Quality of Life Research (2019) 28:841–853
https://doi.org/10.1007/s11136-018-2064-4

To bin or not to bin? A comparison of symptom frequency response formats in the assessment of health-related quality of life

Brooke E. Magnus[1] · Mackenzie Kirkman[1] · Twinkle Dutta[1] · Manpreet Kaur[1] · Nichole Mannchen[1]

Accepted: 22 November 2018 / Published online: 27 November 2018

Abstract

Purpose The goal of this study is to compare three different types of retrospective frequency response formats on the Healthy Days Symptoms Module (HDSM). Responses are compared in terms of intra-individual consistency, psychometric value, and participant feedback about each type of response format.

Methods Respondents each completed three versions of the HDSM, where items were framed to elicit an open-ended frequency, a fixed choice frequency, or a vague quantifier response. Traditional reliability statistics were used to evaluate intra-individual consistency. Differential item functioning (DIF) was used to test for response format effects, and item response theory (IRT) scale scores and standard errors were computed across the three forms to compare psychometric value. Linear mixed modeling was used to examine the associations of IRT scale scores across response formats with respondent characteristics.

Results People are largely consistent in how they respond to items about their health, regardless of the response format, and no DIF was detected between response formats. The IRT scores computed from the "# of days" frequency response formats tend to have better measurement precision than those from vague quantifiers. Open-ended frequencies capture a greater span of individual differences for people reporting fewer symptoms; however, little measurement precision is lost in collapsing the frequencies into categories.

Conclusions Both the open-ended and fixed choice frequency response formats offer more measurement precision than vague quantifiers. While the open-ended frequency response format may capture more individual differences, respondents tend to report more difficulty with exact frequency recall, and thus, prefer the fixed choice frequency format.

Keywords Response format · Symptom frequency · Symptom recall · Count data

In an effort to track population health-related quality of life (HRQOL), the Centers for Disease Control and Prevention (CDC) developed the HRQOL-14 Healthy Days Measure [1]. The HRQOL-14 includes three subscales that assess various aspects of perceived physical and mental health functioning over a 30-day recall period. Of the 14 items that comprise the HRQOL-14, eight have an open-ended frequency response format, such that respondents report the exact number of days in the prior 30 that they have experienced a particular symptom or limitation. Five of these open-ended frequency items comprise the Healthy Days Symptoms Module (HDSM), which addresses five specific health-related symptoms: pain, depression, anxiety, sleeplessness, and vitality. The HDSM is routinely administered on CDC-sponsored questionnaires, including the Behavioral Risk Factor Surveillance System (BRFSS) and the National Health and Nutrition Examination Survey (NHANES) [2, 3].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-018-2064-4) contains supplementary material, which is available to authorized users.

✉ Brooke E. Magnus
brooke.magnus@marquette.edu

Mackenzie Kirkman
Mackenzie.kirkman@marquette.edu

Twinkle Dutta
twinkle.dutta@marquette.edu

Manpreet Kaur
manpreet.kaur@marquette.edu

Nichole Mannchen
nichole.mannchen@marquette.edu

[1] Department of Psychology, Marquette University, P.O. Box 1881, Milwaukee, WI 53201-1881, USA

Items eliciting open-ended frequencies are less commonly used on questionnaires than more general quantifiers (e.g., "Often"), but may offer some advantages in HRQOL research [1, 4]. An exact number of days provides a concrete measure of time that is easily understood by policymakers and, due to its ratio scale, may be more sensitive to short-term changes in health than other response formats [4]. Further, because frequencies are in non-relative units (e.g., 4 days) rather than subjective units (e.g., Rarely), they are more likely to be interpreted consistently across respondents, facilitating score comparison [5–7]. Schneider and Stone found that respondents with a chronic health condition assigned higher open-ended numeric values to the same vague quantifiers than respondents without a chronic health condition [5], likely due in part to the response shift phenomenon in which people with chronic illness positively adapt to their disease and report higher levels of HRQOL, despite having worse health [8, 9]. The objective nature of open-ended frequencies may make them less susceptible to response shift. Finally, the continuous response scale of open-ended frequencies may capture a wider range of individual differences in health than broader categories [1]. For these reasons, some researchers and policymakers have advocated for the use of the open-ended frequency response format on the HDSM [4].

Researchers have examined the reliability and validity of the HDSM from a classical test theory perspective [4, 10–13], finding acceptable levels of reliability ($\alpha = 0.75$–0.90) and moderate to strong associations with other measures of mental and physical functioning (e.g., SF-36) in populations with arthritis [12] and spinal cord injuries [10], as well as the general population [13]; however, the psychometric properties of the HDSM have rarely been studied within a latent variable [i.e., item response theory (IRT)] framework [14–16]. A primary advantage of the IRT framework is that measurement error is not assumed to be constant; rather, the precision with which an item or scale measures individual differences can vary across levels of the latent variable, making it possible to identify subsets of items that may be particularly useful for measurement at different levels of construct severity [17]. While the latent variable framework is appealing, due to the open-ended counts, IRT analyses of the HDSM pose several challenges. Most notably, people have a tendency to report values that are multiples of five, a phenomenon known as heaping [18]. For example, it is more common for people to report experiencing a symptom for 10 days than 9 or 11 days, consistent with the theory that some respondents use cognitive heuristics to estimate frequencies rather than counting over longer recall periods [5, 19–22]. Heaping makes the psychometric analysis of multivariate open-ended frequencies particularly challenging, as the responses do not tend to follow a standard count distribution that can be easily implemented in conventional latent variable software [16].

To circumvent these issues, researchers using IRT to evaluate the HDSM have typically either ignored the heaping [12, 23, 24] or collapsed the 0–30 open-ended frequencies into a smaller subset of responses after data collection [15], a practice commonly referred to as binning [25]. While binning eliminates heaping and offers statistical convenience, it may result in loss of the information that is provided by the raw frequencies—similar to the argument made against the dichotomization of a quantitative variable [26]. Binning also makes assumptions about the way people would respond to the item had it used a different response format—for example, that someone responding 9 days on the open-ended measure would have endorsed 6–10 days on a fixed choice version of the measure. To our knowledge, the assumptions that the response format on symptom frequency measures has no influence on (1) the response itself, and (2) the psychometric properties of the item—and thus, the resulting IRT scale score—have not been empirically tested. Responses that are binned after data collection may or may not be equivalent to those that would arise from a fixed choice response format.

