Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Qual Life Res. Author manuscript; available in PMC 2017 October 01.

Published in final edited form as:

Qual Life Res. 2016 October ; 25(10): 2511–2521. doi:10.1007/s11136-016-1293-7.

The meaning of vaguely quantified frequency response options on a quality of life scale depends on respondents' medical status and age

Stefan Schneider, PhD and **Arthur A. Stone, PhD**

University of Southern California

Abstract

Purpose—Self-report items in quality of life (QoL) scales commonly use vague quantifiers like "sometimes" or "often" to measure the frequency of health-related experiences. This study examined whether the meaning of such vaguely quantified response options differs depending on people's medical status and age, which may undermine the validity of QoL group comparisons.

Methods—Respondents (n = 600) rated the frequency of positive and negative QoL experiences using vague quantifiers (*never, rarely, sometimes, often, always*) and provided open-ended numeric frequency counts for the same items. Negative binomial regression analyses examined whether the numeric frequencies associated with each vague quantifier differed between medical status (no vs. one or more medical conditions) and age (18–40 years vs. 60+ years) groups.

Results—Compared to respondents without a chronic condition, those with a medical condition assigned a higher numeric frequency to the same vague quantifiers for negative QoL experiences; this effect was not evident for positive QoL experiences. Older respondents' numeric frequencies were more extreme (i.e., lower at the low end and somewhat higher at the high end of the response range) than those of younger respondents. After adjusting for these effects, differences in QoL became somewhat more pronounced between medical status groups, but not between age groups.

Conclusions—The results suggest that people with different medical backgrounds and age do not interpret vague frequency quantifiers on a QoL scale in the same way. Open-ended numeric frequency reports may be useful to detect and potentially correct for differences in the meaning of vague quantifiers.

Keywords

Quality of life; chronic illness; age; frequency ratings; vague quantifiers; self-report

Corresponding author: Stefan Schneider, Ph.D., Dornsife Center for Self-Report Science and Center for Economic & Social Research, University of Southern California, 635 Downey Way, Los Angeles, CA 90089-3332, schneids@usc.edu.

Compliance with Ethical Standards

Conflict of interest: S.S. declares that he has no conflict of interest. A.A.S. is a Senior Scientist with the Gallup Organization.

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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Response scales in quality of life (QoL) instruments commonly use vague quantifiers (e.g., *never, rarely, sometimes, often*) to measure health-related experiences. Vague quantifier scales are attractive because they provide an efficient means to communicate the response range in self-report measures and are applicable to a wide variety of experiences and behaviors [1,2]. However, their use has also been considered problematic in that vague quantifiers may have different meanings for different individuals [2]. Volkmann's [3] "rubber band" model suggests that people identify the meaning of rating scale labels relative to the breadth of the stimulus range that comes to mind. As the perceived stimulus range expands or retracts, the meaning of response labels is stretched or compressed accordingly. For example, a “very tall” mouse does not imply the same height as a “very tall” elephant, and going to the gym “a lot” may not denote the same frequency for an athlete than it does for a non-athlete. Hakel [4] states this simply by saying: “one man’s ‘rarely’ is another man’s ‘hardly ever’” (p.533).

The purpose of this study was to investigate whether the meaning of vaguely quantified response options differs depending on people’s medical status and age, and whether this biases group comparisons in QoL research. How people’s QoL changes in response to chronic illness and with increasing age is of substantial interest for research and policy. Accumulating evidence suggests that people often adapt positively to chronic disease and report better QoL than might be expected based on their health conditions [5–7]. Similarly, older people tend to report better wellbeing than younger people despite age-related declines in physiological functioning and physical health [8,9]. However, there is also reason to believe that as people age or live with a chronic illness, they may adjust their internal standards for evaluating their QoL [10–14], a phenomenon that response shift theory has termed *scale recalibration* [13,14]. Scale recalibration may occur as a result of experience with health problems [12,13], changes in salient social comparison groups [15,16], or shifts in aspirations and expectations over time and with increasing age [17,18]. Even though scale recalibration processes can reflect adaptive responses to chronic illness or older age, they may threaten the validity of group differences if participants use different “measuring sticks” and their ratings are not based on a common metric. For example, if patients respond to QoL questions by comparing themselves to other patients, whereas healthy people compare themselves to other healthy people, this may diminish differences in average QoL scores between these groups [10,16]. Similarly, a study by Ubel et al. [19] found that when people of different ages are asked to rate their health, they implicitly adjust the scale endpoint “perfect health” in accordance with their expectations of “perfect health for their own age”. If response scale labels are ambiguous and vaguely quantified, such scale recalibration effects may be more likely to occur because the response options leave the respondents more room for different interpretations based on their own internal standards [13,19].

Our research strategy to investigate respondents’ interpretation of vague quantifiers is to level their ratings against a metric that does not have a relativistic meaning because its units are presumably interpreted by all individuals in the same way [20,21]. While such invariant metrics are not readily available for many aspects of QoL measurement [10], the *frequency* of QoL experiences can be (within limits) assessed in non-relativistic units by asking participants to *count the number of times* these experiences occurred within a given time period [1,22–24]. Thus, the present study compared respondents’ frequency ratings using a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

vaguely quantified response scale (as would be the case in "normal" use of the items) with their open-ended numeric frequency counts for the same QoL experiences. This is not to say that retrospective frequency counts accurately capture people's QoL experiences. A judgment such as "last week, I felt this symptom *five times*" may be distorted by memory biases [25] and by different strategies people use to identify symptom occurrences. However, we assume that numeric frequency counts can be used as a reference point to examine how groups differ in the subjective definition of vaguely quantified response options, because frequency counts are not likely to be impacted by internal standards people apply when they label a frequency as "rarely" or "often" [2].

Conceptually, shifts in the use of vague quantifiers by ill and older people should occur regardless of an item's valence. However, questions that are worded positively (e.g., How often did you sleep well last week?) and negatively (e.g., How often did you sleep poorly last week?) are scored in opposite directions in QoL measures. Accordingly, it might be hypothesized that the effect of scale recalibration should also be operating in opposite directions. For example, respondents who have experienced years of problematic sleep may have adjusted their standards so that having "trouble sleeping 'often'" denotes a higher frequency, and having "a good night's sleep 'often'" denotes a lower frequency for them than for healthy individuals. Alternatively, it might also be argued that scale recalibration effects could operate in the same direction for positively and negatively worded questions. For example, older adults may have a wider internal reference scale than younger adults for what it means to have both positive and negative QoL experiences "often" because they have had more opportunities to experience highly (or frequent) positive and negative states over the life span. To better understand these different possibilities, we examined group (medical status and age) differences in the numeric frequencies associated with vague quantifiers for both negative and positive QoL items in this study.

