Quality of Life Research (2019) 28:1585–1594
https://doi.org/10.1007/s11136-019-02130-y

Patient versus proxy response on global health scales: no meaningful DIFference

Brittany R. Lapin[1,2] · Nicolas R. Thompson[1,2] · Andrew Schuster[2] · Irene L. Katzan[2]

Accepted: 6 February 2019 / Published online: 13 February 2019

Abstract

Purpose Assessment of outcomes from a proxy is often substituted for the patient's self-report when the patient is unable or unwilling to report their status. Research has indicated that proxies over-report symptoms on the patient's behalf. This study aimed to quantify the extent of proxy-introduced bias on the Patient-Reported Outcomes Measurement Information System Global Health (PROMIS GH) scale for mental (GMH) and physical (GPH) scores.

Methods This retrospective cohort study included incident stroke patients seen in a cerebrovascular clinic who completed PROMIS GH between 10/12/15 and 6/6/18. Differential item functioning (DIF) evaluated measurement invariance of patient versus proxy responses. DIF impact was assessed by comparing the initial score to the DIF-adjusted score. Subgroup analyses evaluated DIF within strata of stroke severity, measured by modified Rankin Scale (≤ 1, 2, 3+), and time since stroke (≤ 30, 31–90, > 90 days).

Results Of 1351 stroke patients (age 60.5 ± 14.9, 45.1% female), proxy help completing PROMIS GH was required by 406 patients (30.1%). Proxies indicated significantly worse response to all items. No items for GMH or GPH were identified as having meaningful DIF. In subgroup analyses, no DIF was found by severity or 31–90 days post-stroke. In patients within 30 and > 90 days of stroke, DIF was detected for 2 items. Accounting for DIF had negligible effects on scores.

Conclusions Our findings revealed the overestimation of symptoms by proxies is a real difference and not the result of measurement non-invariance. PROMIS GH items do not perform differently or have spuriously inflated severity estimates when administered to proxies instead of patients.

Keywords Patient-reported outcomes · Proxy · Differential item functioning · PROMIS GH

Introduction

Self-reported health status measures are increasingly utilized to assess outcomes following stroke. As many as 25% of people who have a stroke may be unable to report their status due to language and cognitive impairments [1, 2]. In these cases, assessment of outcomes from a proxy, such as a caregiver or family member, is often substituted for the patient's self-report. Research has demonstrated that proxies tend to rate patient outcomes worse than patients rate their own outcomes [1, 3, 4]. This is particularly pronounced for the more subjective domains like emotional well-being, social functioning, and pain [5–9]. Additionally, agreement between patient and proxy responses has been shown to vary according to demographics and clinical characteristics [10]. Results from a large study of the National Health Interview Survey indicated that proxy-completed responses under-reported disabilities for those aged 18 to 64 years, but over-reported disabilities for those 65 and older [8]. Similar findings have been demonstrated in stroke patients. Studies of stroke patient–proxy validity generally show that there is greater variability in patient–proxy agreement for strokes that occurred recently and as severity of stroke increases [9, 11–15]. Due to these differences between patient and proxy response, inclusion of unbalanced numbers of proxy

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-019-02130-y) contains supplementary material, which is available to authorized users.

✉ Brittany R. Lapin
LapinB@ccf.org

[1] Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Avenue, JJ3-603, Cleveland, OH 44195, USA

[2] Center for Outcomes Research & Evaluation, Cerebrovascular Center, Neurological Institute, Cleveland Clinic, Cleveland, OH, USA

respondents in different treatment groups may bias the analyses of outcomes. Furthermore, patient–proxy disagreement may have an even greater impact on analysis of within-person change [16].

The Patient-Reported Outcome Measurement Information System (PROMIS), funded by the NIH Roadmap Initiative, is a psychometrically validated, dynamic system to measure self-reported health across many domains in patients with a wide range of diseases and demographic characteristics [17]. Although the initial focus was its use in clinical research, there has been increasing interest in using PROMIS tools in clinical care [18, 19]. The International Consortium for Health Outcomes Measurement, a non-profit organization that develops standard sets of outcomes, recommended PROMIS Global Health (GH), a 10-item global health metric, for measuring stroke outcomes [20]. Research by our group has demonstrated the validity and utility of PROMIS GH to track stroke patients' health over time [21]. Assessment of how stroke patients versus proxies respond to PROMIS GH will increase our understanding of stroke outcomes and enhance the utilization of patient-reported outcome measurements (PROMs) to improve care.

To evaluate the reliability of PROMIS GH items as indicators of health-related quality of life (HRQOL) within proxy versus patient self-responses, we utilized differential item functioning (DIF), a targeted item-response theory (IRT) methodology. An item demonstrates DIF when subgroups of respondents with equivalent levels of a symptom have different probabilities of reporting a symptom, or endorsing that item. If participant response to an item on the PROMIS GH is influenced not only by their HRQOL but also by whether or not a proxy is responding, then the item may be under- or overestimating HRQOL within that patient. We hypothesized items that are more difficult for a proxy to accurately assess, such as pain and emotional problems, would be significant sources of DIF in stroke patients. We also predicted that the cumulative effects of DIF would result in deflated/worse HRQOL scores among proxy responses versus patient responses.