To avoid binning, Magnus and Thissen developed a mixture IRT model for open-ended frequency data that also takes into account heaping [16]. The model includes a latent class of individuals who respond to the item according to a count process, as well as a latent class that responds to the item using an estimation method. The authors fit the mixture IRT model to HDSM data from the BRFSS and found that nearly one-third of respondents engaged in some type of estimation behavior, only endorsing multiple-of-five responses. When considered with previous research showing difficulties many participants experience with exact recall [5, 20, 27], these findings bring to light some of the challenges associated with using open-ended frequencies for retrospective assessment. The benefits, and costs, this type of response format may offer in terms of measuring individual differences remain unclear.

Additional research is needed to better understand how the format of response options may influence an individual's response to symptom-related items, as well as to what degree the open-ended frequency response format offers psychometric advantages over its more traditional counterparts. Specifically, how do people's responses and IRT scale scores compare when the item is framed to elicit an open-ended frequency, a fixed choice frequency, or a fixed choice vague quantifier? The primary goal of this study is to compare the performance of these response formats when administered to the same individuals: how consistent are responses across formats, and what measurement value is added by retaining the open-ended count response format rather than using a fixed choice format? Consistency and measurement value

are evaluated in terms of (1) traditional reliability statistics, (2) differential item functioning (DIF), and (3) IRT scale scores and precision. A secondary goal is to use respondent characteristics and feedback to better understand how individuals engage with each type of response format, and whether response format may moderate the relationships between respondent characteristics and IRT scale scores.

Method

Research participants

Data were collected from $N = 950$ respondents over the age of 18 using Amazon's Mechanical Turk (MTurk) online platform. Two hundred and forty-seven ($N = 247$) respondents were excluded from analyses based on evidence of not paying sufficient attention to the survey items, resulting in an analytic sample of $N = 703$. Details of participant exclusion criteria can be found in online Appendix A. Respondents were recruited without regard to age, gender, race, or health status; the only requirement was English proficiency. Volunteers were compensated with $0.75 for their participation, requiring approximately 10 min. IRB approval was obtained through the Office of Research Compliance at Marquette University. The mean respondent age was 37.23 years (SD = 11.89 years); other demographic characteristics of the analytic sample can be found in Table 1.

Table 1 Demographic characteristics of the analytic sample ($N = 703$)

	N (%)
Gender	
Female	316 (45.9)
Male	368 (53.5)
Prefer not to say	4 (0.58)
Education	
Less than a high school degree	3 (0.4)
High school graduate or GED	75 (11.3)
Some college, technical school, or associate degree	213 (31.0)
College degree or advanced degree	397 (57.7)
Race	
White/Caucasian	430 (62.5)
Black/African-American	43 (6.3)
Asian or Pacific Islander	167 (24.3)
Hispanic/Latino	23 (3.3)
Other	25 (3.6)
Global health rating	
Excellent	78 (11.3)
Very good	217 (31.4)
Good	253 (36.7)
Fair	118 (17.1)
Poor	24 (3.5)
Chronic health condition	
Yes	212 (30.7)
No	478 (69.3)

Respondents were free to skip any of the demographic items; thus, sample sizes do not always sum to $N = 703$

Study design

After clicking on the MTurk study link, participants were brought to a study information page, where they clicked on a box indicating they understood all study information, were at least 18 years old, and were willing to participate. Participants could also click on a box declining participation. If they agreed to take part, participants were redirected to a Qualtrics survey where they completed three versions of the HDSM, each using a different type of response format: open-ended frequency, fixed choice frequency, and fixed choice vague quantifier. The item stems and response formats are listed in the margins of Table 2. Each HDSM administration was separated by 4–5 unrelated filler questions to reduce the possibility of participants recalling their responses from a previous form. To minimize the potential influence of one type of response format on subsequent responses, the presentation order was counterbalanced, with participants randomly assigned to one of three order conditions. Items were presented on the screen one at a time; participants could not return to a previous item after submitting a response. Items designed to verify that respondents were paying attention were embedded throughout the survey; failure to correctly answer these items resulted in the participant's automatic exit from the survey, and the exclusion of their responses from analyses (see online Appendix A). After completing all study questionnaires, respondents were asked to answer some questions about demographics, their general health, and their experiences taking the surveys.

Statistical methods

Descriptive statistics, histograms, and boxplots were used to examine the response distribution of each item on all three forms. Because a main goal of the study was to understand respondent (in)consistency, we removed from the sample only those individuals failing the attention checks. To examine the degree to which self-reported open-ended frequencies increase monotonically with self-reported fixed choice frequencies, means and standard deviations of the open-ended frequencies were calculated within each category of the fixed choice frequency and vague quantifier response formats. The percentage of individuals reporting an open-ended frequency falling within the bounds ± 2 days of the corresponding fixed choice frequency categories was also computed (e.g., the percentage of people who reported

Table 2 Upper panel: mean (SD) open-ended number of days reported for each fixed choice frequency category and percentage of people reporting an open-ended frequency within the corresponding bounds ±2 days of the fixed choice frequency categories. Middle panel: mean (SD) open-ended number of days reported for each vague quantifier response. Lower panel: measures of intra-individual consistency between different types of response formats (weighted kappa if equal number of response categories, Spearman correlation if unequal number of response categories)

	During the past 30 days, for about how many days[a] did PAIN make it hard for you to do your usual activities?	During the past 30 days, for about how many days[a] have you felt SAD, BLUE, or DEPRESSED?	During the past 30 days, for about how many days[a] have you felt WORRIED, TENSE, or ANXIOUS?	During the past 30 days, for about how many days[a] have you felt you did NOT get ENOUGH REST or SLEEP?	During the past 30 days, for about how many days[a] have you felt VERY HEALTHY AND FULL OF ENERGY?
Fixed choice frequency					
0 Days	0.08 (0.37) 99%	0.22 (0.75) 96%	0.26 (0.95) 98%	0.64 (2.60) 94%	0.70 (3.99) 73%
1–5 Days	3.67 (2.86) 95%	3.77 (2.71) 96%	3.86 (2.43) 93%	4.50 (3.64) 91%	4.83 (3.34) 74%
6–10 Days	8.20 (4.23) 78%	8.53 (3.81) 84%	9.20 (4.04) 81%	9.05 (3.69) 84%	9.12 (4.08) 83%
11–15 Days	12.73 (4.93) 77%	12.90 (4.55) 80%	12.35 (4.44) 74%	13.80 (5.07) 78%	14.11 (3.81) 74%
16–20 Days	14.76 (7.93) 43%	17.24 (5.05) 79%	17.33 (5.83) 73%	17.95 (4.79) 82%	18.85 (4.00) 62%
21–25 Days	22.81 (4.02) 88%	22.72 (4.97) 88%	22.34 (4.45) 88%	23.08 (3.99) 89%	23.53 (4.02) 72%
26–30 Days	26.74 (2.66) 83%	29.16 (2.00) 97%	29.15 (2.53) 94%	27.75 (5.76) 94%	28.02 (3.95) 97%
Fixed choice vague quantifier					
Never	0.07 (0.47)	0.33 (2.68)	0.66 (1.49)	0.41 (1.14)	1.55 (6.17)
Rarely	3.08 (2.47)	3.48 (2.37)	3.26 (4.39)	4.45 (4.12)	5.26 (5.02)
Sometimes	7.52 (5.26)	8.15 (4.93)	7.26 (6.13)	9.56 (4.81)	11.75 (4.64)
Often	16.29 (8.60)	17.56 (7.39)	13.20 (7.94)	18.17 (7.13)	20.46 (6.22)
Always	24.22 (7.47)	23.65 (9.56)	22.80 (8.38)	25.52 (7.46)	24.78 (7.08)
Weighted kappa (binned open-ended with fixed choice)	0.87	0.92	0.91	0.88	0.92
Spearman correlation (binned open-ended with vague quantifier)	0.82	0.84	0.81	0.79	0.80
Spearman correlation (fixed choice with vague quantifier)	0.81	0.85	0.83	0.80	0.80