Methods

Measures

The self-report questions (i.e., items) chosen for this study were drawn from the Patient-Reported Outcomes Measurement Information System (PROMIS®). The development of PROMIS items involved extensive qualitative item review [26,27], as well as rigorous psychometric evaluation, including tests for differential item functioning across age and illness subgroups [28,29]. Items are calibrated on a T-score metric (mean=50, *SD*=10) relative to the general US population [28].

A subset of PROMIS items employing a 5-point frequency response format with vague quantifiers: *never, rarely, sometimes, often, always,* were administered. All items asked about the "last 7 days". We selected 2 fatigue items with negative wording (*How often did you feel run down? How often did you have to push yourself to get things done because of your fatigue?*) and 2 with positive wording (*How often did you have enough energy to exercise strenuously? How often were you energetic?*). Similarly, we selected 2 sleep disturbance items with negative wording (*I had trouble sleeping; I had trouble staying asleep*) and 2 with positive wording (*I got enough sleep; It was easy for me to fall asleep*). Negative affect was assessed with 2 anger (*I felt angry; I felt like yelling at someone*) and 2

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

depression items (*I felt depressed; I felt lonely*). Positive affect was assessed with 2 Neuro-QOL items (*I felt cheerful; I felt hopeful*)[30].

For each original item with the vague frequency response format, a parallel version of the item was created to obtain numeric frequency counts. Each item was worded to begin with "in the past 7 days, how many times…"; the content and wording of the original items was otherwise left unchanged. Respondents were asked to enter a number in an open-ended response format ("___times"), with possible responses ranging from 0 to 99.

To assess participants' medical history, they were asked to report for each of 16 chronic health conditions, "Has a doctor, nurse, or other health professional ever told you that you had the following?" The list of conditions was taken from PROMIS wave 1 instrument testing [31], and it included cardiac, neurological, psychiatric, pulmonary, and other diagnoses (Table 1).[1]

Participants and procedure

Participants were 600 adults who were drawn from a U.S. national Internet panel of about one million households, hosted by Survey Sampling International (SSI). The opt-in panel consists of Internet users who volunteered to periodically participate in questionnaires for minimal compensation. Invitations to complete the questions were sent to panelists in two age groups (age 18–40 years and 60–93 years, N=300 in each group, stratified by gender) until the targeted sample size was reached. Participants were sampled from younger and older age groups rather than including the full adult age range in order to enhance the variance of age as a predictor variable. Panelists were recruited without regard to health conditions.

Participants completed the questions online and were presented one item at a time. PROMIS items (using vague quantifiers) were administered first, followed by those with open-ended numeric frequency response format; this order was chosen so that providing numeric responses would not impact participants' subsequent interpretation of the vague quantifiers. Sociodemographic characteristics and medical conditions were assessed at the end of the survey.

Analysis plan

Initial descriptive analyses examined the distributions (means, SDs, range) of the numeric frequencies to ensure that they showed sufficient variation in each vague quantifier category for meaningful group comparisons. For purposes of group comparisons by medical status, participants were categorized into those who reported no chronic conditions versus those who reported one or more of the 16 listed conditions. Age groups were defined as younger (18–40 years) and older (60+ years).

Regression analyses were used to determine whether the groups differed in the numeric frequencies associated with the vague quantifiers. To avoid inflation of type 1 error due to multiple comparisons, we fitted two regression models, one for the 8 negatively keyed QoL

[1]PROMIS wave 1 testing included 24 health conditions; the 16 most prevalent were used here.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

items, and one for the 6 positive items, rather than testing each item individually. The models were estimated with clustered robust standard errors in STATA version 13 to accommodate the clustered data structure (multiple items nested in participants).

Negative binomial regression models were estimated because these are more appropriate for modeling count outcome data (i.e., numeric frequency counts) than ordinary linear regressions. We decided to use negative binomial models over Poisson regression models because significant over-dispersion was evident in regressions predicting the numeric frequencies for negative (likelihood ratio test of dispersion parameter alpha $\chi^2(1)= 886.5$, $p <.001$) and positive ($\chi^2(1)= 1062.7$, $p <.001$) QoL items. Zero-inflated models were also considered but not used because there was no evidence of excess zeros [32]. Deviance goodness of fit tests for the final negative binomial models with all predictors were non-significant for negative ($\chi^2(4750)= 4452.5$, p >.90) and positive ($\chi^2(3560)= 3691.7$, $p =.06$) QoL items, indicating that the negative binomial models fit the data well [33].

Specifically, the form of the model equation was

$$\begin{gathered}\log(Y)=intercept\\ +b_{I2}D_{I2}+b_{I3}D_{I3}+....+b_{In}D_{In}\\ +b_{C2}D_{C2}+b_{C3}D_{C3}+b_{C4}D_{C4}+b_{C5}D_{C5}\\ +b_{I2C2}(D_{I2}D_{C2})+b_{I2C3}(D_{I2}D_{C3})+b_{I2C4}(D_{I2}D_{C4})+b_{I2C5}(D_{I2}D_{C5})+\ldots+b_{InC5}(D_{In}D_{C5})\\ +b_{M2}(D_{M2})\\ +b_{A2}(D_{A2})\\ +b_{M2C2}(D_{M2}D_{C2})+b_{M2C3}(D_{M2}D_{C3})+b_{M2C4}(D_{M2}D_{C4})+b_{M2C5}(D_{M2}D_{C5}),\\ +b_{A2C2}(D_{A2}D_{C2})+b_{A2C3}(D_{A2}D_{C3})+b_{A2C4}(D_{A2}D_{C4})+b_{A2C5}(D_{A2}D_{C5}),\end{gathered}$$

where *b*s are estimated regression coefficients for dummy-coded predictors *D*, subscript *I* refers to the PROMIS self-report questions (items), subscript *C* refers to the categories of the response scale within each item (vague quantifiers), subscript *M* refers to medical groups, and subscript *A* refers to age groups.