This study aimed to quantify the extent of proxy-introduced bias on the PROMIS GH scale for mental and physical global health summary scores through assessing DIF in stroke patients or proxies of stroke patients. Furthermore, our study assessed DIF within subgroups known to display differing levels of patient–proxy agreement: age, stroke severity, and time since stroke.

Methods

This retrospective cohort study included incident stroke patients seen in Cleveland Clinic cerebrovascular center who completed PROMIS GH from October 12, 2015 through June 6, 2018. As part of routine care, both patient and clinician-reported scales are collected through the Knowledge Program© (KP), an electronic platform for systematic collection of patient-reported information [22]. Patients were included in the study cohort if they were 18 years and older, completed PROMIS GH during their visit, indicated whether the patient had help completing the survey (proxy respondents), and had a diagnosis of ischemic stroke or intracerebral hemorrhage (ICH) obtained from either visit diagnosis codes (International Classification of Diseases Clinical Modification codes 9th edition: 431, 433.x1, 434.x1, 433.91, and 436 and 10th edition: I61x, I63xx) and/or provider documentation of ischemic stroke or ICH in structured fields of the KP. Patients with history of subarachnoid hemorrhage were excluded.

PROMIS GH produces two summary scores: Global Mental Health (GMH) and Global Physical Health (GPH) [23]. GMH includes 4 items on overall quality of life, mental health, satisfaction with social activities and relationships, and emotional problems, whereas GPH comprises 4-items on physical health, physical functioning, pain intensity, and fatigue. Two PROMIS GH items (general health and social roles) are not used to calculate summary scores. All items, except for the rating of pain, used a 5-category response option. Pain was recoded to 5 categories in this analysis, and all categories were oriented so higher response indicated better HRQOL. Summary scores are centered on the 2000 United States Census with respect to age, sex, education, and race–ethnicity, and are transformed to a T-score metric with a mean of 50 and standard deviation of 10 [24]. A proxy respondent was assessed through the question "Did you receive help completing this questionnaire?" If yes was selected, a second question asked "Although you received help, could you have completed it on your own?" These questions did not distinguish if a proxy completed the questionnaire with input from the patient versus independently. Questionnaires were completed either by the patient or a proxy.

Patient demographics included age, race, sex, marital status, and household income estimated from 2010 census data using ZIP code. Clinicians completed the National Institutes of Health Stroke Scale (NIHSS) and modified Rankin Scale (mRS) during each visit and recorded the date of the last stroke event. The NIHSS is a 12-item measure of neurological impairment with higher scores indicating more severe deficits. The mRS is a 1-item measure of global disability ranging from 0 to 6, with 0 indicating no symptoms, 5 indicating severe disability, and 6 representing death.

Statistical analyses

Demographics, clinical characteristics, and PROMIS GH item scores and summary T-scores were compared between patient and proxy response using Chi-square test for categorical variables and t-test or non-parametric Mann–Whitney U test, as appropriate, for continuous variables. To assess the magnitude of difference between patient and proxy responses, Cohen's D effect sizes with 95% confidence intervals were calculated, where values > 0.2, > 0.5, and > 0.8 represent small, moderate, and large effects [25]. Analyses were conducted using SAS version 9.4 (SAS Institute Inc, Cary, NC).

DIF was analyzed with the R software package Lordif [26]. The Lordif package utilizes an ordinal logistic regression framework, and the graded response model is used for IRT trait estimation [27]. IRT models assume unidimensionality, or that responses to items on a scale can be explained by a single dimension, and local independence, meaning items within each measure are unrelated except for measuring the same underlying trait. The unidimensionality of PROMIS GMH and GPH has been previously established [23]. Single-factor confirmatory factor analysis was conducted to ensure the GMH and GPH were sufficiently unidimensional to appropriately use IRT methodology. The comparative fit index (CFI), Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA) were estimated, with values of ≥ 0.90 for CFI/TLI and < 0.10 for RMSEA indicating adequate model fit [28]. Local independence was evaluated to ensure item responses within each measure were independent except for measuring the same underlying trait. This was assessed by confirming item residuals displayed low correlation relative to other items in the summary measure ($r < 0.20$) [29]. Assumptions were tested using the R package lavaan [30]. The graded response model was used to calibrate item parameter estimates for GMH and GPH.

An item demonstrates uniform DIF when one subgroup consistently has a higher, or lower, probability of endorsing an item at equal levels of the trait. An item shows non-uniform DIF when one subgroup has a higher probability of endorsing the item at low levels and a lower probability at high levels. The presence of uniform and non-uniform DIF was evaluated through constructing three models. The base model included only trait level (theta, θ) to predict patient response to items on a 5-level scale. A second model included one group (patient versus proxy response) in addition to theta to predict patient response (assessing uniform DIF). The third model included an interaction effect of group (patient versus proxy response) and theta (assessing non-uniform DIF). Models were compared to identify meaningful DIF. There are several criteria for identifying DIF, with PROMIS research frequently utilizing the criterion of McFadden's pseudo R^2 change $\geq 2\%$ [31, 32]. Our study used a R^2 change of ≥ 0.02 in the beta coefficient for theta between the base model and model 2 for identifying uniform DIF and between model 2 and model 3 for identifying non-uniform DIF.

