



Careless responding in internet-based quality of life assessments

Stefan Schneider^{1,2}  · Marcella May¹ · Arthur A. Stone¹

Accepted: 12 December 2017 / Published online: 16 December 2017
© Springer International Publishing AG, part of Springer Nature 2017

Abstract

Purpose Quality of life (QoL) measurement relies upon participants providing meaningful responses, but not all respondents may pay sufficient attention when completing self-reported QoL measures. This study examined the impact of careless responding on the reliability and validity of Internet-based QoL assessments.

Methods Internet panelists ($n=2000$) completed Patient-Reported Outcomes Measurement Information System (PROMIS®) short-forms (depression, fatigue, pain impact, applied cognitive abilities) and single-item QoL measures (global health, pain intensity) as part of a larger survey that included multiple checks of whether participants paid attention to the items. Latent class analysis was used to identify groups of non-careless and careless responders from the attentiveness checks. Analyses compared psychometric properties of the QoL measures (reliability of PROMIS short-forms, correlations among QoL scores, “known-groups” validity) between non-careless and careless responder groups. Whether person-fit statistics derived from PROMIS measures accurately discriminated careless and non-careless responders was also examined.

Results About 7.4% of participants were classified as careless responders. No substantial differences in the reliability of PROMIS measures between non-careless and careless responder groups were observed. However, careless responding meaningfully and significantly affected the correlations among QoL domains, as well as the magnitude of differences in QoL between medical and disability groups (presence or absence of disability, depression diagnosis, chronic pain diagnosis). Person-fit statistics significantly and moderately distinguished between non-careless and careless responders.

Conclusions The results support the importance of identifying and screening out careless responders to ensure high-quality self-report data in Internet-based QoL research.

Keywords Quality of life · Patient-reported outcomes · Careless responding · Inattentive responding · Person-fit statistics

Introduction

Recent methodological and technological advances have transformed the measurement of quality of life (QoL) in research and clinical settings. The development of QoL measurement systems with state-of-the-science (e.g., item response theory; IRT) methods has enhanced the standardization, precision, and efficiency of QoL assessments [1, 2]. At the same time, advances in technology have spurred more extensive use of Internet-based data collection as a

cost-efficient way to monitor QoL on a population level and to compare QoL levels across medical diseases and conditions [3]. Internet survey panels also facilitate the development of modern QoL instruments by reducing expenses associated with large-scale data collection. For example, measures by the Patient-Reported Outcomes Measurement Information System (PROMIS®) were administered to large Internet panels for item calibration and to obtain reliability and validity data [2, 4].

Despite these advances, the quality of self-reported QoL data depends upon respondents’ ability and willingness to provide accurate and valid responses. Among a number of self-report biases discussed in the literature (e.g., socially desirable responding, memory bias, faking bad or good), an often overlooked concern is that some respondents may not pay sufficient attention when completing QoL measures. Inattentive or careless responding has been viewed as a form of survey “satisficing,” whereby respondents provide

✉ Stefan Schneider
schneids@usc.edu

¹ University of Southern California, Los Angeles, USA

² Dornsife Center for Self-Report Science and Center for Economic & Social Research, University of Southern California, 635 Downey Way, Los Angeles, CA 90089-3332, USA

answers but do not expend the mental effort required to fully read and interpret the questions or to generate meaningful responses [5, 6]. Johnson [7] argued that Internet-based assessments may make careless responses especially likely given that the physical distance and low personalization of web-based assessments limits participants' accountability. This may be especially problematic in anonymous testing environments, such as population-based data collection, where there are no immediate consequences to participants' response behaviors [7].

Careless responding can manifest in various ways, including completely random responses (wherein a respondent selects a response option at random for each question) or uniformly repeated responses (wherein a respondent may read a few questions and then respond to subsequent questions in the same way without noticing changes in content or direction of items). Numerous indicators of careless responding have been identified and discussed, including bogus items, consistency checks, instructed response items, and inspection of response patterns and response times [for comprehensive summaries, see 8–13]. Notably, carelessness and other forms of invalid responses have sometimes been lumped together in the literature [14, 15], even though careless responding is conceptually distinct from response biases involving intentional misrepresentation, such as impression management and socially desirable responding. Whereas the latter represent response behaviors that are motivated to appear in a particular manner and, therefore, require respondents to pay close attention to the questions, careless responses imply lack of effort and inattention [13].

Theoretically, careless responding poses potentially serious threats to the validity of self-report data in that it can undermine the psychometric properties of survey instruments, reduce statistical power, and obscure meaningful results [8, 13, 16–19]. However, it has been noted that there is a dearth of empirical studies examining the implications of carelessly invalid responses for conclusions derived from “real world” data [14]. In QoL research, much emphasis is placed on evaluating the soundness and precision of measurement *instruments* and the identification of misfitting *items* [20–22]. On the other hand, researchers typically do not routinely screen *individuals* for adequate data quality, so that the prevalence and impact of careless responses in QoL research remain largely unknown.

In the present study, we examine data from an Internet panel survey that administered PROMIS short-forms and single-item QoL measures in a general U.S. population sample. To identify potentially careless responders, the survey included several checks of whether the panelists paid attention to the items. The survey sought to evaluate the prevalence and impact of careless responding (a) on the observed reliability of PROMIS measures, (b) on the correlations between QoL domains (convergent validity), and (c) on the

magnitude of differences in QoL between respondents with and without self-reported disabilities and medical conditions (“known groups” validity). A secondary goal was to examine whether “person-fit” statistics, which have been proposed to detect individuals with aberrant or haphazard response patterns in IRT measures, could be useful to identify inattentive responders [23, 24]. If this were the case, person-fit statistics might serve as a valuable screening tool in setting in which direct inattentiveness assessments are unavailable.