[a]For the vague quantifier response format, the words "how many days" were replaced with "how often"

feeling a symptom for 6–10 days who also provided an open-ended frequency between 4 and 12 days). The open-ended frequencies were collapsed to match the fixed choice frequency categories (e.g., 1, 2, 3, 4, and 5 days were collapsed to 1–5 days), resulting in two sets of items with seven response categories: fixed choice and binned frequencies. Weighted kappa was used as a measure of intra-rater reliability between these two types of response formats. Spearman correlations were used to determine the intra-rater reliability between (1) fixed choice frequencies and vague quantifiers, and (2) post-hoc binned frequencies and vague quantifiers. All descriptive analyses were conducted in R [28].

Graded response models (GRMs, described in online Appendix B) were fit to the data from measures comprising fixed choice frequencies and vague quantifiers [29]. For the open-ended frequency data, we estimated a mixture IRT model that accounts for a class of individuals who respond according to a count process, as well as a class of individuals who use an estimation method [16]. Details and R code for estimating this model can be found in online Appendix C. For all items except the one about energy, a higher frequency response suggests worse HRQOL; thus, for IRT analyses, the energy item was reverse coded to reflect a lack of energy. After calibration, response pattern-based IRT scale scores and standard errors were estimated from all three sets of responses. The mixture IRT model was estimated in R [28]. All other IRT analyses were done in IRTPRO [30].

DIF analysis allows one to examine whether an item has different psychometric characteristics across groups (e.g., those administered open-ended frequency items vs. those administered fixed choice frequency items), after accounting for underlying differences in people's levels of the latent variable [31]. DIF analyses were conducted to determine whether the "# of days" frequency items exhibit different psychometric properties, and thus, yield different scale scores, depending on whether the question was framed to elicit open-ended frequencies that were subsequently collapsed into categories, or whether the question was initially framed with fixed choice frequencies. From the full sample, two independent subsamples were formed based on form administration order. One group comprised the 217 respondents who answered the open-ended frequency items first; the other group comprised the 258 respondents who answered the fixed choice frequency items first. The remaining 228 people who were administered the vague quantifiers first were not included in the DIF analyses. Wald tests were conducted to simultaneously test each item for DIF, treating all items as anchor items.

To compare the precision of IRT scores computed from each type of response format, and to evaluate any potential loss of information that occurs as the result of binning or using vague quantifiers in place of frequencies, we plotted standard errors as a function of IRT scale scores for three sets of scores: (1) open-ended frequencies, (2) fixed choice frequencies, and (3) vague quantifiers.

At the conclusion of all questionnaires, participants were asked to report their recall difficulty and whether they used a counting or an estimation method to arrive at a numeric value. They were also asked to indicate whether they had a preference for one of the response formats that they had encountered, and if so, why. The proportion of individuals preferring each type of response format was calculated, and from the written responses, the first two authors identified common themes in the reasons people provided for preferring a specific type of response format. Themes representing a preference for open-ended frequencies included liking the precision of an exact count and the ability to provide what they considered a more accurate response. Themes representing a preference for fixed choice frequencies included the ease of responding compared to recalling an exact count and the fixed choice frequency offering a good compromise between the exact count frequency, which is too specific and prone to recall error, and the vague quantifier, which is too broad and may carry different meanings across respondents. Themes representing a preference for vague quantifiers included the ease of responding compared to recalling a numeric frequency and the natural tendency for many people to quantify their symptoms more generally rather than on a numeric scale. Two independent raters then classified each written response into these themes. The first author then identified specific responses that were most clearly representative of those themes (i.e., the words included in the responses matched the theme labels). Finally, we used a linear mixed model with subjects as a random effect to examine the associations of IRT scale scores with respondent and form characteristics, including response format, order of form administration, response format preference, whether the respondent used an estimation or counting strategy for frequency recall, and the amount of difficulty the respondent had with frequency recall. Interactions of respondent characteristics with response format were included in the model to examine whether response format moderates any of the relationships between respondent characteristics and IRT scale scores. Because the follow-up questions were specific to items eliciting a numeric frequency, scale scores based on vague quantifier responses were not included in the model. The Benjamini–Hochberg procedure was used to maintain a false discovery rate of 0.05 [32].

Results

Descriptive statistics and intra-individual consistency

Item-level distributions for all three types of response formats are shown in Fig. 1. Descriptive statistics and measures of intra-rater reliability are displayed in Table 2. Several item characteristics are noteworthy. First, for the open-ended frequency response format, heaping at multiples of five is present in all items (Fig. 1). Second, the mean and median open-ended frequency increases across successively higher levels of the fixed choice response options. This is true for both the fixed choice frequencies and the vague quantifiers. Third, the mean open-ended frequencies generally fall within the bounds of the corresponding fixed choice frequencies, with the exception of the 16–20 days response option for the item about pain. Fourth, according to aggregate measures of agreement, people show a high degree of intra-individual consistency in responding to open-ended frequency, fixed choice frequency, and vague quantifier response formats, with agreement statistics ranging from 0.79 to 0.92; however, in examining specific response categories (e.g., 0 days, 1–5 days, etc.), percent agreement tends to dip for moderate to severe symptoms (11–25 days) and then rise for the most severe response category (26–30 days). For the four items where smaller frequencies reflect better HRQOL, percent agreement is over 90% through 1–5 days, drops to around 70–90% for the moderate to severe responses, and typically returns to over 90% for 26–30 days. Agreement tends to be lowest for the energy item, likely due to the reversal of the wording from negative to positive. Taken together, these results suggest that people are overall consistent in how they respond to the item, whether the item is framed as an open-ended frequency, a fixed choice frequency, or a vague quantifier, and that responses tend to be more consistent at higher (better) levels of HRQOL.