If items were identified with DIF, the impact was evaluated by first recalibrating any item found to have DIF to a 2-parameter graded response model and then item parameters were re-estimated separately across level of subgroup (patient or proxy), yielding group-specific item parameters. T-scores were then estimated based on the recalibration resulting in "DIF-corrected," or theta-adjusted, scores. The difference in scores before and after DIF correction quantified the impact of DIF on PROMIS GH summary scores. This is indicated throughout as the difference between adjusted theta and unadjusted theta. Since the detection of statistically significant DIF is largely influenced by sample size, a priori criteria were established to evaluate the clinical relevance of detected DIF. A difference between adjusted and unadjusted theta scores of at least 1 standard error of measurement (SEM) or greater than 0.3 theta units was used to identify salient DIF [33].

Subgroup analyses

Analyses at the person level were conducted to evaluate the impact of proxy respondents across subgroup. Subgroup analyses evaluated the mean difference between self- and proxy response and measurement invariance within strata of age (18–64 versus 65+), stroke severity measured by mRS (≤ 1, 2, 3+), and by time since stroke (≤ 30, 31–90, > 90 days). DIF analyses as described above were conducted within each strata.

Compliance with ethical standards

This study was approved by the Cleveland Clinic Institutional Review Board. Because the study consisted of analyses of pre-existing data, the requirement for patient informed consent was waived.

Results

Of 1351 stroke patients (mean age 60.5 ± 14.9, 45.1% female), proxy help completing PROMIS GH was required by 406 patients (30.1%), whereas 945 (69.9%) patients responded on their own (Table 1). The study did not include self- and proxy responses on the same patient. Of the patients who had proxy help completing PROMIS GH, 233 (57.4%) could have answered on their own. Proxy assistance was more often found for older patients, with 57.4% of patients ≥ 65 years having proxy-reported PROMs versus 37.2% self-reported PROMs, $p < 0.01$. Patients with proxies had significantly more disability as indicated by

Table 1 Demographics and clinical characteristics of patients with self-reported PROMs versus patients with proxy-reported PROMs, $n = 1351$

Study characteristics	Total *N* (%)	Self-reported PROMs *N* (%)	Proxy-reported PROMs *N* (%)	*P* value
Total number of patients	1351	945 (69.9)	406 (30.1)	
Demographics				
Female	609 (45.1)	419 (44.3)	190 (46.8)	0.40
Age (years), Mean ± SD	60.5 ± 14.9	58.4 ± 14.4	65.4 ± 15.1	< 0.001
65 + years	585 (43.3)	352 (37.2)	233 (57.4)	< 0.001
Race				
White	1076 (79.6)	764 (80.9)	312 (76.8)	0.24
Black	193 (14.3)	126 (13.3)	67 (16.5)	
Other	82 (6.1)	55 (5.8)	27 (6.7)	
Married	848 (64.9)	599 (65.2)	249 (64.0)	0.67
Household income (x $10 k), median (Q1, Q3)	4.97 (4.01, 6.26)	5.04 (4.18, 6.33)	4.68 (3.89, 5.96)	< 0.001
Clinical characteristics				
Ischemic stroke (vs ICH)	1162 (86.0)	836 (88.5)	326 (80.3)	< 0.001
Days since stroke, Median (Q1, Q3)	106 (40, 456)	112 (40, 476)	95 (42, 382)	0.32
0–30 days	249 (18.7)	181 (19.4)	68 (17.0)	0.023
31–90 days	388 (29.1)	251 (26.9)	137 (34.3)	
> 90 days	695 (52.2)	501 (53.7)	194 (48.6)	
Rankin score, median (Q1, Q3)	1 (0, 2)	1 (0, 2)	2 (1, 3)	< 0.001
0–1	844 (64.2)	679 (73.9)	165 (41.8)	< 0.001
2	280 (21.3)	181 (19.7)	99 (25.1)	
3+	190 (14.5)	59 (6.4)	131 (33.2)	
NIHSS, median (Q1, Q3)	0 (0, 1)	0 (0, 1)	1 (0, 3)	< 0.001

PROMs patient-reported outcome measures, *SD* standard deviation, *Q* quartile, *ICH* intracerebral hemorrhage, *NIHSS* National Institutes of Health Stroke Scale

clinician-reported mRS and NIHSS. The median time since stroke was 106 days.

Proxies indicated significantly worse response to all PROMIS GH items as compared to self-responses (Table 2). Proxy-reported GMH and GPH were 41.6 ± 8.4 and 39.8 ± 7.9, respectively, as compared to self-reported GMH and GPH scores of 47.5 ± 9.0 and 45.8 ± 9.2, $p < 0.01$ for both. Effect sizes were large for the items assessing mental health, physical function, and social roles ($d = 0.72$, 0.84, 0.77, respectively). The smallest effect size was for the item on pain ($d = 0.25$). Figure 1 depicts the mean difference between patient and proxy responses. The largest difference was for physical function and social roles (0.99 points and 0.84 points, respectively). All of the items have 95% confidence intervals greater than zero, indicating proxies report significantly worse health on all items compared to patients.

Results of the confirmatory factor analysis substantiated that a one-factor solution had acceptable goodness of fit as assessed by CFI and TLI ≥ 0.95 for GMH and GPH overall and when stratified by patient–proxy response (Supplemental Table 1). GPH also reached adequate goodness of fit based on the RMSEA of 0.07 (90% CI 0.04–0.11); however, GMH exceeded the threshold with RMSEA of 0.18 (95% CI 0.15–0.21). Local independence was established for both GMH and GPH (data available upon request).