Methods

Participants and procedure

Participants were 2000 adults from a U.S. national opt-in Internet panel comprising about one million households. The panel consists of members volunteering to periodically participate in online questionnaires for minimal rewards (e.g., bonus miles). Recruitment from the panel was stratified by age (balanced groups of 21–44, 45–64, and 65+ years of age), gender (50% male), and education (50% high school graduate or less, 50% some college or more). Invitations were sent to panelists until the targeted sample size in each of the strata was reached. Participants completed the QoL measures for this study as part of a larger online survey (~150 items) generated using Qualtrics (Provo, UT) that included questions about various attitudes and life concerns not considered here. All participants provided electronic informed consent before completing the survey. Respondents had the option of skipping questions, but few used this option; the average missing value rate was 0.66% (median = 0.00%; interquartile range = 0.71%). Thirteen participants with more than 10% missing values were eliminated from the data; thus, the analysis sample consisted of 1987 participants (study data and materials are publically available [25]).

Indicators of careless responding

Individual indicators of carelessness are prone to false positives (e.g., participants misunderstanding an item) and false negatives (e.g., participants selecting the correct option by chance); therefore, multiple inattentiveness checks were implemented.

Instructed response item

The most direct type of screening for inattentiveness is the use of instructed response items [8, 13]. An item that explicitly instructed participants to “move to the next item without clicking any of the answers” was administered at a random location of the survey. The response was coded as correct

if the item was skipped as instructed and as incorrect if any response option was selected.

Consistency in age reports

To capture inconsistent responses to a factual question, respondents were asked to indicate their age twice in the survey, once at the beginning (asking for age in years using an open response format) and once near the end of the survey (using a forced choice response format with 10-year age brackets). The responses were coded as inconsistent if the open-ended and forced-choice responses did not match.

Vocabulary items

Another attention screening technique involves administering items with a clearly correct answer [9]. While these items require some cognitive effort, they are relatively “easy” so that attentive respondents should respond correctly. Three vocabulary items were administered at random locations in the survey that asked for words similar in meaning to the words *permit*, *talk*, and *couch*, respectively. Out of five response choices, the correct one was a synonym (*allow*, *speak*, and *sofa*, respectively), whereas incorrect choices represented clearly unrelated words.

Figure matching

A figure matching task was included as a non-verbal attentiveness check. The task consisted of 20 items from the “Identical Pictures” test [26, 27]; ten were presented near the beginning and ten at the end of the survey. Instructions and two practice items were provided before the test items. For each item, a simple target figure appeared on the left side of the screen, and five response alternatives (1 matching and 4 clearly non-matching) on the right side of the screen. Subjects could spend up to 10 s to select the figure matching the target figure. The number of incorrectly selected figures was counted as an indicator of inattentive responding.

Response time

Short response times have been suggested to indicate low response effort and lack of cognitive processing when completing survey items [8, 9, 28]. For each respondent, we calculated the median of the response times to all survey items. Using the median was preferable over the average of response times because some respondents spent hundreds of minutes on some pages, possibly because they took a break from the survey. The distribution of median item response times was positively skewed and was normalized via log transformation.

Quality of life measures

PROMIS short-form measures for four QoL domains were administered: depression (8 items) [29], fatigue (7 items) [30], pain interference (8 items) [31], and applied cognitive abilities (8 items) [32]. PROMIS instruments have undergone rigorous psychometric evaluation and have demonstrated excellent measurement precision [2].

Single-item measures of general health and of pain intensity were also administered. General health was assessed with the SF-36 [33] item “In general, would you say your health is” (1 = excellent; 5 = poor). Pain intensity was assessed with the Brief Pain Inventory [34] average pain item (0 = no pain; 10 = pain as bad as you can imagine). All measures were administered and scored in the direction of the construct, such that higher scores indicate better health and cognitive abilities, but more severe depression, fatigue, pain interference, and pain intensity.

Data analysis

Identification of careless responders

Prior studies have shown that multiple indicators of careless responding do not always measure the same construct [8, 9]. Accordingly, initial analyses examined the dimensionality of the various careless response indicators in exploratory and confirmatory factor analysis models. Factor model fit was evaluated with the comparative fit index (CFI, > 0.95 for good fit), non-normed fit index (NNFI, > 0.95 for good fit), and root mean square error of approximation (RMSEA, < 0.06 for good fit) [35]. Latent class analysis was then used to identify potential subgroups of responders based on the multivariate profile of scores on the careless response indicators. While latent class analysis is an exploratory procedure that cannot conclusively classify careless responders and does not impose restrictions on the shape of profiles in each class, we expected to find latent classes separated by higher versus lower scores across all indicators, in line with what would be expected if one subgroup was consistently careless. The number of latent classes to be retained was determined based on information criteria (Bayesian Information Criterion, BIC; Akaike Information Criterion, AIC) and on the substantive interpretability of the extracted profiles [36–38]. Analyses were conducted in *Mplus* version 7.4. [39], using a normal distribution specification for continuous (response time), a logistic distribution function for binary (instructed response item, vocabulary items, response inconsistency), and a Poisson distribution specification for count (figure matching) indicator variables. Missing values were accommodated with maximum likelihood estimation.

Sociodemographic predictors of careless responding

Logistic regression models were used to examine whether subgroups of careless and non-careless responders identified in the latent class analysis differed on demographic characteristics (age, gender, race, ethnicity, marital status, education, income). Each demographic variable was first tested individually as a predictor of careless responder status, followed by a multiple logistic regression with all demographic variables entered simultaneously.