The numeric frequency responses *Y* served as dependent variable[2]. Dummy codes for the self-report items (D_{I2} − D_{In}, where n=8 for negative and n=6 for positive items), vague quantifier response categories (D_{C2} − D_{C5}), and their interaction ($D_{I2}D_{C2}$ − $D_{In}D_{C5}$) were entered first as categorical predictors to control for differences between items and ensure that effects of group reflected differences from those that *would be expected* based on the vague quantifier responses. The focal predictors entered next were the main effects of medical and age groups (D_{M2} and D_{A2}) and the group X vague quantifier interaction terms ($D_{M2}D_{C2}$ − $D_{M2}D_{C5}$ and $D_{A2}D_{C2}$ − $D_{A2}D_{C5}$): a main effect indicates that the frequency responses of the groups differ uniformly across the vague quantifier response categories, and a group X vague quantifier interaction indicates that the group difference in frequency responses depends upon the vague quantifier response category. Additional higher-order

[2]The model could also have been specified using the vague quantifiers as categorical dependent variable and the numeric frequencies as continuous predictor variable. The selected model had the advantage that we could use the vague quantifier categories as dummy-coded predictors, thereby not imposing any assumptions about the functional form (linear, quadratic, etc.) of the relationship between numeric frequencies and vague quantifier categories.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

terms (i.e., the item X group 2-way interactions and item X vague quantifier X group 3-way interactions) were also examined but found non-significant and are not reported. An alpha level of .05 was used for all comparisons. Nagelkerke's pseudo R^2 values [34] are reported to quantify the degree to which each covariate improved the prediction of the numeric frequency responses.

In addition to testing the statistical significance of the effects, we were also interested in the magnitude of the potential impact that differences in the frequency counts underlying the vague quantifiers could have for QoL group comparisons. To investigate this, we compared the medical and age-groups on T-scores for each PROMIS QoL domain (fatigue, sleep disturbance, positive affect, anger, depression) before and after adjustment for differences in the numeric frequencies associated with vague quantifiers. Specifically, we saved the residuals from negative binomial regressions predicting the numeric frequencies from vague quantifier responses (i.e., we saved the deviations of the numeric frequencies from what would be expected based on the vague quantifier response) for each item; these were used to estimate propensity scores (using logistic regression analysis) for the imbalance in the groups' numeric frequencies for all items of a given QoL domain. T-scores for each QoL domain were then compared between medical and age-groups before and after conditioning on the propensity scores [35].

Finally, we examined whether computing QoL "scale scores" from the numeric frequency reports would yield similar conclusions about medical and age group differences as PROMIS (vague quantifier) scale scores. For this purpose, the numeric frequency scores were log transformed and summated across the items of each QoL domain. The numeric frequency and PROMIS scale scores for each domain were converted into z-scores to obtain standardized differences (Cohen's *d*) between medical and age groups for both types of response formats. Whether the standardized effects differed between numeric frequency and PROMIS scale scores was examined by testing the interaction term of medical or age group (between-subjects) by type of response format (within-subjects) in 2x2 ANOVA models.

Results

Participants had a mean age of 48.3 years (*SD*=21.8, range=18–93), with mean ages of 27.4 (*SD*=6.3) years in the younger and 69.2 (*SD*=6.2) years in the older age groups. Most participants where White (79%) and non-Hispanic (86%). The rate of respondents with some college education or more was 70%, exceeding the national rate of 57% reported by the U.S. Census for these age groups [36]. About one fourth (26%) of the sample reported no medical condition, and 74% reported one or more conditions (Table 1). The most common conditions were hypertension (38%), arthritis (30%), anxiety (25%), depression (22%), migraines (18%), asthma (16%), and sleep disorder (15%); prevalence rates were comparable to those reported in the PROMIS wave 1 testing sample [31].

Descriptive characteristics of the numeric frequency responses for each item are shown in Table 2. The distributions of responses were positively skewed, as is typical for frequency counts [32]. The minimum frequency was zero for all items, and maximum frequencies ranged between 20 and 99. Negative QoL experiences were generally less frequently

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

reported than positive QoL experiences, with mean frequencies ranging from 2.2 to 3.3 for negative and from 2.8 to 5.5 for positive items. Average frequency reports by vague quantifier category are also shown in Table 2, and they showed the expected pattern of higher numeric frequencies associated with higher vague quantifiers. Importantly, there was pronounced variation (*SD*s ranging from 0.4 to 11.6, with a median *SD* of 2.1) around the mean frequency response for each vague quantifier category, including variation in the frequencies for "never" and "always"; thus, it was reasonable to examine evidence for systematic group differences underlying this variation.

Does the numeric frequency associated with vague quantifiers differ by medical status?

Results of the negative binomial regressions are shown in Table 3. For negative QoL items, after controlling for the effects of items, vague quantifiers, and their interaction (R^2_{pseudo}= .602) medical status was a highly significant (p <.001; ΔR^2_{pseudo}= .008) predictor of the numeric frequencies; the medical status X vague quantifier interaction was marginally significant (p =.06; ΔR^2_{pseudo}= .002). As illustrated in Figure 1, people with a medical condition assigned a *higher* numeric frequency to a given vague quantifier response category than people without a medical condition for negative QoL experiences. This effect was nonsignificant for the frequency associated with response category "never" (p =.29) and became significant and increasingly more pronounced for vague quantifier categories "rarely" (p =.01) "sometimes" (p <.01), "often" (p <.001), and "always" (p =.01). Compared to people with no condition, those with a medical condition reported a 1.26 times greater numeric frequency for "rarely", 1.19 times greater numeric frequency for "sometimes", a 1.32 times greater frequency for "often", and a 1.39 times greater frequency associated with "always" having negative QoL experiences.

For positively keyed QoL items, neither the main effect of medical status (p =.30) nor the interaction with vague quantifiers (p =.41) were significant (Table 3 and Figure 1).

To determine whether these results were affected by demographic differences between the medical status groups, we also estimated negative binomial regression models with gender, education, race, ethnicity, and marital status included as covariates. Main effects of ethnicity ($\chi^2(1)$=5.13, p =.02) and marital status ($\chi^2(1)$=9.59, p <.01) were significant for positive QoL items (*ns* for negative items), with Hispanic and non-married participants associating the vague quantifiers with higher frequencies compared to non-Hispanic and married participants, respectively. Controlling for the demographic variables did not alter the results for medical status groups. As a further robustness check, we re-ran the models with numeric frequency data trimmed at the 90th percentile to rule out that the effects were driven by potential outlier responses; the effects remained unchanged.