Table 3 displays the parameter estimates and associated standard errors for PROMIS GH items by self- versus proxy response. GMH item slope estimates for self-respondents ranged from 1.39 to 3.61, indicating variation in their level of discrimination. Items on social discretionary and mental health were the most likely to differentiate between patients at different trait levels while emotional problems was the least likely. The GPH estimates for self-respondents ranged from 1.35 to 2.42, with items on physical health and function differentiating patients to the highest extent, and pain to the lowest extent. Proxy responses had lower discrimination on all items as compared to self-responses. Proxy responses to the pain item contributed the least amount of information. Supplemental Figure 1 highlights that items with greater discrimination contribute larger information, with self-respondents contributing more information than proxy respondents. Mental health and social discretionary provided the most information to GMH with emotional problems providing the least. For GPH, physical health and function provided the most information with pain providing the least, consistent across most theta levels.

Table 2 Unadjusted PROMIS GH item and summary scores by self-report versus proxy report, $n = 1351$

PROMIS GH items	Self-reported Mean ± SD	Proxy-reported Mean ± SD	Effect size of difference (95% CI)
Total number of patients	945	406	1351
1. General health	3.02 ± 0.93	2.56 ± 0.87	0.51 (0.40, 0.61)
2. Quality of life	3.33 ± 1.00	2.82 ± 0.97	0.51 (0.40, 0.62)
3. Physical health	2.88 ± 0.93	2.42 ± 0.84	0.50 (0.40, 0.61)
4. Mental health	3.35 ± 1.05	2.60 ± 1.02	0.72 (0.61, 0.83)
5. Social discretionary	3.34 ± 1.08	2.68 ± 1.02	0.62 (0.51, 0.73)
6. Physical function	3.92 ± 1.17	2.93 ± 1.21	0.84 (0.73, 0.95)
7. Pain	3.82 ± 1.05	3.56 ± 1.09	0.25 (0.14, 0.36)
8. Fatigue	3.40 ± 0.91	3.01 ± 0.90	0.43 (0.33, 0.54)
9. Social roles	3.22 ± 1.11	2.38 ± 1.02	0.77 (0.66, 0.88)
10. Emotional problems	3.49 ± 1.03	3.09 ± 1.07	0.38 (0.27, 0.49)
GMH summary score	47.5 ± 9.0	41.6 ± 8.4	0.67 (0.56, 0.78)
GPH summary score	45.8 ± 9.2	39.8 ± 7.9	0.68 (0.57, 0.79)

GMH global mental health, *GPH* global physical health; Higher scores indicate better health-related quality of life

$p < 0.001$ for all self- versus proxy-reported comparisons

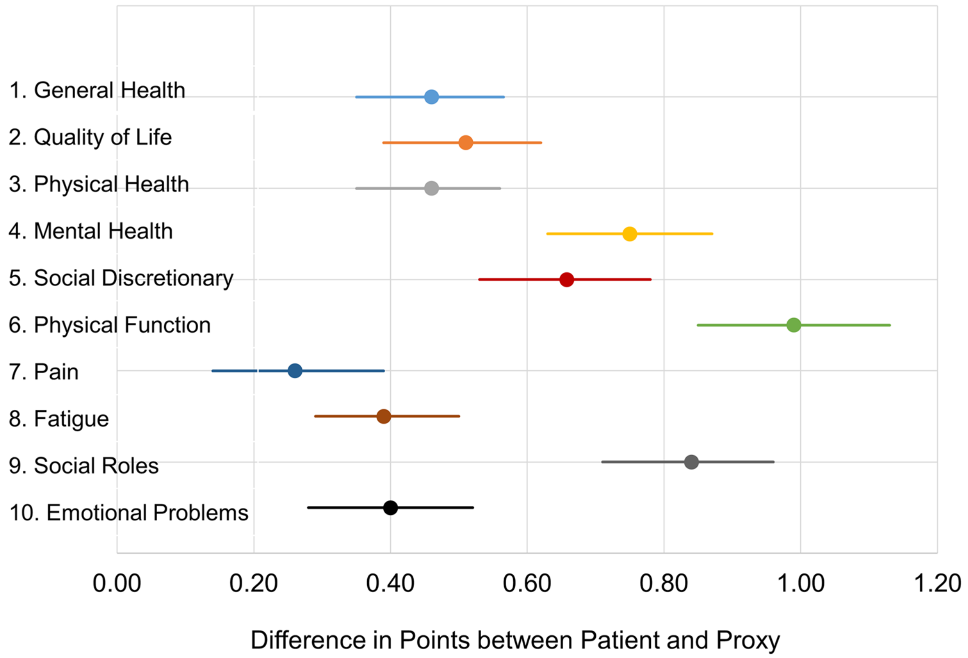

Fig. 1 Mean patient–proxy difference in PROMIS GH item scores. Horizontal bars represent 95% confidence intervals around the mean difference

The item location indicates the value of theta where the probability is 50% for endorsing that item or higher. As expected, larger locations were associated with higher categories on all items, indicating patients with higher/better levels of the trait were endorsing higher/better category responses (Table 3). For all items but pain, proxy locations were higher, meaning a higher/better trait level is required for a proxy to endorse the item as compared to a self-report. This corresponds to the theta distributions for self- versus proxy response portrayed in Supplemental Figure 2. There is broad overlap in the distributions although proxy respondents demonstrate lower/worse trait levels of GMH and GPH than self-respondents (mean theta for GMH = − 0.44 versus 0.18; mean theta for GPH = − 0.43 versus 0.18, respectively).