Impact of careless responding on psychometric properties of QoL measures

The impact of careless responding on measurement reliability was tested only for PROMIS short-forms, given that multi-item scales are required for these analyses. The impact of careless responding on correlations between QoL domains and on known-groups differences in QoL was evaluated using both PROMIS and single-item measures.

Measurement reliability Following established PROMIS psychometric testing procedures, measurement precision was evaluated with test information functions derived from IRT models (Samejima's graded response model) fitted to each PROMIS domain [20, 29–32]. Information functions depict the reliability of a measure across the levels of the underlying continuum (e.g., depression, fatigue), and were descriptively compared between careless and non-careless responder groups.

Correlations between QoL domains The matrix of inter-correlations between QoL measures was computed for participants classified as careless and non-careless responders. Tests for significant group differences in the correlations were conducted using Fisher r -to- z transformed coefficients [40]. The magnitude of differences was evaluated using Cohen's effect size q , where values of 0.10, 0.30, and 0.50 indicate small, medium, and large differences between correlations, respectively [41].

Known-groups differences in QoL Standardized mean differences (Cohen's d , where 0.20, 0.50, and 0.80 indicate small, medium, and large effects, respectively) were computed for comparisons of QoL scores between groups defined by self-reported disability status (receiving disability benefits or not) and medical status (presence or absence of depression diagnosis and chronic pain diagnosis). Whether the effect sizes differed between careless and non-careless responders was evaluated by testing interaction terms (disability/diagnosis by responder status) using standardized QoL scores as outcome variables in ANOVA models.

Detection of careless responding from person-fit statistics

Person-fit statistics evaluate the degree to which a participant's item scores conform to a specified IRT model, such that the pattern of item responses corresponds with the pattern of "difficulty" (or severity) represented by the items (e.g., a person endorsing a moderate severity item should also endorse a low severity item) [23, 24]. The *perfit* R package [42] was used to obtain three major non-parametric person-fit statistics averaged across the PROMIS short-forms: the number of Guttman errors, the normed Guttman statistic, and the generalized U3 statistic (see [23, 42, 43] for details). Receiver Operating Characteristic (ROC) analyses were then conducted to examine the extent to which the person-fit indices discriminated between careless and non-careless responder groups [44]; the area under the ROC curve (AUC) was used to summarize diagnostic accuracy (where $0.50 < AUC \leq 0.70$ suggests low, $0.70 < AUC \leq 0.90$ moderate, and $AUC > 0.90$ high accuracy) [44, 45].

Results

Participants had a mean age of 53.2 years ($SD = 16.4$, range 21–89). Most participants were White (88.6%) and non-Hispanic (95.0%), and almost two-thirds (59.4%) were married (see Table 1). One-eighth (12.3%) reported receiving disability, 21.4% reported having received a diagnosis of depressive disorder, and 25.5% reported a chronic pain condition.

Identification of careless responders

Factor analysis of the indicators of inattentiveness showed that a 1-factor model fitted the data well ($CFI = 0.98$; $NNFI = 0.98$; $RMSEA = 0.048$ [90% CI 0.038–0.059]; all loadings ≥ 0.35), suggesting a single underlying construct. An exploratory 2-factor model did not converge. In latent class models conducted with 1–5 classes, while BIC and AIC values continuously improved as more classes were extracted, inspection of "elbow plots" suggested that improvements in model fit leveled off markedly when more than 2 classes were extracted, and the 3-class model resulted in one very small (2%, $n = 44$) class. Thus, the 2-class model was preferred for parsimony and statistical reasons [36–38]. As shown in Table 2, respondents in class 2 had vastly higher rates of incorrect responses on all inattentiveness items, and showed two times faster response times, compared to those in class 1. Accordingly, class 2 ($n = 148$; 7.4%) was labeled "careless" responders, and class 1 ($n = 1839$; 92.6%) "non-careless" responders.

The latent class model showed a high entropy of 0.97 (where values closer to 1.0 imply better class separation),

Table 1 Demographic characteristics of the study sample

	<i>n</i>	%
Age		
Mean = 53.2, SD = 16.4		
21–44	690	34.7
45–64	638	32.1
65+	659	33.2
Gender		
Male	994	50.0
Female	993	50.0
Race		
African-American	121	6.1
Asian	38	1.9
Native American	11	0.6
Pacific Islander	5	0.3
White	1761	88.6
More than one race	19	1.0
Other	32	1.6
Ethnicity		
Hispanic	100	5.0
Non-Hispanic	1887	95.0
Marital status		
Never married	346	17.4
Living with partner	117	5.9
Married	1180	59.4
Separated	23	1.2
Divorced	201	10.1
Widowed	120	6.0
Education		
Up to 8th grade	10	0.5
8th–12th grade	78	3.9
High school graduate	905	45.6
Some college	315	15.9
College graduate	409	20.6
Master's degree	210	10.6
Doctoral degree	60	3.0
Yearly income		
< \$20,000	299	15.1
\$20,000–\$34,999	374	18.8
\$35,000–\$49,999	336	16.9
\$50,000–\$74,999	385	19.4
\$75,000+	588	29.6

which means that respondents could be assigned to the latent classes with very little uncertainty [46]. As shown in Table 2, indicator-specific entropies (which evaluate how well each indicator by itself can accurately identify the latent classes) were moderate to high for all indicators, with figure matching being the most successful (univariate entropy = 0.93) and inconsistent age reports being the least successful class indicator (univariate entropy = 0.63) [47].