The results also point to the subjective nature of the vague quantifier, especially for the "Sometimes" and "Often" categories. The spread of the open-ended frequencies across different levels of the fixed choice response categories is more easily seen in the boxplots in Fig. 2. While the median

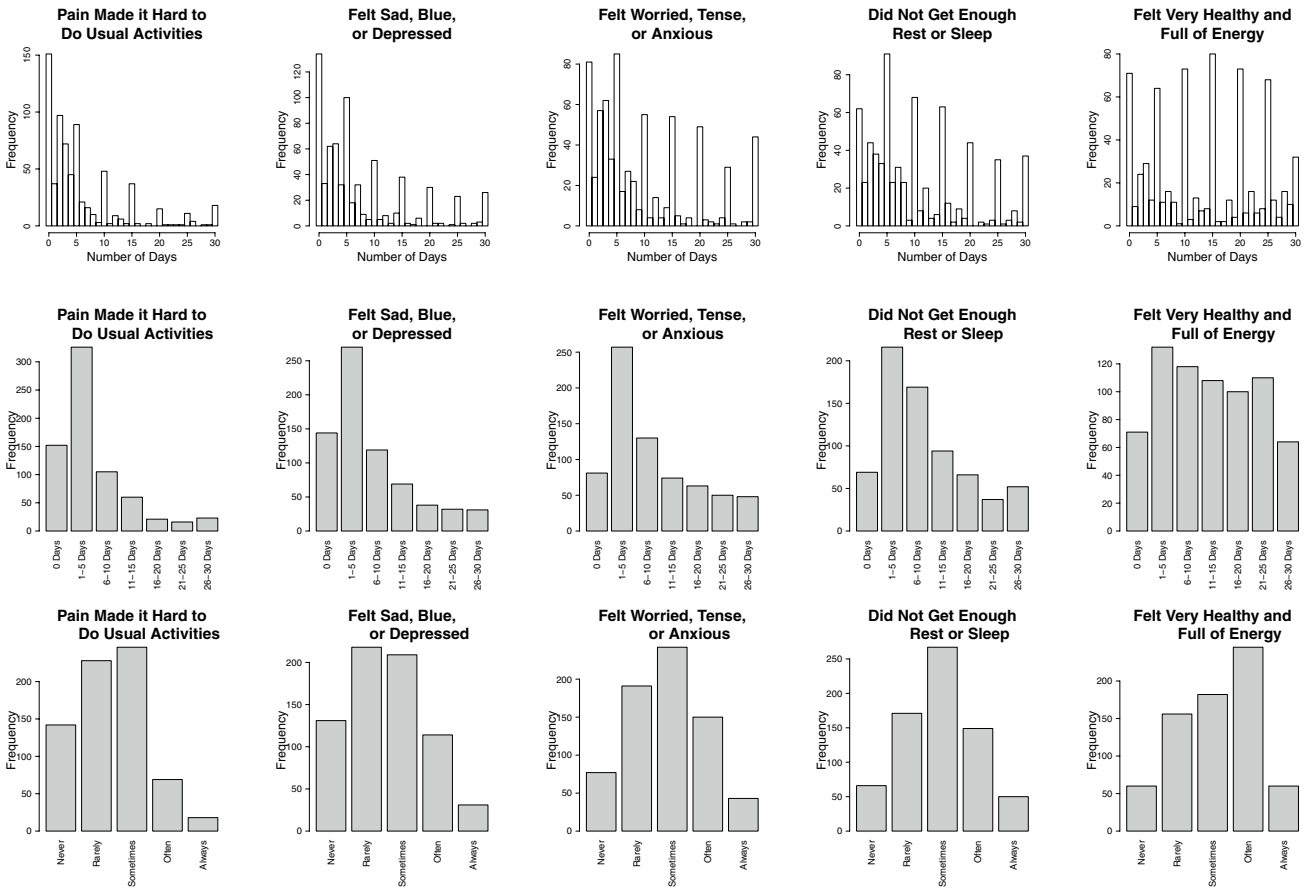

Fig. 1 Item response distributions across different response formats. Upper panel: open-ended frequency. Middle panel: fixed choice frequency. Lower panel: vague quantifier