Differential item functioning

DIF was not detected across self- versus proxy report. The criterion of a change in pseudo $R^2 \geq 0.02$ resulted in no items

Table 3 Parameter estimates with standard errors for PROMIS GH items by self- versus proxy report

PROMIS GH items	Respondent	a	b1	b2	b3	b4
GMH						
2. Quality of life	Self	2.59 (0.16)	−2.41 (0.12)	−1.00 (0.15)	0.23 (0.12)	1.33 (0.71)
	Proxy	2.10 (0.20)	−1.70 (0.14)	−0.54 (0.14)	1.04 (0.30)	2.12 (6.12)
4. Mental health	Self	3.31 (0.25)	−2.02 (0.09)	−0.89 (0.16)	0.13 (0.11)	1.19 (1.02)
	Proxy	2.84 (0.32)	−1.30 (0.11)	−0.08 (0.11)	1.13 (0.83)	1.97 (14.4)
5. Social discretionary	Self	3.61 (0.28)	−1.83 (0.08)	−0.88 (0.15)	0.16 (0.10)	1.13 (1.18)
	Proxy	3.15 (0.37)	−1.33 (0.10)	−0.18 (0.13)	0.98 (0.67)	1.90 (17.9)
10. Emotional problems	Self	1.39 (0.09)	−3.13 (0.21)	−1.57 (0.20)	0.06 (0.14)	1.43 (0.30)
	Proxy	1.36 (0.14)	−2.60 (0.26)	−0.83 (0.20)	0.67 (0.18)	2.00 (1.10)
GPH						
3. Physical health	Self	2.42 (0.18)	−2.01 (0.11)	−0.45 (0.12)	0.85 (0.17)	2.17 (6.14)
	Proxy	2.23 (0.31)	−1.56 (0.14)	0.24 (0.09)	1.73 (2.24)	2.81 (46.3)
6. Physical function	Self	2.42 (0.19)	−2.74 (0.17)	−1.26 (0.21)	−0.47 (0.18)	0.16 (0.15)
	Proxy	1.75 (0.21)	−1.78 (0.17)	−0.34 (0.14)	0.67 (0.17)	1.51 (0.99)
7. Pain	Self	1.35 (0.10)	−4.26 (0.38)	−1.73 (0.32)	−0.60 (0.29)	0.79 (0.29)
	Proxy	0.88 (0.13)	−5.01 (0.79)	−1.82 (0.51)	−0.09 (0.41)	1.51 (0.56)
8. Fatigue	Self	1.98 (0.14)	−2.78 (0.17)	−1.42 (0.20)	0.16 (0.14)	1.63 (0.73)
	Proxy	1.49 (0.18)	−2.67 (0.28)	−0.94 (0.24)	0.95 (0.22)	2.69 (3.91)

Parameter estimates with standard error presented from graded response model; a = discrimination coefficient (slope); b_i = parameter (location) for 5 response categories minus one

on GMH or GPH being detected as having uniform DIF (R^2 change values ranged from 0.0003 to 0.0119) or non-uniform DIF (R^2 change values ranged from 0 to 0.0003) (Table 4).

Table 4 Uniform and non-uniform DIF of PROMIS GH for self- versus proxy report

PROMIS GH	Uniform[a] ΔR^2	Non-uniform[b] ΔR^2
GMH scale		
Item 2. Quality of life	0.0006	0.0000
Item 4. Mental health	0.0022	0.0001
Item 5. Social discretionary	0.0005	0.0000
Item 10. Emotional problems	0.0003	0.0003
GPH scale		
Item 3. Physical health	0.0021	0.0000
Item 6. Physical function	0.0119	0.0000
Item 7. Pain	0.0031	0.0002
Item 8. Fatigue	0.0013	0.0000
Threshold for detecting meaningful DIF	≥ 0.02	

GMH global mental health, *GPH* global physical health

[a] Model 1 (intercept + theta) versus Model 2 (Model 1 + respondent group)

[b] Model 2 (Model 1 + respondent group) versus Model 3 (Model 2 + theta * group)

Subgroup analyses

Mean differences between self- and proxy response within subgroups of interest are presented in Table 5. Patients aged 65 + years had considerably larger discrepancies in self- versus proxy response for the item assessing mental health and low differences for the pain item (0.17 (95% CI 0.01–0.34)) versus 0.44 (0.26–0.63) for younger patients, < 65 years. Average item scores were generally similar between older and younger patients; however, older patients had better scores on items assessing pain and emotional problems as compared to younger patients (mean score 3.83 ± 1.10 versus 3.67 ± 1.02, $p < 0.01$; 3.54 ± 0.98 versus 3.24 ± 1.10, $p < 0.01$, respectively), and younger patients had better scores on physical function (3.76 ± 1.24 versus 3.44 ± 1.27, $p < 0.01$). No DIF was found by self- versus proxy report for patients < 65 or 65 + years of age (Supplemental Table 2).