Socioeconomic characteristics associated with careless responding

In bivariate logistic regressions, participants assigned to the careless responder class were more likely to be younger (OR 0.93 per 1-year increase in age; $p < .001$), male (OR 3.01; $p < .001$), Hispanic (OR 2.75; $p < .001$), unmarried (OR 1.62; $p < .01$), college educated (OR 1.55; $p < .05$), and to have a higher income (OR 2.15 for income \geq \$50,000 vs. less; $p < .001$) than those assigned to the non-careless responder class. When all demographic predictors were simultaneously entered in a multiple logistic regression, only age ($p < .001$) and gender ($p < .001$) uniquely predicted the responder classes. As shown in Fig. 1, rates of careless responders were roughly 30% (for men) and 10% (for women) at age 30, and they were below 1% at age 75.

Careless responding and reliability of PROMIS short-forms

Unidimensional graded response models fitted to the PROMIS short-forms in each responder class yielded model fit statistics consistent with prior results for PROMIS measures, with CFI > 0.98 (non-careless) and > 0.98 (careless), NNFI > 0.98 (non-careless) and > 0.97 (careless), and RMSEA values ranging from 0.12 to 0.19 (non-careless) and from 0.11 to 0.16 (careless) [30, 31, 48]. Test information functions of the measures showed high reliabilities in the non-careless responder class, in line with previously established results [29–32]. As shown in Fig. 2, while the reliabilities obtained for careless responders were somewhat lower for some PROMIS domains (depression, cognitive abilities, and pain impact), the test information functions still indicated reliabilities above 0.90 over wide ranges of the underlying constructs, suggesting that careless responding had a very small impact on reliability.

Careless responding and correlations between QoL domains

We next examined the impact of careless responding on the correlations between QoL domains (see Table 3). The subsample of non-careless responders showed the expected pattern of moderately positive correlations among domains that were scaled in the same direction (positive with positive and negative with negative), and negative correlations among domains that were scaled in opposite directions. By comparison, the class of careless responders showed almost uniformly more positive (i.e., closer to $r = 1.0$) correlations for domains scaled in the same direction, and less negative (i.e., closer to $r = 0.0$) correlations for domains scaled in opposite directions. The correlations significantly differed between careless

Table 2 Latent class profiles on indicators of careless responding

Careless response indicator	Latent class #1 <i>N</i> = 1839 (92.6%)	Latent class #2 <i>N</i> = 148 (7.4%)	Indicator-specific latent class separation ^b			
			Univariate entropy	Proportion correctly assigned	Negative predictive value	Positive predictive value
Incorrect instructed response item	4.2%	71.5%	0.74	0.94	0.98	0.59
Inconsistent age reports	0.7%	11.1%	0.63	0.93	0.93	0.60
Incorrect vocabulary item 1 (permit—allow)	4.5%	64.6%	0.73	0.93	0.97	0.54
Incorrect vocabulary item 2 (talk—speak)	0.7%	40.2%	0.71	0.95	0.95	0.82
Incorrect vocabulary item 3 (couch—sofa)	0.8%	47.8%	0.73	0.95	0.96	0.83
Incorrectly matched figures (out of 20)	6.7%	41.7%	0.93	0.98	0.98	0.97
Median response time per item (s) ^a	5.85	2.85	0.78	0.95	0.96	0.77

^aLog-transformed item response time was used in the analysis and was back-transformed for the presentation of means in the table

^bIndicator-specific latent class separation indices were calculated from univariate latent class models conducted for each indicator variable with all parameter values fixed at those obtained from the full model. Univariate entropy values can range from 0 to 1; higher values denote that the indicator is more informative at separating the latent classes. The proportion of subjects correctly assigned to latent classes, negative predictive values (i.e., proportion of subjects assigned to class 1 that truly belonged to latent class 1), and positive predictive values (i.e., proportion of subjects assigned to class 2 that truly belonged to latent class 2) were assessed by comparing true latent class memberships with observed class memberships (based on the highest posterior class probability)

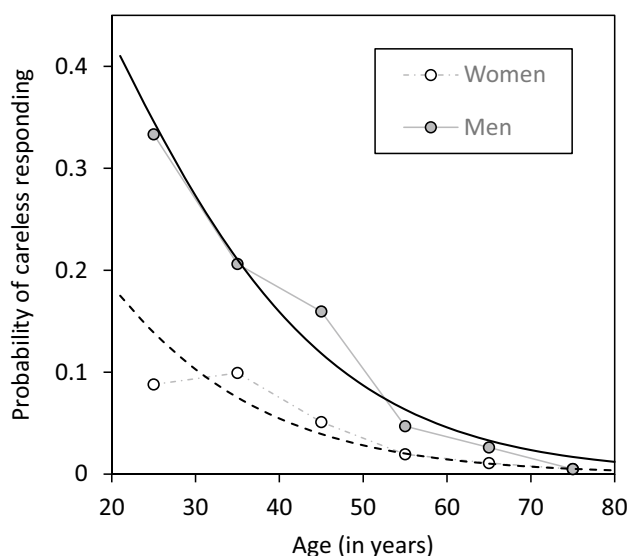


Fig. 1 Probability of careless responding by age and gender. Fitted lines are based on linear effects in logistic regression analysis. Connected dots represent observed proportions in 10-year age groups

and non-careless responder classes for each domain pair (all p s < .05), except for the correlation between fatigue and pain intensity ($p = .10$); the magnitude of differences in correlations ranged from Cohen's $q = 0.10$ to $q = 0.78$ (median $q = 0.38$). To evaluate the effect that excluding careless responders from the overall sample would have, we also compared the correlations in the full sample (i.e., all participants) with those in the non-careless responder subsample: the median difference in correlations was $q = 0.07$ (range $q = 0.00$ – 0.10 ; see Table 3).