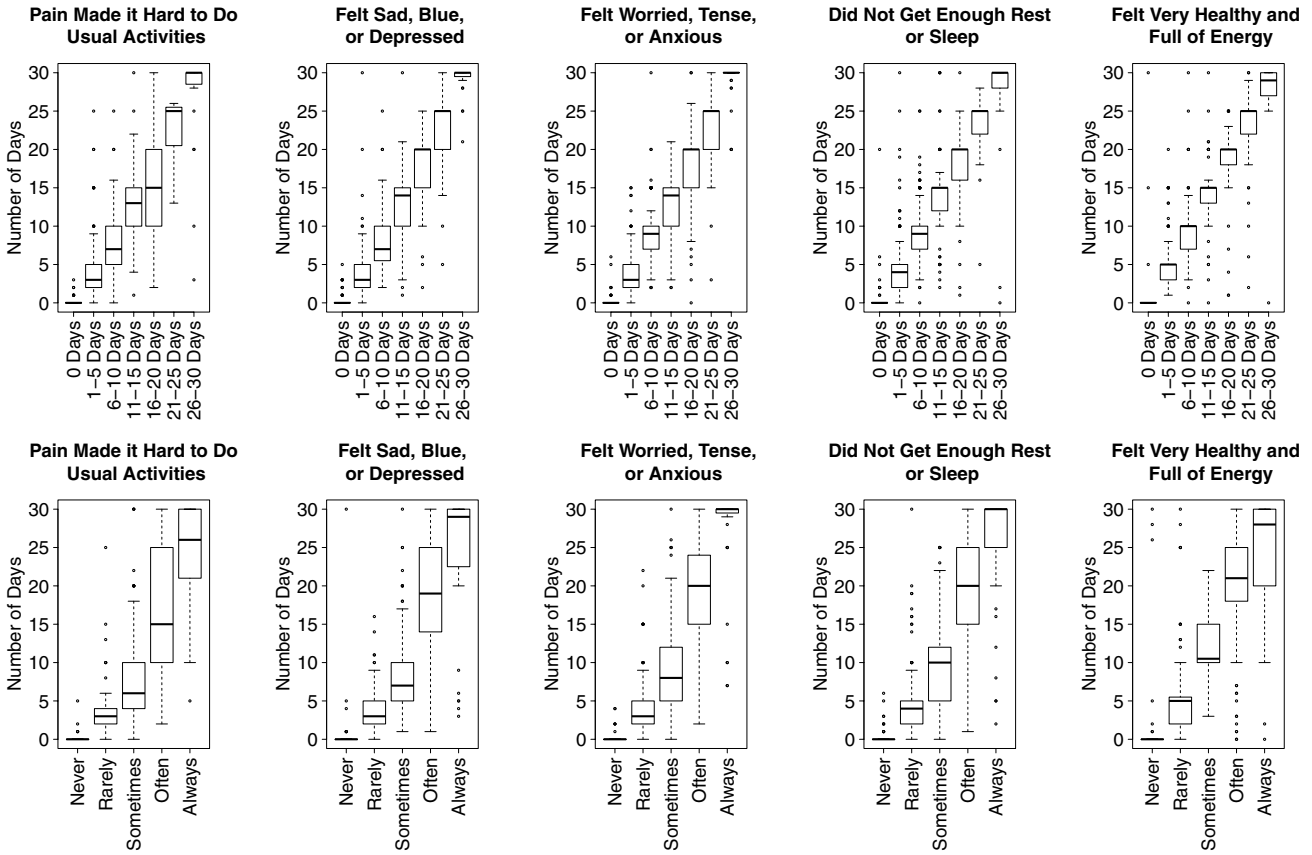

Fig. 2 Boxplots of open-ended response frequencies as a function of fixed choice frequency category (upper panel) and vague quantifier category (lower panel)

open-ended frequency increases over the fixed choice response options, there is substantial variability in these frequencies, particularly for the categories associated with more frequent symptoms.

Item parameter estimates and differential item functioning

Item parameter estimates and standard errors for the three sets of item responses are shown in Table 3. Values of the RMSEA ranged from 0.03 to 0.04, indicating that the HDSM is sufficiently unidimensional. For all response formats, the items with the highest discrimination parameters (a), and thus the items that can best differentiate among individuals at varying levels of HRQOL, include those about depression and anxiety; the items about pain, rest, and energy exhibit noticeably weaker discrimination values, suggesting that the HDSM primarily measures mental health, with physical health being secondary. The large range of threshold values (b) suggests that the scale is able to measure individual differences across many levels of HRQOL. To formally evaluate the assumption that the binning of open-ended frequencies into categories is equivalent to fixed choice responses, DIF analyses were conducted using two independent samples of open-ended frequencies after binning and fixed choice frequencies; results are shown in Table 4. None of the items exhibit significant discrimination or threshold DIF. Thus, there is no evidence that framing the question as an open-ended frequency and then collapsing the frequencies into categories results in different parameter estimates than framing the question with fixed choice frequency categories; the resulting scale scores should be equivalent.

IRT scale scores

To compare (1) the measurement range of the scale, and (2) the measurement precision across the three response formats, IRT scale scores and their associated standard errors were computed for open-ended frequencies, fixed choice frequencies, and vague quantifiers, shown in Fig. 3. The observed lower bound of the scale scores is slightly expanded when scores are computed from the preserved open-ended frequencies ($\theta = -2.60$) compared to either of the fixed choice responses ($\theta = -2.22$ for fixed choice frequencies, $\theta = -2.27$ for vague quantifiers), suggesting that the open-ended frequency response format is able to capture

Table 3 Item parameter estimates (SE) for open-ended frequency (negative binomial mixture IRT model), fixed choice frequency (GRM—7 categories), and vague quantifier (GRM—5 categories) responses

	Pain			Depressed			Anxious			Rest			Energy		
	O-E	FC	VQ	O-E	FC	VQ	O-E	FC	VQ	O-E	FC	VQ	O-E	FC	VQ
a	0.65 (0.06)	1.10 (0.10)	1.08 (0.10)	1.15 (0.06)	3.61 (0.34)	3.33 (0.34)	1.05 (0.05)	3.02 (0.23)	2.79 (0.24)	0.68 (0.05)	1.40 (0.11)	1.21 (0.10)	0.46 (0.04)	1.56 (0.11)	1.16 (0.10)
b_1	− 2.34 (0.23)	− 1.44 (0.14)	− 1.55 (0.15)	− 1.30 (0.09)	− 0.91 (0.06)	− 1.00 (0.06)	− 1.76 (0.11)	− 1.39 (0.08)	− 1.46 (0.08)	− 3.54 (0.28)	− 2.08 (0.16)	− 2.31 (0.18)	− 6.00 (0.48)	− 2.00 (0.14)	− 2.49 (0.20)
b_2	–	0.83 (0.10)	0.12 (0.08)	–	0.25 (0.05)	0.00 (0.05)	–	− 0.04 (0.05)	− 0.34 (0.05)	–	− 0.36 (0.08)	− 0.71 (0.09)	–	− 0.96 (0.09)	− 0.31 (0.08)
b_3	–	1.72 (0.15)	2.14 (0.19)	–	0.77 (0.06)	0.93 (0.06)	–	0.51 (0.05)	0.70 (0.06)	–	0.58 (0.08)	0.97 (0.10)	–	− 0.36 (0.07)	0.88 (0.10)
b_4	–	2.54 (0.22)	3.82 (0.36)	–	1.16 (0.07)	1.92 (0.11)	–	0.86 (0.06)	1.81 (0.10)	–	1.19 (0.10)	2.57 (0.21)	–	0.20 (0.07)	2.47 (0.20)
b_5	–	3.00 (0.26)	–	–	1.47 (0.08)	–	–	1.25 (0.07)	–	–	1.79 (0.13)	–	–	0.84 (0.08)	–
b_6	–	3.54 (0.32)	–	–	1.89 (0.10)	–	–	1.71 (0.09)	–	–	2.31 (0.17)	–	–	1.90 (0.13)	–

$N = 703$ for all calibrations

IRT item response theory, *GRM* graded response model, *O-E* open-ended frequency, *FC* fixed choice frequency (RMSEA = .04), *VQ* vague quantifier (RMSEA = .03). For the open-ended frequencies, only the negative binomial IRT parameters are shown for the mixture IRT model

Table 4 Differential item function (DIF) statistics comparing fixed choice frequencies ($N=217$) versus open-ended frequencies after binning ($N=258$)