We also explored whether the results would generalize across medical conditions or whether the effects were illness-specific. To do this, we estimated a separate regression model for the 7 most prevalent conditions (i.e., those with a prevalence rate >15%, see Table 1) comparing those people who reported the medical condition versus those who did not. For negative QoL items, the effect was evident across most of the conditions (see Figure 2); main effects of condition were significant for anxiety ($\chi^2(1)$=9.24, p <.01), depression ($\chi^2(1)$=28.62, p <.001), migraines ($\chi^2(1)$= 3.88, p <.05), asthma ($\chi^2(1)$=3.90, p <.05) and sleep disorder

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

($\chi^2(1)=6.47$, $p=.01$), and nonsignificant for hypertension ($\chi^2(1)=1.14$, $p=.29$) and arthritis ($\chi^2(1)=0.19$, $p=.66$); the condition X vague quantifier interactions were not significant (ps >.05). For positive QoL items, condition-specific effects were nonsignificant.

Does the numeric frequency associated with vague quantifiers differ by age?

Age differences in the numeric frequency reports are also shown in Table 3. For negative QoL items, the main effect of age was highly significant ($p<.001$, $\Delta R^2_{pseudo}= .009$), as was the age X vague quantifier interaction ($p<.001$, $\Delta R^2_{pseudo}= .014$). For positive QoL items, no main effect of age emerged ($p=.09$), but the age X vague quantifier interaction was highly significant ($p<.001$, $\Delta R^2_{pseudo}= .012$).

The nature of the age effects is shown in Figure 3. For negative QoL items, older people indicated *lower* numeric frequencies associated with response options “never” ($p<.001$) and “rarely” ($p<.001$) than younger people; frequency responses associated with the remaining vague quantifiers were comparable between the age-groups (ps >.07). For positive QoL items, older people indicated *lower* frequencies associated with “never” ($p=.10$) and “rarely” ($p<.001$), and significantly *higher* frequencies associated with the response option “always” ($p<.001$), than younger people. Thus, the frequency responses associated with the vague quantifiers followed a pattern that was more “stretched out” for older people compared to younger people, especially for positive QoL items.

Quantifying the impact of group differences in the use of vague quantifiers

Table 4 shows the group differences in PROMIS T-scores for each of the QoL domains before and after adjustment for numeric frequency responses. We start with the illness groups. Before adjustment, participants with a medical condition scored between 2.61 and 2.98 T-scores (ps <.001) higher on fatigue, sleep disturbance, and depression, 1.62 T-scores ($p=.01$) lower on positive affect, and 0.15 T-scores higher on anger ($p=.85$) than participants without a diagnosis. Adjustment for numeric frequency reports magnified the differences between the medical status groups by varying degrees; the group difference increased by 0.34 (fatigue), 0.02 (sleep disturbance), 0.93 (anger), 1.18 (depression), and 0.02 (positive affect) T-scores after conditioning on propensity scores. Turning now to the age groups, participants in the older age group had more favorable QoL scores than younger participants on all domains, in magnitudes ranging from 2.57 to 7.31 T-scores; adjustment for numeric frequency responses did not noticeably affect the age-group differences.

When QoL scores for each domain were computed directly from the numeric frequency reports, the resulting standardized medical and age group differences were similar in magnitude to those obtained from PROMIS measures. For the comparison between medical status groups, effect sizes ranged from $d = 0.02$ (anger) to $d = 0.37$ (fatigue) for PROMIS scores and from $d = 0.06$ (anger) to $d = 0.44$ (fatigue) for numeric frequency QoL scores, with no significant differences in effect sizes between the reporting formats (p >.08 for all QoL domains). For age comparisons, numeric frequencies yielded a significantly ($p<.01$) larger effect ($d = 0.59$) than PROMIS scores ($d = 0.44$) for depression (with older people scoring lower than younger people); age effects did not differ between the reporting formats for other QoL domains (ps >.07).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Discussion

Repeated findings of favorable QoL among chronically ill and older populations have spurred debates among researchers and there is a lingering concern that such results may be a methodological artifact that is due to differences in the interpretation and "recalibration" of the response scale. In this study, we examined the hypothesis that group differences in the meaning of vaguely quantified response scales might obscure the impact of chronic illness and age on self-reported QoL.

We start with the results for the medical groups, whose results were partially in line with this hypothesis. Specifically, people with a medical condition assigned higher numeric frequency values to the same vague quantifiers for *negatively* framed QoL experiences than those without a condition, a finding that replicated across several medical diagnoses. This pattern of results is consistent with the notion that having experienced a greater number (or severity) of symptoms over time has shifted or recalibrated the use of vague quantifiers in people with a medical history relative to those without a condition [13]. Of course, we cannot confirm this explanation with the present data.

On the other hand, for *positively* framed QoL experiences, we found that respondents with and without a diagnosis associated the vague quantifiers with similar frequencies. We do not have an explanation for this finding. One possibility is that presumably opposite poles of a content dimension – for example, fatigue (negative) and energy (positive) – do not behave symmetrically with regard to scale recalibration, maybe because they differ in the cognitive processes involved in mapping subjective information onto the response scale [2].

It is important to note, however, that negatively keyed items are especially common in QoL measures, which tend to focus on patients' symptoms and health impairments. A possible implication of this finding is that including both positive and negative items in QoL measures might mitigate the effect of differences in the interpretation of vague quantifiers across medical groups. In fact, when QoL scores were adjusted for the differences in frequency responses in the current study, the impact of the adjustment was negligible for fatigue and sleep disturbance, which included both positive and negative QoL items, whereas it was more pronounced for anger and depression scores, which included only negative items.

While our results are in line with the idea that the use of vague quantify response scales underestimate the negative impact of chronic illness on QoL, the magnitude of this effect was very small for most QoL domains. Compared to unadjusted scores, adjustment for differences in frequency responses increased the group level average differences by 0.9 T-scores for anger, 1.2 T-scores for depression, and less for other QoL domains. These effects on group level differences do not exceed thresholds of minimally important differences established for PROMIS, which range from 2.5 to 6.0 T-scores [37].