Careless responding and known-groups validity

The subsamples of non-careless and careless responders did not significantly differ in self-reported rates of depression diagnosis (non-careless: 21.2%, careless: 24.3%; OR 1.19, $p = .37$) and chronic pain diagnosis (non-careless: 25.4%, careless: 25.7%; OR 1.02, $p = .94$). Fewer non-careless (11.8%) than careless (18.9%; OR 1.75, $p < .05$) responders reported being on disability. Table 4 shows the effect sizes for the known-group differences in QoL scores by responder class. For non-careless responders, all effect sizes were moderate to large (ranging from $d = 0.29$ to $d = 1.26$) in expected directions, and highly significant ($p < .001$). For careless responders, all effect sizes were non-significant ($p > .05$) and 15 out of 18 effect sizes were significantly lower than those for non-careless responders (p s < .05 for the interaction terms of disability/diagnosis by responder class in ANOVA models, see Table 4). The median difference in effect sizes between non-careless and careless responder classes was $d = 0.59$ (range 0.23–1.23). To evaluate the effect of eliminating careless responders from the data, we compared the effects sizes in the full sample with those in the class of non-careless responders: eliminating the class of careless responders increased the effect sizes in ranges from 5.0 to 15.9%.

Detection of careless responding from person-fit statistics

In secondary analyses, the ability of person-fit statistics to detect careless responding was examined. All three person-fit statistics significantly ($p < .001$) distinguished between

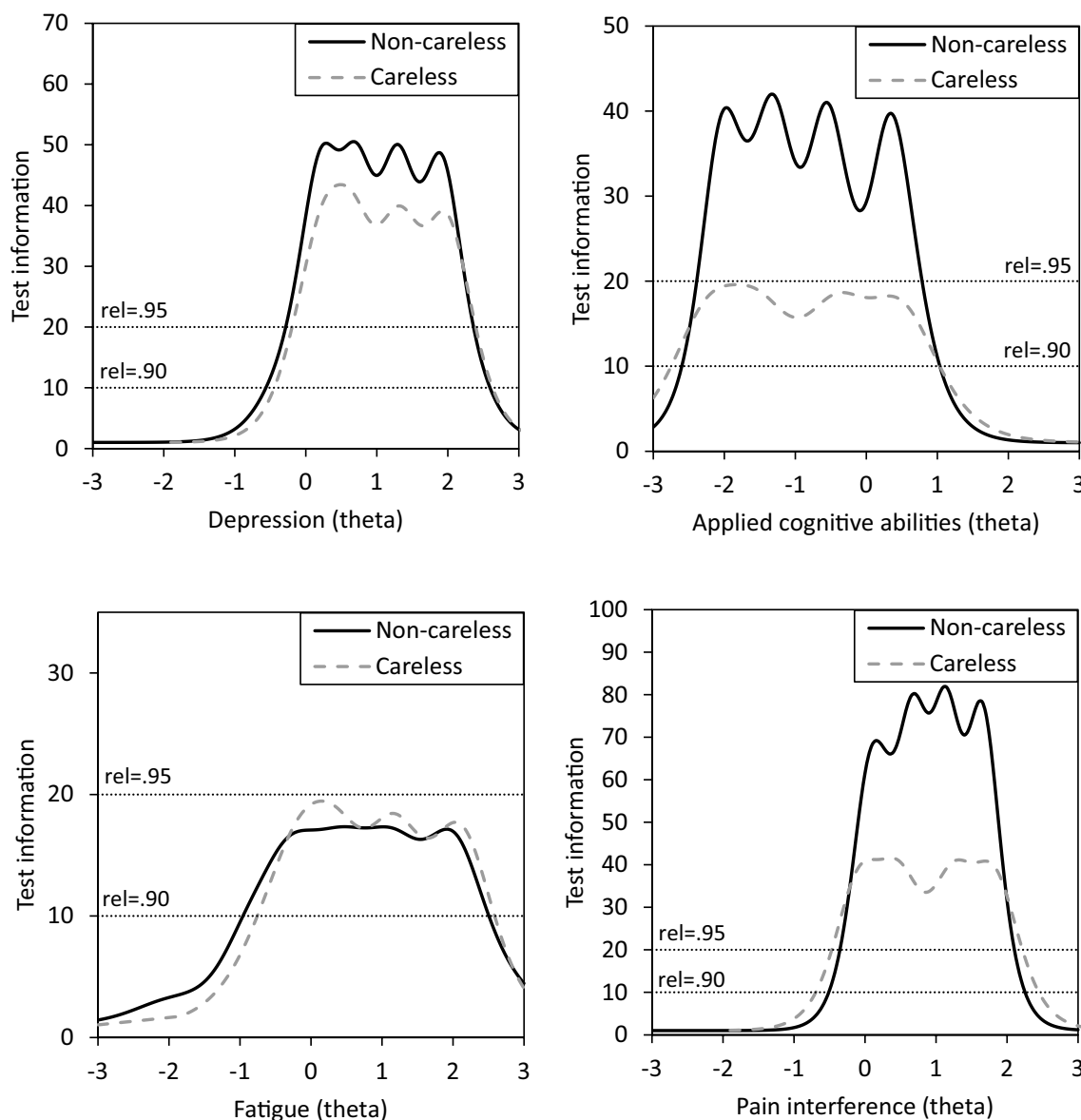


Fig. 2 Test information functions for PROMIS depression, applied cognitive abilities, fatigue, and pain interference in careless and non-careless responder groups. The horizontal-dotted lines indicate values of the test information functions comparable with internal consistency reliabilities of 0.90 and 0.95. Areas in which the test information exceeds these benchmarks represent the range of scale scores that are commonly considered precise individual scores

ency reliabilities of 0.90 and 0.95. Areas in which the test information exceeds these benchmarks represent the range of scale scores that are commonly considered precise individual scores

non-careless and careless responder classes. The AUC values were 0.79 (95% CI 0.74–0.83) for the number of Guttman errors, 0.73 (95% CI 0.68–0.77) for the normed Guttman statistic, and 0.70 (95% CI 0.65–0.74) for the U3 index, indicating moderate classification accuracy of all fit statistics (see Fig. 3). Statistical comparison of the ROC curves [49] showed that the number of Guttman errors was more accurate than the normed Guttman statistic ($\chi^2[1]=43.3, p < .001$), which in turn was more accurate than the U3 index ($\chi^2[1]=42.3, p < .001$), in distinguishing between the responder classes.