Item	Total χ^2 (df)	p-value	$\chi^2_{discrim}$ (df)	p-value	$\chi^2_{threshold}$ (df)	p-value
Pain	6.4 (7)	0.50	0.5 (1)	0.50	5.9 (6)	0.44
Depression	9.8 (7)	0.12	0.0 (1)	0.83	9.8 (6)	0.14
Anxiety	5.6 (7)	0.59	0.6 (1)	0.43	5.0 (6)	0.55
Rest	9.7 (7)	0.21	0.1 (1)	0.77	9.6 (6)	0.14
Energy	6.3 (7)	0.51	0.5 (1)	0.33	5.3 (6)	0.51

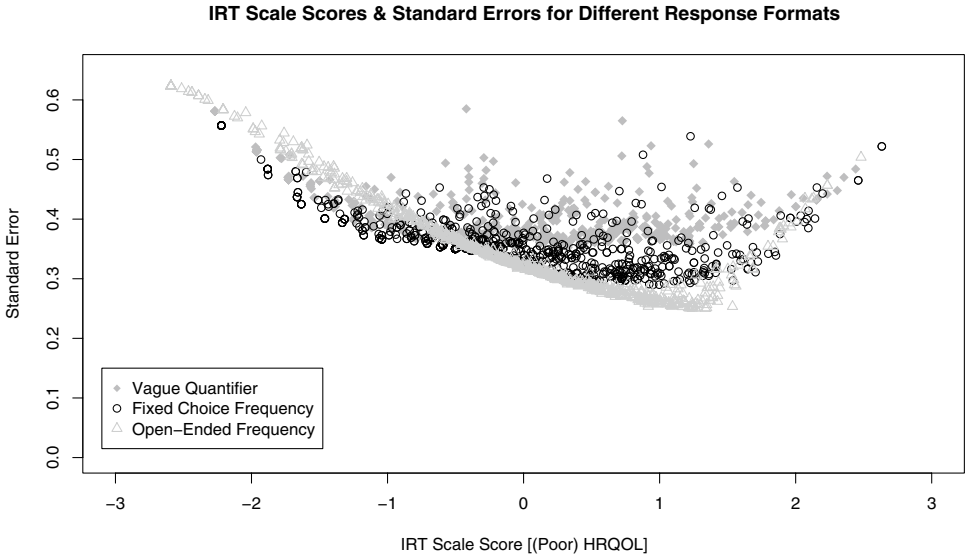

Fig. 3 Standard errors as a function of IRT scale scores for open-ended frequency (open triangle), fixed choice frequency (open circle), and vague quantifier (filled diamond) response formats

a larger span of individual differences among people reporting fewer symptoms; however, the associated measurement precision is noticeably worse for these same individuals when the open-ended response format is used. This can be seen in Fig. 3, where for $\theta<-1$, the standard errors associated with the open-ended response format are larger than for either of the other two response formats. Items eliciting a frequency, whether open-ended or fixed choice, almost always yield smaller standard errors than items that elicit a vague quantifier, particularly in the mid-to-upper range of the scale scores where standard errors are reduced by as much as 15%. Importantly, this end of the distribution represents individuals with more severe symptoms and thus more likely to endorse the "Sometimes," "Often," and "Always" response categories. At the lower end of the scale score distribution ($\theta<0$, where respondents experience milder symptoms and are more likely to endorse "Never" and "Rarely"), the standard errors of the vague quantifier scale scores are comparable to those obtained from the other types of response formats. Thus, it is only at higher levels of the latent variable that information is sacrificed using vague rather than more specific frequency quantifiers.

In comparing the open-ended and fixed choice frequency response options, there is a tradeoff in score precision between lower and upper ends of the latent variable distribution. For individuals exhibiting moderately severe symptoms ($1<\theta<2$), the open-ended frequency response format yields very slightly smaller standard errors than the fixed choice frequencies, and thus, measures individual differences with more precision. The cost for this gain in precision is considerably larger standard errors for the open-ended response format for individuals experiencing only mild symptoms ($\theta<-1$).

Respondent characteristics and preferences

Table 5 shows the proportions of individuals preferring each type of response format, as well as some representative explanations for those preferences. Nearly half of the sample (41%) preferred the fixed choice response format, followed by approximately one quarter of the sample (25.8%) preferring vague quantifiers. Table 6 displays estimates of the fixed effects from the linear mixed model used to examine the relationships between respondent characteristics, response format, and IRT scale scores. After controlling for age and gender, scale scores increase with successively worse ratings of overall health ($\chi^2(3)=205.14$, $p<.001$). There is an overall effect of recall difficulty ($\chi^2(2)=23.64$, $p<.001$); those who reported having some ($\beta=.27$, $p<.001$) or a great deal ($\beta=.54$, $p=.003$) of difficulty with numeric frequency recall

Table 5 Participant response format preferences and representative sample explanations

Preferred response format		
Open-ended frequency	Fixed choice frequency	Vague quantifier
$N = 123$ (17.8%)	$N = 283$ (41.0%)	$N = 178$ (25.8%)
Representative explanations for preference		
"Within 30 days we can exactly predict how many days we don't get sleep, or something. So I think [it's] always better to give exact answers."	"I feel that it gives me the ideal balance between numeric accuracy and the possibility to give approximations."	"I found it much easier to choose an option rather than come up with a specific number."
"I prefer to have a large range of response options."	"[Compared to a numeric frequency] 'Rarely' can mean a lot of things."	"I don't usually quantify the frequency of my symptoms."
"I like to make my choices in an open ended response to show how I truly feel."	"I think it helped me put things into perspective better, but it wasn't too vague."	"It's easier to recall and interpret my symptoms in terms of never, sometimes, often, rather than numeric values, because I'm just estimating."

$N = 106$ respondents (15.4%) had no response format preference

also tend to have worse HRQOL compared to those who report having no difficulty. For the open-ended frequency response format, people who reported using an estimation strategy for frequency recall show significantly worse HRQOL than those who indicated that they used a counting method ($\beta = .21$, $p = .04$); however, this effect is moderated by response format, such that the difference in HRQOL between those who use a counting versus estimation strategy is smaller (and non-significant) when the fixed choice frequency response format is used ($\beta = .12$, $p = .11$). In summary, respondents reporting greater symptom frequencies tend to have worse overall ratings of health, report greater difficulty with symptom recall, and rely on estimation methods rather than counting in recalling their symptom frequencies—particularly for the open-ended frequency format.

Discussion

The goals of this study were to examine the potential effects of different response formats on how people answer questions about their HRQOL, as well as to what extent the open-ended frequency response format offers psychometric advantages over more traditional response scales. Results suggest that individuals are overall consistent in reporting their numeric symptom frequencies when the item is framed either as (1) an open-ended frequency, or (2) a fixed choice frequency. The open-ended count responses tend to map onto the corresponding fixed choice responses, and intra-rater reliability is high. While there is a high degree of consistency across response formats, there is substantial variability in the open-ended frequencies that people provide within each of the vague quantifier categories. This is particularly true for the vague quantifiers that tend to be associated with moderate to severe symptoms ("Sometimes" and "Often"). These findings are consistent with prior research suggesting that people may interpret vague quantifiers differently based on individual differences [5].