Results for the comparison of age groups yielded a pattern suggesting that relative to younger respondents, older respondents used vague quantifiers at the low end of the scale as indicating lower numeric frequencies and those at the high end of the scale with somewhat higher numeric frequencies (for positive items). We did not find that adjusting scores for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

differential use of vague quantifiers affected age differences in QoL, possibly because the effects at the high and low ends of the scale canceled each other out. Consistent with prior research, older respondents reported less fatigue [38], better subjective sleep quality [39], less frequent anger and depression [8], and more positive affect [40], and adjustments for the frequency responses did not change these results. Nevertheless, biases may occur in studies examining age-effects in specific subpopulations, especially when comparing groups who are either very high or very low on the spectrum of QoL experiences (that is, where discrepancies in vague quantifier use would not be counterbalanced by others at the opposite end of the scale).

This study has several limitations. The medical conditions in this study were based on self-reports; while evidence suggests that people are reasonably accurate reporters of medical diagnosis [41,42], it would have been ideal to obtain confirmation of the diagnoses from clinicians or medical tests. An additional limitation is the cross-sectional study. Of interest, though, is the possibility of using the current approach to examine shifts in the meaning of vague quantifiers over time, with changing age or changing medical status. Furthermore, due to the retrospective nature of the reports, both the vague quantifier ratings and numeric frequency reports may have been affected by memory biases, and we cannot say that either report provided an ecologically valid assessment of people's actual QoL experiences [25]. A potentially promising alternative is to ask people to reconstruct their experiences using the Day Reconstruction Method [43] or event history calendar techniques [44], which provides an assessment of numeric frequency counts while presumably improving recall accuracy [1,43,45]. In addition, this study was limited to the examination of frequency response scales. Notably, PROMIS measures often combine several different response formats in the same instrument [27], and the present findings may not generalize to other types of response scales that are commonly used in practice, such as intensity or severity ratings, agree/disagree responses, amounts (e.g., few—a lot), or ratings of the difficulty of physical activities. The use of only 2 to 4 items per QoL domain further limits definite conclusions about the impact of vague quantifiers on PROMIS scale scores, and the results may not generalize to other items from the larger PROMIS item banks [28]. Finally, while this study found different effects for positive and negative items, future research should examine additional distinctions (e.g. items tapping behaviors vs. physical or emotional symptoms) to gain a more fine-grained understanding of the impact of vague quantifiers in QoL research.

Category rating scales with vaguely quantified response options are ubiquitous in QoL research and practice, and it is timely to consider methods to reduce potential biases and response ambiguities associated with vague quantifiers. Modern (item response theory) test construction approaches have led to substantial improvements in QoL measures with category rating scales such as those developed by PROMIS [28], including the development of instruments that are calibrated based on item response functions, and the possibility to detect and recalibrate specific items that show "differential item functioning" and are not interpreted in the same way across different groups [46,47]. Additionally, techniques involving anchoring vignettes [48,49], modulus based assessments [50], and joint evaluation procedures [51], have been suggested to reduce incomparabilities in the use of response scales. Pending further research on this topic, open-ended numeric frequency reports may fruitfully expand the repertoire of diagnostic tools that are available to identify whether

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

people use response options differently by elucidating individuals' subjective meaning of vague quantifiers. An important avenue for future research will be to compare the benefits of these different approaches [see 47,52] and their ability to facilitate unbiased comparisons of QoL between groups and individuals on the same metric.

Acknowledgments

We would like to thank Joan Broderick, PhD, Doerte Junghaenel, PhD, and Alicia Bolton, PhD, for helpful discussions in preparation of this manuscript.

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

References

1. Schwarz N, Oyserman D. Asking questions about behavior: Cognition, communication, and questionnaire construction. American Journal of Evaluation. 2001; 22:127–160.
2. Tourangeau, R., Rips, LJ., Rasinski, K. The psychology of survey response. Cambridge: Cambridge University Press; 2000.
3. Volkman, J. Scales of judgment and their implications for social psychology. In: Rohrer, JH., Sherif, M., editors. Social Psychology at the Crossroads. New York: Harper; 1951. p. 273-294.
4. Hakel MD. How often is often? American Psychologist. 1968; 23:533–534. [PubMed: 5660355]
5. Ahmed S, Mayo NE, Corbiere M, Wood-Dauphinee S, Hanley J, Cohen R. Change in quality of life of people with stroke over time: True change or response shift? Quality of Life Research. 2005; 14:611–627. [PubMed: 16022056]
6. Andrykowski MA, Hunt JW. Positive psychosocial adjustment in potential bone marrow transplant recipients: Cancer as a psychosocial transition. Psycho-Oncology. 1993; 2:261–276.
7. Sieff EM, Dawes RM, Loewenstein G. Anticipated versus actual reaction to HIV test results. The American journal of psychology. 1999; 112:297–311. [PubMed: 10696276]
8. Carstensen LL, Pasupathi M, Mayr U, Nesselroade JR. Emotional experience in everyday life across the adult life span. Journal of Personality and Social Psychology. 2000; 79:644–655. [PubMed: 11045744]
9. Stone AA, Schwartz JE, Broderick JE, Deaton A. A snapshot of the age distribution of psychological well-being in the United States. Proc Natl Acad Sci U S A. 2010; 107:9985–9990. [PubMed: 20479218]
10. Lacey HP, Fagerlin A, Loewenstein G, Smith DM, Riis J, Ubel PA. Are they really that happy? Exploring scale recalibration in estimates of well-being. Health Psychology. 2008; 27:669–675. [PubMed: 19025261]
11. Schwarz, N. Measurement: Aging and the psychology of self-report. In: Carstensen, LL., Hartel, CR., editors. When I'm 64. Washington, D.C: The National Academies Press; 2006. p. 219-230.
12. Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. Quality of life research. 2003; 12:599–607. [PubMed: 14516169]
13. Schwartz CE, Andresen EM, Nosek MA, Krahn GL. Response Shift Theory: Important Implications for Measuring Quality of Life in People With Disability. Archives of Physical Medicine and Rehabilitation. 2007; 88:529–536. [PubMed: 17398257]
14. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Social science & medicine. 1999; 48:1507–1515. [PubMed: 10400253]
15. Kaplan G, Baron-Epel O. What lies behind the subjective evaluation of health status? Social science & medicine. 2003; 56:1669–1676. [PubMed: 12639584]
16. Gibbons FX. Social comparison as a mediator of response shift. Social science & medicine. 1999; 48:1517–1530. [PubMed: 10400254]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