Discussion

The problem of careless responding to survey items has not yet found widespread attention in QoL research, despite the theoretical importance of the topic for ensuring psychometrically sound self-report measurement. To date, the phenomenon has been most extensively studied in personality research conducted with student samples [8, 16, 18, 50]. Thus, even though web-based data collection is popular in population health research and modern IRT-based QoL test development, little is known about the prevalence of careless

Table 3 Correlations among quality of life measures for the full sample and for subgroups of careless and non-careless responders

Pairs of outcome measures	Correlations			Difference between correlations (Cohen's <i>q</i>)	
	Careless responders	Non-careless responders	Full sample	Careless versus non-careless	Full sample versus non-careless
Pain interference—Pain intensity	0.55 _c	0.74 _c	0.74	−0.34	0.00
Depression—Fatigue	0.83 _c	0.67 _c	0.71	0.37	0.07
Fatigue—Pain interference	0.74 _b	.59 _b	0.63	0.28	0.06
Fatigue—Pain intensity	0.60	0.50	0.54	0.14	0.05
Depression—Pain interference	0.72 _c	0.45 _c	0.51	0.41	0.08
Depression—Pain intensity	0.55 _a	0.39 _a	0.44	0.20	0.07
Cognitive abilities—General health	0.30 _c	0.39 _c	0.38	−0.10	−0.01
Pain intensity—Cognitive abilities	0.13 _c	−0.30 _c	−0.26	0.44	0.04
Depression—General health	−0.05 _c	−0.40 _c	−0.31	0.37	0.09
Pain interference—Cognitive abilities	0.21 _c	−0.36 _c	−0.32	0.60	0.05
Pain intensity—General health	−0.07 _c	−0.42 _c	−0.35	0.38	0.09
Pain interference—General health	−0.07 _c	−0.44 _c	−0.36	0.41	0.10
Fatigue—General health	−0.06 _c	−.047 _c	−0.38	0.45	0.10
Depression—Cognitive abilities	0.10 _c	−0.53 _c	−0.47	0.70	0.08
Fatigue—Cognitive abilities	0.14 _c	−.56 _c	−0.50	0.78	0.09

Variable pairs are sorted in descending order of correlations in the full sample. Correlations sharing the same subscript significantly differ between careless and non-careless responders at $p < .05$ (subscript _a), $p < .01$ (subscript _b); $p < .001$ (subscript _c)

responding in general population samples and its impact on the quality of Internet-based QoL data.

The findings from the current study support the position that identifying careless responders may have important implications for web-based QoL research. Based on multiple attention checks interspersed throughout the survey, latent class analysis categorized 7.4% of survey panelists as careless responders. This prevalence is only slightly lower than the 10–12% careless responders documented in research conducted with undergraduate students [8]. Careless responding was much more prevalent among younger and male panelists compared to older and female participants. These demographic effects are strikingly different from those found for other response biases, such as acquiescent and extreme response styles, which have been observed more frequently in older and female respondents [51, 52]. This suggests that careless responding represents a distinctive phenomenon in survey research and should not be viewed as interchangeable with other response biases.

Careless responding had little impact on the test information functions of PROMIS measures in contrast to a previous study that found pronounced effects of careless responding on measurement reliability [13]. A potential explanation is that uniform response patterns predominated over completely random patterns among careless responders. Whereas random responses increase measurement error and attenuate reliabilities, uniform patterns, in which participants

repeatedly use the same response option, can artificially inflate internal consistency estimates [53]. This interpretation is also consistent with the observed impact of careless responding on the correlations between QoL domains, where correlations of QoL measures scaled in the same direction were inflated and correlations of measures scaled in opposite directions were greatly diminished or even reversed in the careless responder group: this effect is exactly what would be expected to result from uniformly repeated responses. While other samples of careless responders may display different compositions of artificial response patterns [8], the results show that even a relatively small number (7.4%) of careless participants can generate complex and meaningful biases in the psychometric properties of QoL measures.

With respect to “known groups” differences in QoL, the results showed that the presence of careless responders in the dataset consistently weakened the effect sizes for comparisons between groups with and without disability and medical conditions. While one might have suspected that single-item QoL measures may be especially susceptible to this given their lower reliability compared with multi-item scales, the effects were similar for both single-item and PROMIS short-form measures, suggesting that careless responders were consistently inattentive across multi-item scales. An interesting question is whether excluding careless responders from the data increases the nominal statistical power to detect the examined group differences. In other words, does

Table 4 Effect sizes for known-groups differences in quality of life by careless responder status