The item discrimination parameters across all three types of response formats suggest that the HDSM primarily measures HRQOL as it pertains to mental health, with physical health being a secondary construct. While we chose to treat the HDSM as "unidimensional enough," the pattern of item discrimination parameters also supports prior research that has split the measure into mental and physical health subscales [15]. The DIF results offer evidence that binning open-ended frequencies after data collection does not yield appreciable differences in item parameter estimates compared to initially framing the item with fixed choice frequencies. This finding is important because it supports the implicit assumption that binning does not alter the psychometric properties of the items, and consequently, that the IRT scale scores computed from the item responses do not depend on the initial response format. Thus, the results of this study provide justification for the practice of binning open-ended frequencies into a smaller number of response categories after data collection (e.g., [15, 25]).

The scores from the fixed choice frequencies nearly always exhibit better measurement precision than those from vague quantifier responses, and more often than not, the open-ended frequency responses as well. While open-ended frequencies are able to capture a broader span of individual differences at the healthier end of the latent variable continuum, relatively little score precision is lost in using the fixed choice frequencies. The location on the latent variable where open-ended frequencies provide the greatest improvement in measurement precision is in the moderate to severe range, which represents individuals suffering from higher symptom frequencies. While this response format offers slight improvement in measurement precision

Table 6 Fixed effects from a linear mixed model predicting IRT scale scores (based on open-ended and fixed choice frequencies) from form and respondent characteristics

Respondent or form characteristic	Fixed effects	95% CI
General health		
Excellent (reference)	–	–
Very good	0.23**	[0.03, 0.43]
Good	0.77***	[0.57, 0.96]
Fair	1.24***	[1.01, 1.46]
Poor	1.80***	[1.45, 2.15]
Age	−0.01***	[−0.02, −0.01]
Gender		
Male (reference)	–	
Female	0.22**	[0.10, 0.34]
Form type		
Open-ended frequency (reference)	–	
Fixed choice frequency	0.06	[0.00, 0.12]
Form order		
First (reference)	–	
Second	0.04	[0.00, 0.07]
Third	0.05	[0.01, 0.09]
Difficulty with frequency recall		
No difficulty (reference)	–	
Some difficulty	0.27**	[0.14, 0.39]
A great deal of difficulty	0.54**	[0.19, 0.89]
Counting versus estimation method		
Counting (reference)	–	
Estimation	0.21*	[0.07, 0.36]
Response format preference		
Open-ended frequency (reference)	–	
Fixed choice frequency	−0.04	[−0.20, 0.12]
Vague quantifier	0.04	[−0.13, 0.22]
No preference	−0.18	[−0.38, 0.02]
Form type × counting versus estimation method		
Fixed choice frequency × counting (reference)	–	
Fixed choice frequency × estimation	−0.10*	[−0.16, 0.03]

Due to the volume of possible interactions, non-significant interactions are not shown

IRT item response theory

*$p < .05$; **$p < .01$; ***$p < .001$ after applying the Benjamini–Hochberg correction; higher IRT scale scores indicate worse HRQOL

for these individuals, those experiencing worse HRQOL also expressed having more difficulty with symptom recall and reliance on estimation heuristics, particularly with the open-ended frequency response format, calling into question the accuracy of these exact responses. This finding is not unexpected, as previous research suggests that retrospective recall may be distorted by memory bias and differences in estimation strategies that people use to recall exact frequencies [19, 20].

For these reasons, the use of open-ended frequencies over a 30-day recall period may be a less reliable method of symptom assessment compared to fixed choice response formats, especially for individuals who suffer from poor HRQOL. Open-ended frequencies may be able to capture a wider span of individual differences, and sometimes with more precision, but the accuracy of these measured individual differences is unclear. Further, when the open-ended frequency response format is used, HRQOL scores differ depending on recall strategy, such that people with worse HRQOL are more likely to use estimation strategies. The same is not true of the fixed choice frequency response format.

Qualitative respondent feedback can also offer insight into the response process: Due to the higher cognitive demand required of the open-ended frequency response format, many people prefer the less memory taxing alternatives—in

particular, the fixed choice frequency option, which respondents tend to describe as a desirable balance between an exact count and an overly vague categorization. Taking the present findings into consideration, for symptom recall over a 30-day period, we recommend using either (1) the fixed choice frequency response format, or (2) collapsing open-ended count data into categories after data collection. Rather than requiring a mixture model to accommodate the heaping that results from retrospectively reported open-ended counts, the fixed choice frequencies can be readily analyzed using conventional IRT models and software, and minimal measurement precision is lost in using this response format over its open-ended counterpart. Further, the fixed choice frequency response format offers greater measurement precision than vague quantifiers, particularly for people experiencing moderate to severe symptoms. These are the same individuals who report greater difficulty with symptom recall and reliance on estimation rather than counting. Thus, the open-ended frequency response format may place unnecessary burden primarily on people suffering from poor HRQOL, providing additional support for using fixed choice frequencies.

It is important to note that our recommendation is specific to the 30-day recall period and may not generalize to shorter recall periods. Prior research has found that respondents are more likely to rely on cognitive heuristics and show memory bias when tallying events over longer recall periods [33, 34]; heaping in open-ended frequency data is likely an artifact of using such cognitive heuristics. With shorter recall periods (e.g., 7 days), respondents may be less inclined to engage in estimation shortcuts, making heaping and recall bias less prevalent in open-ended frequency responses. If heaping and recall bias are reduced with shorter recall periods, the fixed choice frequency response format may no longer offer advantages over its open-ended counterpart.

This study is not without limitations. Due to the repeated measures design of this study, there is a potential for carryover effects. Thus, it is possible that after the first administration, participants were more attuned to their symptom frequencies on subsequent administrations, inflating intra-individual consistency. While the filler questions were intended to minimize such carryover effects, ideally, future studies could allow more time to elapse between administrations (but not enough for HRQOL to have changed). Further, while the form order was randomized here, future studies may also consider randomizing the order of items within each form, which was not done in the present study. Future research should also seek to replicate these findings with a more nationally representative sample than what can typically be obtained through MTurk. A further limitation is that all assessment methods considered here rely on retrospective self-report. Thus, it is not possible to know someone's "true" symptom frequencies over the last 30 days, and there is no gold standard to which to compare each of these response formats. Future research could examine the intra-individual consistency of retrospective recall with tabulated end-of-day assessments, as the daily diary approach may be a more reliable method of symptom assessment [20]. Finally, as previously discussed, the present study considered only a 30-day recall period, so the results and recommendations generalize only to similar types of items and recall periods. Future studies could compare frequency response formats when used with shorter time frames (e.g., 7 days, 14 days). Because the current results suggest that open-ended frequencies can capture a wider range of individual differences among people who report fewer symptoms, it is plausible that open-ended frequencies may be more useful for shorter recall periods during which fewer symptoms tend to be reported. This remains an important area for future HRQOL measurement research.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval 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.