17. Carver CS, Scheier MF. Scaling back goals and recalibration of the affect system are processes in normal adaptive self-regulation: understanding 'response shift'phenomena. Social science & medicine. 2000; 50:1715–1722. [PubMed: 10798327]
18. Daltroy LH, Larson MG, Eaton HM, Phillips CB, Liang MH. Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. Social science & medicine. 1999; 48:1549–1561. [PubMed: 10400256]
19. Ubel PA, Jankovic A, Smith D, Langa KM, Fagerlin A. What is perfect health to an 85-year-old?: Evidence for scale recalibration in subjective health ratings. Medical care. 2005; 43:1054–1057. [PubMed: 16166876]
20. Biernat M, Manis M, Nelson TE. Stereotypes and Standards of Judgment. Journal of Personality and Social Psychology. 1991; 60:485–499.
21. Bartoshuk LM, Fast K, Snyder DJ. Differences in our sensory worlds: invalid comparisons with labeled scales. Current Directions in Psychological Science. 2005; 14:122–125.
22. Knäuper B, Schwarz N, Park D. Frequency Reports Across Age Groups. Journal of Official Statistics. 2004; 20:91–96.
23. Wänke M. Conversational norms and the interpretation of vague quantifiers. Applied Cognitive Psychology. 2002; 16:301–307.
24. Wright DB, Gaskell GD, O'Muircheartaigh CA. How much is 'Quite a bit'? Mapping between numerical values and vague quantifiers. Applied Cognitive Psychology. 1994; 8:479–496.
25. Schneider S, Stone AA. Ambulatory and diary methods can facilitate the measurement of patient-reported outcomes. Quality of life research. 2016; 25:497–506. [PubMed: 26101141]
26. Christodoulou C, Junghaenel DU, DeWalt DA, Rothrock N, Stone AA. Cognitive interviewing in the evaluation of fatigue items: results from the patient-reported outcomes measurement information system (PROMIS). Quality of life research. 2008; 17:1239–1246. [PubMed: 18850327]
27. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Medical care. 2007; 45:S12–S21. [PubMed: 17443114]
28. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol. 2010; 63:1179–1194. [PubMed: 20685078]
29. Cook KF, Bamer AM, Amtmann D, Molton IR, Jensen MP. Six patient-reported outcome measurement information system short form measures have negligible age-or diagnosis-related differential item functioning in individuals with disabilities. Archives of physical medicine and rehabilitation. 2012; 93:1289–1291. [PubMed: 22386213]
30. Gershon RC, Lai JS, Bode R, Choi S, Moy C, Bleck T, et al. Neuro-QOL: quality of life item banks for adults with neurological disorders: item development and calibrations based upon clinical and general population testing. Quality of life research. 2012; 21:475–486. [PubMed: 21874314]
31. Rothrock NE, Hays RD, Spritzer K, Yount SE, Riley W, Cella D. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). Journal of clinical epidemiology. 2010; 63:1195–1204. [PubMed: 20688471]
32. Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychological bulletin. 1995; 118:392–404. [PubMed: 7501743]
33. Fahrmeir, L., Tutz, G. Multivariate statistical modelling based on generalized linear models. New York: Springer; 2013.
34. Nagelkerke NJ. A note on a general definition of the coefficient of determination. Biometrika. 1991; 78:691–692.
35. Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological methods. 2010; 15:234–249. [PubMed: 20822250]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

36. United States Census Bureau. [Accessed 22 February 2016] Current Population Survey Data on Educational Attainment. 2014. http://www.census.gov/hhes/socdemo/education/data/cps/index.html
37. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six PROMIS-Cancer scales in advanced-stage cancer patients. Journal of clinical epidemiology. 2011; 64:507–516. [PubMed: 21447427]
38. Junghaenel DU, Christodoulou C, Lai JS, Stone AA. Demographic correlates of fatigue in the US general population: Results from the patient-reported outcomes measurement information system (PROMIS) initiative. Journal of psychosomatic research. 2011; 71:117–123. [PubMed: 21843744]
39. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry research. 1989; 28:193–213. [PubMed: 2748771]
40. Schneider S, Stone AA. Mixed Emotions Across the Adult Life Span in the United States. Psychology and aging. 2015; 30:369–382. [PubMed: 25894487]
41. Barlow J, Turner A, Wright C. Comparison of clinical and self-reported diagnoses for participants on a community-based arthritis self-management programme. Rheumatology. 1998; 37:985–987.
42. El Fakiri F, Bruijnzeels MA, Hoes AW. No evidence for marked ethnic differences in accuracy of self-reported diabetes, hypertension, and hypercholesterolemia. Journal of clinical epidemiology. 2007; 60:1271–1279. [PubMed: 17998082]
43. Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA. A survey method for characterizing daily life experience: The day reconstruction method. Science. 2004; 306:1776–1780. [PubMed: 15576620]
44. Belli RF. The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. Memory. 1998; 6:383–406. [PubMed: 9829098]
45. Schneider S, Stone AA. Distinguishing between frequency and intensity of health-related symptoms from diary assessments. Journal of psychosomatic research. 2014; 77:205–212. [PubMed: 25149030]
46. Keding A, Böhnke JR, Croudace TJ, Richmond SJ, MacPherson H. Validity of single item responses to short message service texts to monitor depression: an mHealth sub-study of the UK ACUDep trial. BMC medical research methodology. 2015; 15:56. [PubMed: 26224088]
47. Cameron IM, Scott NW, Adler M, Reid IC. A comparison of three methods of assessing differential item functioning (DIF) in the Hospital Anxiety Depression Scale: ordinal logistic regression, Rasch analysis and the Mantel chi-square procedure. Quality of life research. 2014; 23:2883–2888. [PubMed: 24848597]
48. Crane M, Rissel C, Greaves S, Gebel K. Correcting bias in self-rated quality of life: an application of anchoring vignettes and ordinal regression models to better understand QoL differences across commuting modes. Quality of life research. 2016; 25:257–266. [PubMed: 26254800]
49. Salomon JA, Tandon A, Murray CJ. Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. Bmj. 2004; 328:258. [PubMed: 14742348]
50. Hsee CK, Tang JN. Sun and water: on a modulus-based measurement of happiness. Emotion. 2007; 7:213. [PubMed: 17352577]
51. Lacey HP, Loewenstein G, Ubel PA. Compared to what? A joint evaluation method for assessing quality of life. Quality of life research. 2011; 20:1169–1177. [PubMed: 21293930]
52. Visser MR, Oort FJ, Sprangers MA. Methods to detect response shift in quality of life data: a convergent validity study. Quality of life research. 2005; 14:629–639. [PubMed: 16022057]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