Average item scores were significantly worse across strata of mRS from ≤ 1, 2, to 3+ (data available upon request). However, there was less divergence between self- and proxy report as severity increased from mRS ≤ 1 to 3+ (Table 5). Despite the difference in means across strata of mRS, no DIF was found within any of the severity subgroups (Supplemental Table 3).

By time since stroke, all item scores, except for fatigue, were slightly better for patients within 30 days of their stroke. Fatigue was rated similarly across time since stroke groups. All item scores were similar between patients

Table 5 Mean Differences in PROMIS GH item scores and mean summary scores between self- and proxy report within subgroups

	Age group		mRS			Days since stroke		
	< 65 years *N* = 766	65 + years *N* = 585	0–1 *N* = 844	2 *N* = 280	3+ *N* = 190	0–30 *N* = 249	31–90 *N* = 388	> 90 *N* = 695
PROMIS GH items, mean difference (95% CI)								
1. General health	0.47 (0.31 to 0.64)	0.49 (0.34 to 0.63)	0.45 (0.29 to 0.60)	0.19 (−0.01 to 0.40)	0.05 (−0.21 to 0.30)	0.49 (0.24 to 0.74)	0.58 (0.39 to 0.77)	0.35 (0.20 to 0.51)
2. Quality of life	0.57 (0.40 to 0.75)	0.49 (0.34 to 0.65)	0.48 (0.32 to 0.65)	0.11 (−0.13 to 0.34)	0.14 (−0.16 to 0.44)	0.55 (0.28 to 0.81)	0.54 (0.33 to 0.76)	0.47 (0.31 to 0.63)
3. Physical health	0.43 (0.27 to 0.59)	0.50 (0.36 to 0.64)	0.46 (0.31 to 0.62)	0.08 (−0.12 to 0.29)	0.09 (−0.17 to 0.36)	0.63 (0.39 to 0.88)	0.53 (0.33 to 0.72)	0.34 (0.19 to 0.49)
4. Mental health	0.67 (0.49 to 0.85)	0.88 (0.71 to 1.04)	0.81 (0.64 to 0.99)	0.34 (0.09 to 0.60)	0.55 (0.21 to 0.89)	0.94 (0.69 to 1.19)	0.93 (0.70 to 1.15)	0.56 (0.38 to 0.73)
5. Social discretionary	0.66 (0.47 to 0.86)	0.70 (0.54 to 0.86)	0.66 (0.48 to 0.84)	0.20 (−0.05 to 0.45)	0.37 (0.04 to 0.71)	0.78 (0.51 to 1.06)	0.69 (0.47 to 0.91)	0.57 (0.39 to 0.75)
6. Physical function	0.89 (0.69 to 1.09)	1.03 (0.84 to 1.22)	0.82 (0.64 to 1.00)	0.26 (−0.02 to 0.54)	0.29 (−0.05 to 0.63)	0.95 (0.64 to 1.27)	0.87 (0.61 to 1.12)	1.07 (0.87 to 1.26)
7. Pain	0.44 (0.26 to 0.63)	0.17 (0.01 to 0.34)	0.18 (0.01 to 0.36)	0.07 (−0.20 to 0.34)	0.15 (−0.21 to 0.50)	0.30 (0.00 to 0.59)	0.27 (0.04 to 0.50)	0.22 (0.05 to 0.39)
8. Fatigue	0.34 (0.18 to 0.51)	0.47 (0.34 to 0.61)	0.42 (0.27 to 0.58)	0.18 (−0.04 to 0.39)	0.15 (−0.13 to 0.44)	0.47 (0.22 to 0.72)	0.44 (0.25 to 0.63)	0.34 (0.18 to 0.49)
9. Social roles	0.82 (0.63 to 1.02)	0.87 (0.71 to 1.04)	0.81 (0.63 to 0.99)	0.30 (0.07 to 0.54)	0.34 (0.01 to 0.67)	0.88 (0.59 to 1.16)	0.84 (0.61 to 1.08)	0.79 (0.61 to 0.97)
10. Emotional problems	0.42 (0.23 to 0.60)	0.53 (0.37 to 0.69)	0.44 (0.27 to 0.62)	0.06 (−0.20 to 0.33)	0.34 (0.01 to 0.66)	0.40 (0.13 to 0.66)	0.54 (0.33 to 0.76)	0.30 (0.12 to 0.48)
PROMIS GH summary scores, mean ± SD								
GMH—self	46.7 ± 9.3	48.8 ± 8.2	48.7 ± 8.8	44.4 ± 8.6	43.4 ± 8.7	49.5 ± 7.7	47.8 ± 9.2	46.4 ± 9.2
Proxy	40.8 ± 9.2	42.1 ± 7.7	42.5 ± 8.2	42.5 ± 7.9	39.9 ± 8.8	42.8 ± 7.4	41.0 ± 8.9	41.6 ± 8.3
GPH—self	45.6 ± 9.5	46.1 ± 8.5	47.4 ± 8.9	42.3 ± 8.3	38.6 ± 7.5	46.8 ± 8.7	45.5 ± 9.5	45.4 ± 9.1
Proxy	39.6 ± 8.7	39.9 ± 7.3	42.0 ± 7.8	40.6 ± 7.8	36.7 ± 7.1	40.3 ± 8.2	39.5 ± 8.0	39.9 ± 7.8