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

References

1. Centers for Disease Control and Prevention (CDC). (2000). *Measuring healthy days*. Atlanta, GA: CDC.
2. Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). (1999). *National Health and Nutrition Examination Survey Questionnaire*. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
3. Centers for Disease Control and Prevention (CDC). (1984). *Behavioral Risk Factor Surveillance System Survey Questionnaire*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
4. Moriarty, D. G., Zack, M. M., & Kobau, R. (2003). The Centers for Disease Control and Prevention's Healthy Days Measures—Population tracking of perceived physical and mental health over time. *Health and Quality of Life Outcomes,1*, 37.
5. Schneider, S., & Stone, A. A. (2016). The meaning of vaguely quantified frequency response options on a quality of life scale depends on respondents' medical status and age. *Quality of Life Research, 25*(10), 2511–2521.
6. Hakel, M. D. (1968). How often is often? *American Psychologist, 27*(3), 533–534.
7. Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported

behavior and comparative judgments. *Public Opinion Quarterly, 49*, 388–395.
8. Ahmed, S., Mayo, N. E., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R. (2005). Change in quality of life of people with stroke over time: True change or response shift? *Quality of Life Research, 14*, 611–627.
9. Schwartz, C. E., Andresen, E. M., Nosek, M. A., Krahn, G. L., & RRTC Expert Panel on Health Status Measurement. (2007). Response shift theory: Important implications for measuring quality of life in people with disability. *Archives of Physical Medicine and Rehabilitation, 88*, 529–536.
10. Andresen, E. M., Fouts, B. S., Romeis, J. C., & Brownson, C. A. (1999). Performance of health-related quality of life instruments in a spinal cord injured population. *Archives of Physical Medicine and Rehabilitation, 80*, 877–884.
11. Ôunpuu, S., Chambers, L. W., Chan, D., & Yusuf, S. (2001). Validity of the US Behavioral Risk Factor Surveillance System's Health Related Quality of Life survey tool in a group of older Canadians. *Chronic Diseases in Canada, 22*, 93–101.
12. Mielenz, T., Jackson, E., Currey, S., DeVellis, R., & Callahan, L. F. (2006). Psychometric properties of the Centers for Disease Control and Prevention Health-Related Quality of Life (CDC HRQOL) items in adults with arthritis. *Health and Qual of Life Outcomes, 4*, 66.
13. Horner-Johnson, W., Krahn, G., Andresen, E., Hall, T., & RRTC Expert Panel on Health Status Measurement. (2009). Developing summary scores of health-related quality of life for a population-based survey. *Public Health Reports, 124*, 103–110.
14. Bann, C. M., Kobau, R., Lewis, M. A., Zack, M. M., & Thompson, W. W. (2012). Development and psychometric evaluation of the public health surveillance well-being scale. *Quality of Life Research, 21*(6), 1031–1043.
15. Mielenz, T. J., Callahan, L. F., & Edwards, M. C. (2016). Item response theory analysis of Centers for Disease Control and Prevention Health-Related Quality of Life (CDC HRQOL) items in adults with arthritis. *Health and Quality of Life Outcomes*, 14, 43.
16. Magnus, B. E., & Thissen, D. (2017). Item response modeling of multivariate count data with zero inflation, maximum inflation, and heaping. *Journal of Educational and Behavioral Statistics, 42*(5), 531–558.
17. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
18. Wang, H., & Heitjan, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine, 27*, 3789–3804.
19. Burton, S., & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly, 55*, 50–79.
20. Schneider, S., & Stone, A. A. (2016). Ambulatory and diary methods can facilitate the measurement of patient-reported outcomes. *Quality of Life Research, 25*(3), 497–506.
21. Shiffman, S. (2009). How many cigarettes did you smoke? Assessing cigarette consumption by global report, time-line follow-back, and ecological momentary assessment. *Health Psychology, 28*(5), 519–526.
22. Stone, A. A., Broderick, J. E., Shiffman, S. S., & Schwartz, J. E. (2004). Understanding recall of weekly pain from a momentary assessment perspective: Absolute agreement, between- and within-person consistency, and judged change in weekly pain. *Pain, 107*, 61–69.
23. Barile, J. P., Reeve, B. B., Smith, A. W., Zack, M. M., Mitchell, S. A., Kobau, R. et al. (2013). Monitoring population health for Healthy People 2020: Evaluation of the NIH PROMIS(R) Global Health, CDC Healthy Days, and satisfaction with life instruments. *Quality of Life Research, 22*(6), 1201–1211.
24. Yin, S., Njai, R., Barker, L., Siegel, P. Z., & Liao, Y. (2016). Summarizing health-related quality of life (HRQOL): Development and testing of a one-factor model. *Population Health Metrics, 14*, 22.
25. McGinley, J. S., & Curran, P. J. (2014). Validity concerns with multiplying ordinal items defined by binned counts: An application to a quantity-frequency measure of alcohol use. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences, 10*(3), 108–116.
26. MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40.
27. Schechter, S., Beatty, P., & Willis, G. B. (1998). Asking survey respondents about health status: Judgment and response issues. In N. Schwartz, D. Park, B. Knauper, & S. Sudman (Eds.), *Cognition, aging and self-reports*. Philadelphia: Psychology Press.
28. R Development Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
29. Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x.
30. Cai, L., Thissen, D., & du Toit, S. H. C. (2017). *IRTPRO 2.1 for windows*. Lincolnwood, IL: Scientific Software International.
31. Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48.
32. 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, 57*(1), 289–300.
33. Broderick, J. E., Schwartz, J. E., Vikingstad, G., Pribbernow, M., Grossman, S., & Stone, A. A. (2008). The accuracy of pain and fatigue items across different reporting periods. *Pain, 139*(1), 146–157.
34. Schmier, J. K., & Halpern, M. T. (2004). Patient recall and recall bias of health state and health status. *Expert Review of Pharmacoeconomics & Outcomes Research, 4*(2), 159–163.