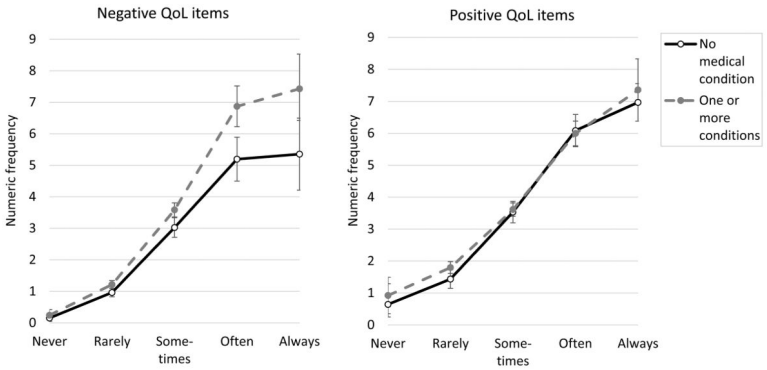

Figure 1.

Predicted numeric frequencies by response category for individuals with and without a medical condition. Error bars represent 95% confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

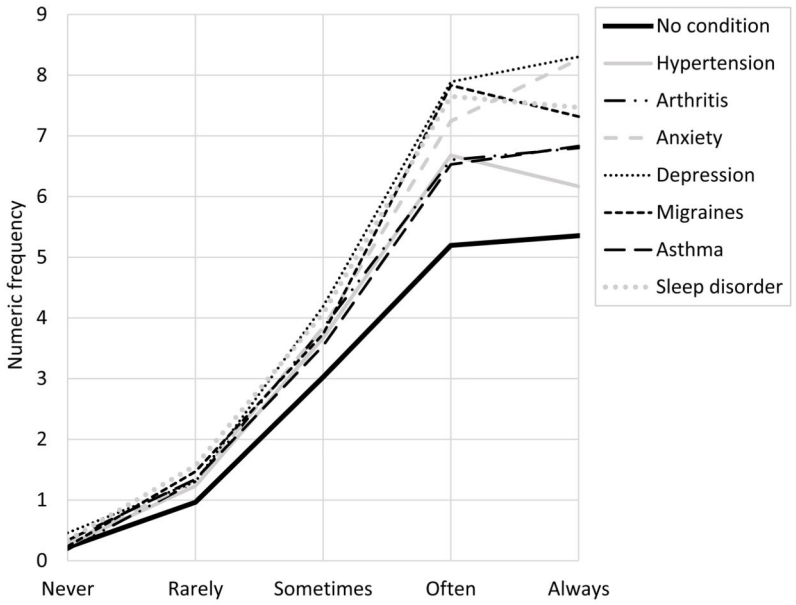

Figure 2.
Predicted numeric frequencies by response category for disease groups (negative QoL items).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

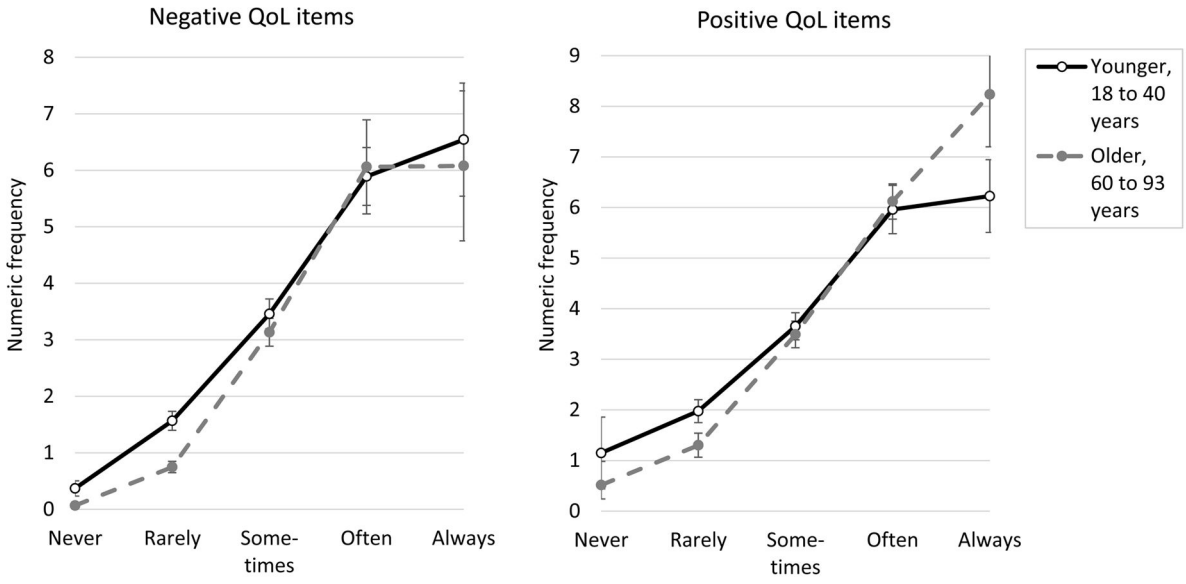

Figure 3.
Predicted numeric frequencies by response category for younger and older age groups. Error bars represent 95% confidence intervals.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Demographic and medical characteristics of study participants (N = 600)

Characteristics	N	%
Education		
Less than high school degree	26	4.4
High school graduate or GED	156	26.0
Some college, technical school, associate degree	208	34.7
College degree /advanced degree	210	35.0
Race		
White	474	79.0
African American	43	7.2
Native American/Alaska Native	5	0.8
Native Hawaiian/Pacific Islander	3	0.5
Asian	29	4.8
Other/multiple	46	7.6
Ethnicity		
Hispanic	86	14.3
Marital status		
Never married	171	28.5
Married/living with partner	320	53.3
Separated /divorced	58	9.7
Widowed	51	8.5
Gender		
Female	300	50.0
Age		
Mean (*SD*)	48.3 (21.8)	
Range	18 – 93	
Number of chronic conditions		
0	158	26.3
1	114	19.0
2	134	22.3
3	76	12.7
4	41	6.8
5 or more	77	12.8
Type of chronic condition		
Hypertension	229	38.2
Arthritis	181	30.2
Anxiety	152	25.3
Depression	130	21.7
Migraines	108	18.0
Asthma	94	15.7
Sleep disorder	91	15.2