Comparison groups	ES (Cohen's <i>d</i>)			<i>F</i> -value for difference in ES between careless and non-careless responders	% difference in ES between non-careless responders and full sample
	Careless responders ^a	Non-careless responders ^b	Full sample ^b		
Disability					
Depression	-0.02	0.41	0.38	3.74‡	7.9%
Cognitive abilities	-0.10	-0.33	-0.30	1.04	7.0%
Fatigue	-0.06	0.54	0.50	7.17**	8.3%
Pain interference	0.14	0.75	0.71	7.40**	6.0%
General health	-0.16	-0.74	-0.64	6.79**	15.6%
Pain intensity	0.01	0.66	0.61	8.58**	8.9%
Depression diagnosis					
Depression	0.37	1.15	1.04	13.67***	10.5%
Cognitive abilities	-0.13	-0.81	-0.75	11.14***	7.2%
Fatigue	0.30	1.05	0.95	12.78***	10.0%
Pain interference	0.19	0.77	0.71	8.19**	9.0%
General health	-0.26	-0.67	-0.62	3.89*	8.2%
Pain intensity	0.41	0.67	0.64	1.56	5.0%
Chronic pain diagnosis					
Depression	-0.05	0.37	0.32	4.62*	14.6%
Cognitive abilities	0.28	-0.29	-0.25	8.58**	15.9%
Fatigue	0.12	0.71	0.64	8.82**	11.6%
Pain interference	0.27	1.40	1.23	30.25***	13.5%
General health	-0.36	-0.82	-0.77	5.18*	6.6%
Pain intensity	0.33	1.56	1.36	34.52***	14.6%

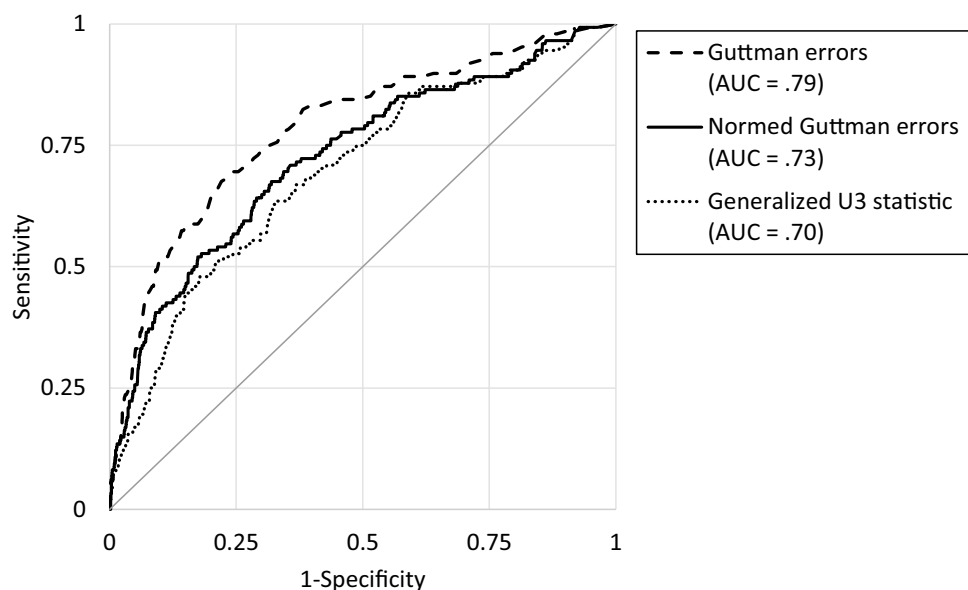
‡*p* < .06; **p* < .05; ***p* < .01; ****p* < .001.

^aEffect sizes are all non-significant (*p* > .05) for careless responders

^bEffect sizes are all significant (*p* < .001) for non-careless responders and for the full sample

ES effect size

Fig. 3 ROC curves for the prediction of careless responders from person-fit indices



the gain in power due to effect size increases outweigh the loss in power due to sample size reductions resulting from the exclusion of careless responders? Simple power calculations [41] suggest that this is the case: excluding careless responders reduces the sample size required to detect the observed effect sizes with 80% power by more than 7.4% (i.e., by more than the sample size reduction resulting from excluding careless responders) for all conducted comparisons and QoL outcomes. This suggests that identifying and screening out careless responders may reduce false negatives (type 2 errors) in QoL research.

In terms of practical implementation of these results, we recognize that including all of the attention checks used in this survey may be impractical. However, as suggested by the moderate to high indicator-specific entropy values, and consistent with prior research [13], inserting even one or two questions into a survey may help detect careless responders with reasonably high accuracy.¹ Furthermore, IRT-based person-fit statistics obtained from PROMIS measures were moderately associated with careless responding, with the raw number of Guttman errors slightly outperforming other fit statistics. This is consistent with simulation studies showing that a simple count of the Guttman errors is effective in detecting serious person misfit [23]. Thus, even though person-fit statistics are aimed at identifying aberrant response patterns of any kind and are not specific to inattentive response behaviors, they may be a useful tool for flagging potentially careless responders when more direct indicators of inattentiveness are not available.

Several limitations need to be considered in interpreting the study results. Most importantly, the QoL measures were administered as part of a longer questionnaire, which may have increased the prevalence of careless responses relative to short surveys [8]. Moreover, the data were collected from an opt-in Internet panel. Lower data quality of opt-in panels compared to probability panels has been documented [54], even though the extent to which careless responses are less prevalent or impactful in probability panels needs to be examined. Some panelists may have spotted designated inattentiveness checks that look different from other items, which may have missed some careless responders. It is also

¹ To evaluate how well 2 selected indicators would recover the full latent class solution, we fitted a series of additional latent class models considering all 21 combinations of indicator pairs. Most indicator pairs were reasonably successful at replicating the class assignments when judged by the proportion of respondents correctly assigned (range 0.80–0.96), positive predictive values (PPV range 0.59–1.0), and negative predictive values (NPV range 0.94–0.97). Pairs involving “inconsistent age reports” tended to perform the least well. Combinations of “median response time” with a vocabulary item, instructed response item, or figure matching task tended to most successfully recover the class assignments from the full model (proportions correctly assigned ≥ 0.95 , PPV > 0.70 , NPV > 0.96).

unclear whether the present results generalize from anonymous survey settings to clinical settings, in which patients completing QoL measures may be much more attentive to the questions, especially in settings where responses may be perceived as impacting treatments. Finally, participant disability and medical status were determined via self-reports, and we cannot say whether careless responders paid insufficient attention to these items to answer them correctly. The impact of careless responding on known-groups differences may be less pronounced in studies obtaining confirmation of medical status from clinicians or medical tests, and the results may not generalize to studies that do not rely on self-reported medical diagnoses.

