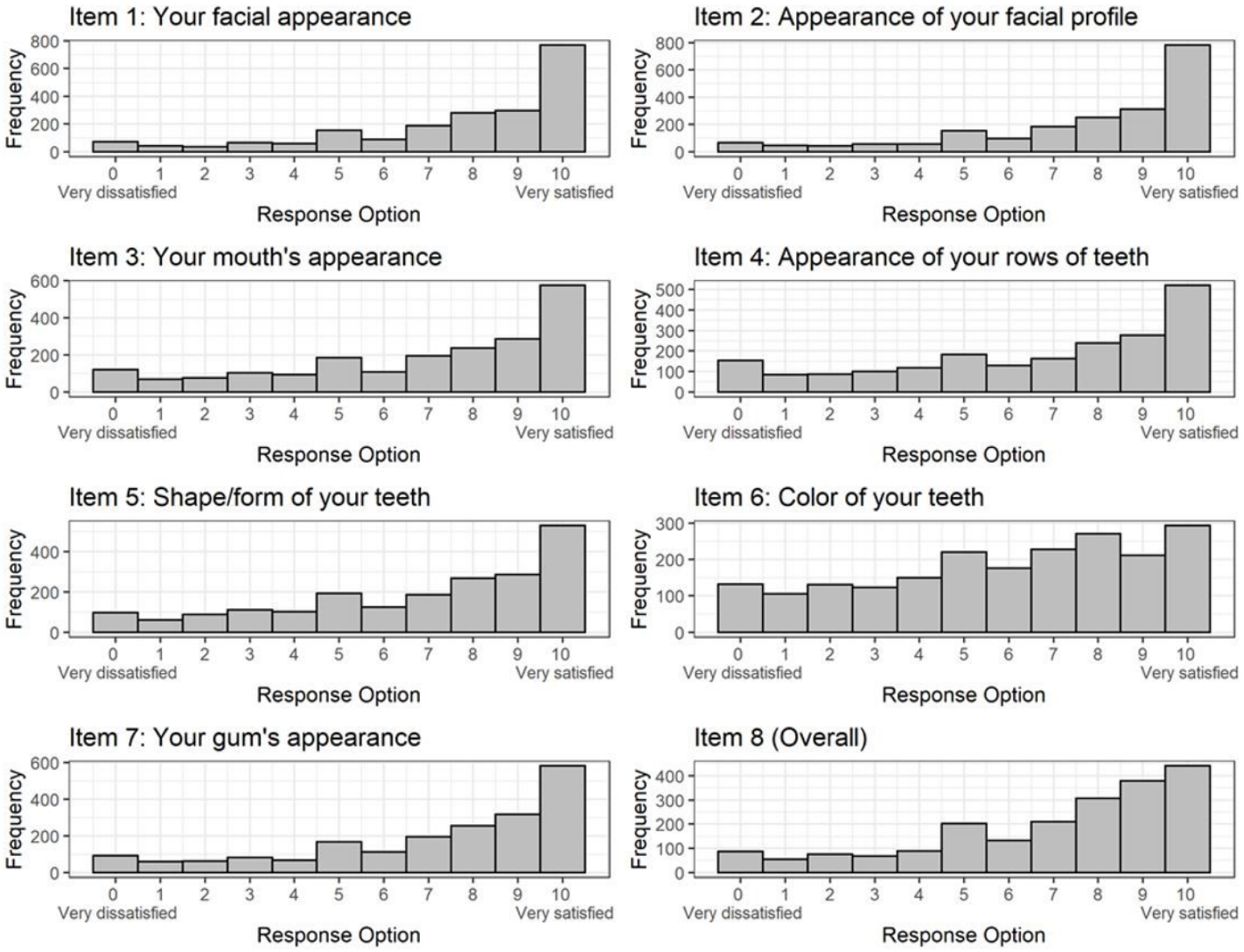

Figure 2

Histograms of Items 1-8 on a 11-point scale