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Characteristics	N	%
Diabetes	77	12.8
COPD, chronic bronchitis, emphysema	47	7.8
Angina	45	7.5
Cancer	36	6.0
Stroke, transient ischemic attack	25	4.2
Spinal cord injury	24	4.0
Coronary artery disease	22	3.7
Heart failure, congestive heart failure	22	3.7
Liver disease, hepatitis, cirrhosis	20	3.3

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Descriptive statistics of numeric frequency counts by vague quantifier for each item

	Overall			Means (*SD*s) by vague quantifier response category				
In the past 7 days, how many times…	**Mean (*SD*)**	**Skew**	**Range**	**Never**	**Rarely**	**Sometimes**	**Often**	**Always**
Negative items								
…did you feel run down?	3.29 (3.28)	2.35	0–25	0.21 (0.62)	1.43 (1.52)	3.44 (2.45)	6.63 (3.74)	7.40 (4.61)
…did you have to push yourself to get things done because of your fatigue?	3.10 (3.31)	2.42	0–25	0.41 (1.02)	1.46 (2.45)	3.49 (2.09)	6.01 (4.08)	6.30 (3.30)
…did you have trouble sleeping?	2.92 (2.68)	1.20	0–20	0.17 (0.60)	1.16 (1.22)	3.31 (1.87)	5.20 (1.75)	6.67 (3.43)
…did you have trouble staying asleep?	2.74 (2.83)	1.64	0–21	0.27 (0.96)	1.19 (1.38)	3.33 (1.88)	5.49 (2.53)	7.15 (3.52)
…did you feel angry?	2.41 (3.31)	3.33	0–30	0.20 (0.87)	1.13 (1.29)	3.74 (3.30)	7.25 (5.12)	4.00 (2.92)
…did you feel like yelling at someone?	2.52 (5.34)	8.26	0–85	0.18 (0.91)	1.16 (2.06)	3.55 (2.39)	9.40 (11.64)	9.89 (11.03)
…did you feel depressed?	2.24 (3.65)	4.24	0–40	0.14 (0.66)	1.04 (1.03)	3.73 (2.59)	6.94 (5.96)	9.56 (6.36)
…did you feel lonely?	2.28 (3.63)	4.20	0–40	0.07 (0.35)	0.82 (0.95)	3.23 (2.47)	6.16 (5.37)	8.83 (5.06)
Positive items								
…did you have enough energy to exercise strenuously?	2.77 (2.96)	2.41	0–28	0.53 (1.72)	1.41 (1.91)	2.76 (1.57)	4.62 (1.95)	7.18 (5.55)
…were you energetic?	4.43 (5.81)	9.43	0–99	0.90 (1.97)	1.82 (1.74)	3.64 (2.31)	6.96 (9.08)	7.93 (6.69)
…did you get enough sleep?	4.18 (2.51)	0.50	0–20	1.00 (1.96)	1.83 (1.48)	3.68 (1.76)	5.68 (2.00)	6.42 (1.77)
…was it easy for you to fall asleep?	4.24 (2.81)	1.60	0–32	0.63 (1.41)	1.90 (3.10)	3.63 (1.95)	5.84 (1.41)	6.62 (1.91)
…did you feel cheerful?	5.47 (4.61)	5.15	0–50	2.20 (2.86)	2.30 (1.66)	4.14 (3.71)	6.86 (4.56)	7.89 (5.86)
…did you feel hopeful?	5.49 (5.78)	5.76	0–64	1.30 (2.26)	1.63 (1.46)	3.93 (3.85)	6.54 (4.87)	9.12 (8.94)

Qual Life Res. Author manuscript; available in PMC 2017 October 01.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Results of negative binomial model predicting numeric frequency responses

	Negative QoL items				Positive QoL items			
	df	χ^2	R^2_{pseudo}	IRR	*df*	χ^2	R^2_{pseudo}	IRR
Control variables								
Item	7	30.5***	.012		5	31.7***	.075	
Vague quantifier category	4	1109.7***	.591		4	727.5***	.444	
Item X vague quantifier	28	103.0***	.602		20	37.2*	.451	
Focal predictors								
Medical status group	1	22.2***	.610		1	1.1	.451	
Age group	1	53.8***	.619		1	2.9	.451	
Medical status X vague quantifier	4	8.9	.621		4	3.9	.452	
Never				1.15				1.35
Rarely				1.26*				1.25
Sometimes				1.19**				1.02
Often				1.32***				0.99
Always				1.39*				1.06
Age group X vague quantifier	4	74.5***	.635		4	26.9***	.464	
Never				0.18***				0.45
Rarely				0.48***				0.66***
Sometimes				0.91				0.96
Often				1.03				1.03
Always				0.93				1.32***

Notes: R^2_{pseudo} = Nagelkerke Pseudo-R^2. IRR= incidence rate ratio for the difference between groups by vague quantifier category; values exceeding 1.0 imply higher incidence rates for people with a medical condition compared to people without a condition, and higher incidence rates for older than for younger people.

*$p < .05$;

**$p < .01$;

***$p < .001$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Group differences in fatigue, sleep disturbance, and emotional wellbeing (T-scores) before and after adjustment for numeric frequency responses

	Fatigue	Sleep disturbance	Anger	Depression	Positive affect
Medical status					
No medical condition	50.33	49.29	52.08	50.82	52.08
With medical condition	53.31	52.26	52.23	53.43	50.46
Difference, unadjusted	2.98***	2.96***	0.15	2.61**	−1.62*
Difference, adjusted [a]	3.32***	2.98***	1.08	3.79**	−1.64*
Age					
Younger age-group	54.15	53.51	55.85	54.69	49.60
Older age-group	50.90	49.45	48.54	50.80	52.60
Difference, unadjusted	−3.25***	−4.05***	−7.31***	−3.89***	2.57***
Difference, adjusted [a]	−3.25***	−3.91***	−7.27***	−3.92***	2.58***

[a]Group difference after propensity score adjustment.

*$p < .05$;

**$p < .01$;

***$p < .001$.