In conclusion, the study results support the importance of identifying and screening out careless responders to ensure high-quality self-report data in Internet-based QoL research. More research is needed to examine ways to avoid or prevent careless response tendencies before they occur. In addition, while the present study examined the role of careless response behaviors as a measurement confound, they may also represent a substantive individual differences marker (e.g., a manifestation of respondent personality, see [55]). An interesting avenue for future research is to examine whether careless responding can indicate motivational or cognitive deficits that may be substantively implicated in impaired QoL.

Acknowledgements We would like to thank Margaret Gatz, PhD, and Doerte U. Junghaenel, PhD, for their comments on the study design and helpful discussions in preparation of this manuscript.

Funding This work was supported by a grant from the National Institute on Aging (R01 AG042407).

Compliance with ethical standards

Conflict of interest A.A.S. is a Senior Scientist with the Gallup Organization and a consultant with Adelphi Values, inc. S.S. and M.M. declare that they have no conflict of interest.

Ethical approval The study was approved by the University of Southern California Institutional Review Board. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Cella, D., Gershon, R., Lai, J.-S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*, 133–141.
2. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave

- of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*, 1179–1194.
3. Eysenbach, G., & Wyatt, J. (2002). Using the Internet for surveys and health research. *Journal of Medical Internet Research*, *4*, e13.
 4. Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the patient-reported outcomes measurement information system internet panel. *Journal of Clinical Epidemiology*, *63*, 1169–1178.
 5. Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539–569.
 6. Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.
 7. Johnson, J. A. (2005). Ascertain the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*, 103–129.
 8. Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455.
 9. Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.
 10. Godinho, A., Kushnir, V., & Cunningham, J. A. (2016). Unfaithful findings: Identifying careless responding in addictions research. *Addiction*, *111*, 955–956.
 11. Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.
 12. Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99–114.
 13. Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.
 14. McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*, 450–470.
 15. Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78*, 582–593.
 16. Osborne, J. W., & Blanchard, M. R. (2010). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, *1*, 220.
 17. Miura, A., & Kobayashi, T. (2016). Survey satisficing inflates stereotypical responses in online experiment: The case of immigration study. *Frontiers in Psychology*, *7*, 1563.
 18. Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on internet-based surveys. *Computers in Human Behavior*, *48*, 554–568.
 19. Fervaha, G., & Remington, G. (2013). Invalid responding in questionnaire-based research: Implications for the study of schizotypy. *Psychological Assessment*, *25*, 1355–1360.
 20. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks - Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care*, *45*, S22–S31.
 21. Zhao, Y. (2017). Impact of IRT item misfit on score estimates and severity classifications: An examination of PROMIS depression and pain interference item banks. *Quality of Life Research*, *26*, 555–564.
 22. Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., et al. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research*, *18*, 313–333.
 23. Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224–247.
 24. Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, *20*, 159–168.
 25. Schneider, S. (2017). Careless responding. *osf.io/um9d3*.
 26. Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.
 27. Lindenberger, U., Mayr, U., & Kliegl, R. (1993). Speed and intelligence in old age. *Psychology and Aging*, *8*, 207–220.
 28. Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.
 29. Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., et al. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS (R)): Depression, anxiety, and anger. *Assessment*, *18*, 263–283.
 30. Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, *92*, S20–S27.
 31. Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W. H., Choi, S., Revicki, D., et al. (2010). Development of a PROMIS item bank to measure pain interference. *Pain*, *150*, 173–182.
 32. Becker, H., Stufbergen, A., Lee, H., & Kullberg, V. (2014). Reliability and validity of PROMIS cognitive abilities and cognitive concerns scales among people with multiple sclerosis. *International Journal of MS Care*, *16*, 1–8.
 33. Ware, J. E. Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, *30*, 473–483.
 34. Cleeland, C. (1994). Pain assessment: Global use of the Brief Pain Inventory. *Annals of Academic Medicine Singapore*, *23*, 129–138.
 35. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.
 36. Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 345–369). Thousand Oaks, CA: Sage.
 37. Nylund, K. L., Asparoutiov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling-A Multidisciplinary Journal*, *14*, 535–569.
 38. Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods*, *8*, 369–377.
 39. Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (7th edn.). Los Angeles, CA: Muthén & Muthén.
 40. Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
 41. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Erlbaum.
 42. Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2015). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*, 1–27.

43. Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.
44. Greiner, M., Pfeiffer, D., & Smith, R. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, *45*, 23–41.
45. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.
46. Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*, 195–212.
47. Asparouhov, T., & Muthén, B. (2014). Variable-specific entropy contribution. Retrieved June 30, 2017, from <https://www.statmodel.com/download/UnivariateEntropy.pdf>.
48. Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, *18*, 447–460.
49. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.
50. Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*, 596–612.
51. Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*, 195–217.
52. Schneider, S. (2018). Extracting response style bias from measures of positive and negative affect in aging research. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *73*, 64–74.
53. Fong, D. Y., Ho, S., & Lam, T. (2010). Evaluation of internal reliability in the presence of inconsistent responses. *Health and Quality of Life Outcomes*, *8*, 27.
54. Callegaro, M., Villar, A., Krosnick, J., & Yeager, D. (2014). A critical review of studies investigating the quality of data obtained with online panels. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective*. Hoboken, NJ: Wiley.
55. Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*, 218–229.