Mean differences with 95% confidence intervals presented for the difference between self- and proxy report on PROMIS GH items. Positive differences indicate worse global health for proxy-reported scores as compared to self-reported scores. Larger values indicate greater discrepancy

GMH global mental health, *GPH* global physical health

31–90 days from their stroke and > 90 days (data available upon request). For all items except physical function, similarity between self- and proxy-response was highest in patients > 90 days out from their stroke (Table 5). In patients within 30 days of stroke, uniform DIF was detected for GMH item mental health ($\Delta R^2 = 0.057$) (Supplemental Table 4). Assessment of salient DIF, or the impact of DIF, on GMH scores for patients within 30 days of their stroke was calculated as the difference between theta scores when DIF is ignored versus accounted for by exclusion of the mental health item. The difference was negligible (median difference = 0.027 on the theta scale). No DIF was found for either GMH or GPH for patients with strokes that occurred between 31 and 90 days. For patients with strokes over 90 days prior, uniform DIF was found for GPH item physical function ($\Delta R^2 = 0.059$). Assessment of DIF impact did not demonstrate salient DIF (median difference = 0.025 on the theta scale).

Discussion

Our findings revealed proxy respondents report worse mental and physical global health as compared to self-respondents; however, this overestimation of symptoms by proxies is a real difference and not the result of measurement non-invariance. While our study found proxy reports provide less ability to discriminate, lower information, and have larger location thresholds compared to self-reports, DIF was not identified in our primary study analysis. An item is considered to have DIF when item parameters are different in subgroups after controlling for the trait level. By demonstrating no DIF, we can conclude patients and proxies are interpreting the items in the same way. Since our study demonstrated no DIF in proxy responses, proxy responses can be interpreted along with patient self-responses of global health. Furthermore, the impact of the items with DIF within subgroups (age groups, severity, and time since stroke) is negligible and can be ignored when patient versus proxy comparisons are based on all items comprising PROMIS GH. Non-uniform DIF was not identified within any of our subgroup comparisons. Our study therefore supports the practical utility of PROMIS GH in clinical settings where both patients and proxies complete PROMs.

Overall, PROMIS GH scores were worse in the patient group with proxy-reported PROMIS GH than the group with self-reported PROMIS GH, even within the strata of clinician-reported disability. Although our study was unable to directly assess this, it is possible that proxies' perceptions of patients' health status are worse than the patients would have reported themselves, which has been shown in prior inter-rater agreement studies in stroke patients [3, 9, 34]. A 2010 systematic review of 13 studies of stroke patients and their proxies found proxies overestimated quality-of-life impairments compared to patient self-report, yet there was substantial-to-excellent agreement for surveys asking about activities of daily living [9]. They concluded agreement is better for more concrete, observable domains such as physical functioning and self-care than for less observable domains like emotion, pain, and social functioning [9]. In addition, domain agreement between patient and proxy responses has been found to vary according to age [8, 35, 36]. Results from a large study of the National Health Interview Survey indicated that proxy-completed responses under-reported disabilities for those aged 18 to 64 years, but over-reported disabilities for those 65 and older [8].

Another likely explanation for the worse PROMIS scores obtained by proxies compared to patient self-report seen in our study is that many of the patients with proxy completions in our study, especially older patients, had deficits that prevented them from completing patient questionnaires, presumably resulting in worse responses on PROMIS GH items than patients who completed the questionnaires themselves. This is supported by the finding that the greatest discrepancy between scores was with physical function rather than the more subjective items such as mental health, which have been shown to have greater disagreement between patient and proxy responses in studies of inter-rater agreement. Our study does not support DIF being a meaningful factor in differences in patient and proxy responses seen in these inter-rater reliability studies. In the few instances where DIF was present, it did not appreciably impact GPH summary scores.

An additional finding from our study is that for all items except physical function, similarity between self- and proxy-response was highest in patients > 90 days out from their stroke. This agrees with studies of stroke patient–proxy reliability, which have consistently shown there is more agreement as time since stroke increases [9, 11–15]. Research has indicated that as more time passes since the stroke, patients and their proxies have more time living with and observing symptoms, leading to an improvement in patient–proxy agreement [5, 12, 34, 37]. Interestingly, our study found DIF for mental health in those within 1 month of stroke, and for physical function in those > 90 days out from stroke. The impact on summary scores was negligible however, as our study examined the effect of removing DIF items on the total theta and found that removing items with DIF did not influence the effect on theta.

Our study has many strengths including the ability to investigate the differences in global health reporting across patients and proxies within a large representative cohort of stroke patients. A robust flexible hybrid method using logistic regression and IRT was utilized which enabled investigation of both uniform and non-uniform DIF. There are, however, some noteworthy limitations. We did not have self- and proxy-response data on the same patient. Patients with proxy responses instead of self-responses were more likely to be older and have more severe health. As DIF was not found in our study, the significantly worse scores demonstrated by proxies may be real and not due to measurement invariance. Worse scores by proxies could be due to the greater severity in patients who require proxy assistance and thus reflect real differences. We attempted to mitigate this effect by assessing patient–proxy differences for patients with greater severity. There were fewer discrepancies between self- and proxy-reported scores for patients with higher severity as measured by modified Rankin scale scores ≥ 3. It is possible that patients with clinician-reported mild disability had deficits that were not fully appreciated by the clinician. However, in this study, we could not directly assess whether the differences found in patient–proxy scores are true differences, and future studies are necessary to evaluate self- and proxy-response to PROMs within the same patient. Another limitation is there is no information on the proxy respondent or the relationship between the patient and proxy. Similarity between patient and proxy HRQOL domain scores may be affected by proxy characteristics, which we were unable to assess in this study. Lastly, DIF detection is highly influenced by multidimensionality. If unidimensionality of the construct is rejected, items may be flagged with DIF, even when no real DIF exists. The RMSEA for GMH exceeded the threshold for adequate model fit in our analyses so the items comprising the GMH may not reflect a unidimensional construct. Our primary study analysis did not identify meaningful DIF for any items comprising the GMH, so false positives were not an issue in our analysis.