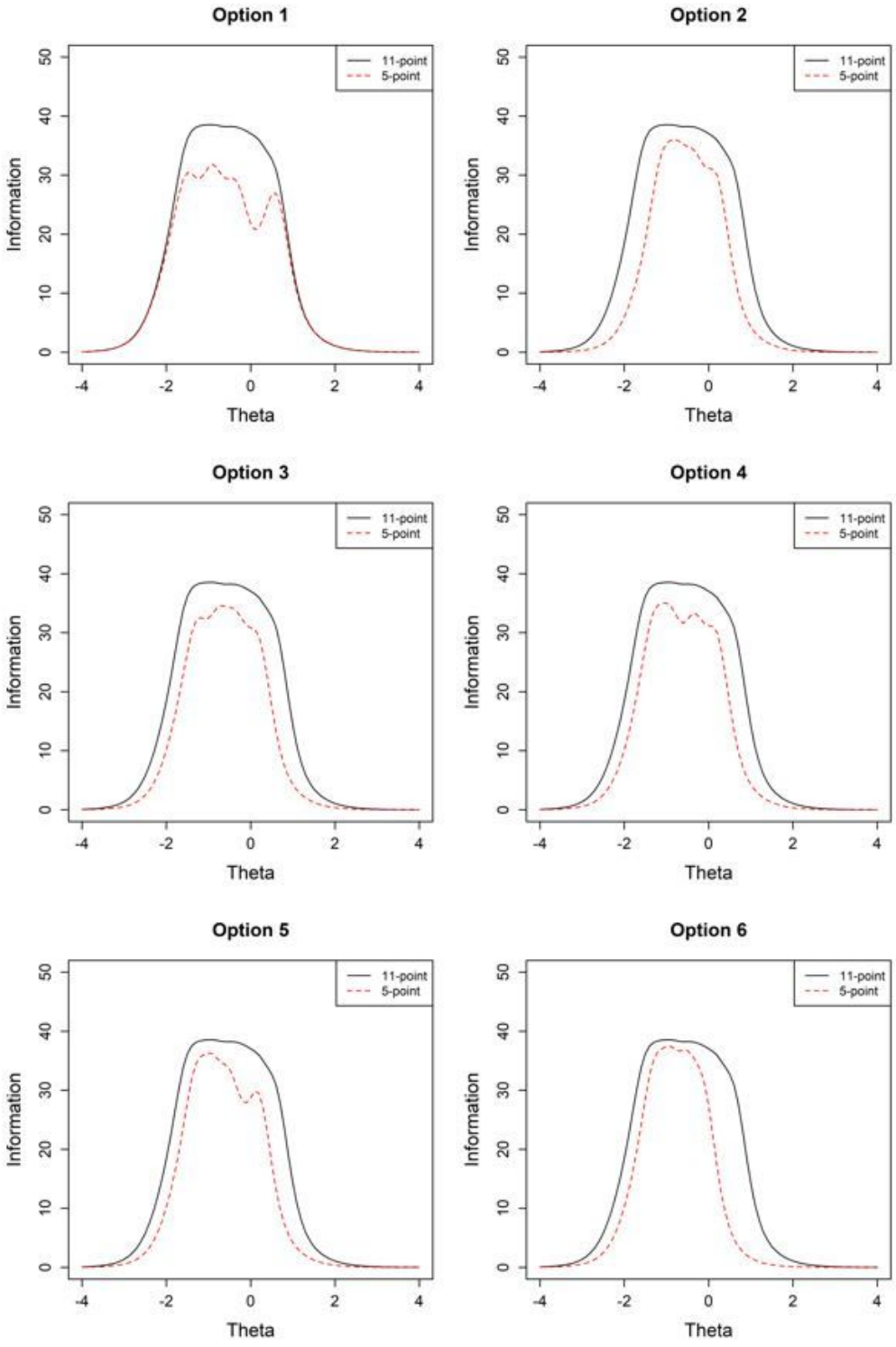

Figure 3

Test information function curves