In conclusion, our study revealed patients and proxies perceive the meaning of items on PROMIS GH consistently.

PROMIS GH items do not perform differently or have spuriously inflated severity estimates when administered to proxies instead of patients. As the number of frail, comorbid, and elderly patients who may be unable to communicate their health status increases, there will be greater reliance on proxy information. Findings from this study greatly enhance our ability to understand the impact of proxy responses on PROMs following stroke and allow more effective utilization of PROMs in clinical research and patient care.

Compliance with ethical standards

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

References

1. Williams, L. S., Bakas, T., Brizendine, E., Plue, L., Tu, W., Hendrie, H., et al. (2006). How valid are family proxy assessments of stroke patients' health-related quality of life? *Stroke, 37*(8), 2081–2085. https://doi.org/10.1161/01.STR.0000230583.10311.9f.
2. Pedersen, P. M., Jorgensen, H. S., Nakayama, H., Raaschou, H. O., & Olsen, T. S. (1995). Aphasia in acute stroke: Incidence, determinants, and recovery. *Annals of Neurology, 38*(4), 659–666. https://doi.org/10.1002/ana.410380416.
3. Duncan, P. W., Lai, S. M., Tyler, D., Perera, S., Reker, D. M., & Studenski, S. (2002). Evaluation of proxy responses to the Stroke Impact Scale. *Stroke, 33*(11), 2593–2599.
4. Epstein, A. M., Hall, J. A., Tognetti, J., Son, L. H., & Conant, L. Jr. (1989). Using proxies to evaluate quality of life. Can they provide valid information about patients' health status and satisfaction with medical care? *Medical Care, 27*(3 Suppl), S91–S98.
5. Dorman, P. J., Waddell, F., Slattery, J., Dennis, M., & Sandercock, P. (1997). Are proxy assessments of health status after stroke with the EuroQol questionnaire feasible, accurate, and unbiased? *Stroke, 28*(10), 1883–1887.
6. Kozlowski, A. J., Singh, R., Victorson, D., Miskovic, A., Lai, J. S., Harvey, R. L., et al. (2015). Agreement between responses from community-dwelling persons with stroke and their proxies on the NIH neurological quality of life (Neuro-QoL) short forms. *Archives of Physical Medicine and Rehabilitation, 96*(11), 1986–1992 e1914. https://doi.org/10.1016/j.apmr.2015.07.005.
7. Hays, R. D., Vickrey, B. G., Hermann, B. P., Perrine, K., Cramer, J., Meador, K., et al. (1995). Agreement between self reports and proxy reports of quality of life in epilepsy patients. *Quality of Life Research, 4*(2), 159–168.
8. Todorov, A., & Kirchner, C. (2000). Bias in proxies' reports of disability: Data from the National Health Interview Survey on disability. *American Journal of Public Health, 90*(8), 1248–1253.
9. Oczkowski, C., & O'Donnell, M. (2010). Reliability of proxy respondents for patients with stroke: A systematic review. *Journal of Stroke and Cerebrovascular Diseases, 19*(5), 410–416. https://doi.org/10.1016/j.jstrokecerebrovasdis.2009.08.002.
10. Brandon, T. G., Becker, B. D., Bevans, K. B., & Weiss, P. F. (2017). Patient-reported outcomes measurement information system tools for collecting patient-reported outcomes in children with Juvenile Arthritis. *Arthritis Care & Research, 69*(3), 393–402. https://doi.org/10.1002/acr.22937.
11. Carod-Artal, F. J., Coral, F., Trizotto, L. Stieven, D., & Moreira, M., C (2009). Self- and proxy-report agreement on the Stroke Impact Scale. *Stroke, 40*(10), 3308–3314. https://doi.org/10.1161/STROKEAHA.109.558031.
12. Pickard, A. S., Johnson, J. A., Feeny, D. H., Shuaib, A., Carriere, K. C., & Nasser, A. M. (2004). Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and Health Utilities Index. *Stroke, 35*(2), 607–612. https://doi.org/10.1161/01.STR.0000110984.91157.BD.
13. Hilari, K., Owen, S., & Farrelly, S. J. (2007). Proxy and self-report agreement on the Stroke and Aphasia Quality of Life Scale-39. *Journal of Neurology, Neurosurgery & Psychiatry, 78*(10), 1072–1075. https://doi.org/10.1136/jnnp.2006.111476.
14. Sangha, R. S., Caprio, F. Z., Askew, R., Corado, C., Bernstein, R., Curran, Y., et al. (2015). Quality of life in patients with TIA and minor ischemic stroke. *Neurology, 85*(22), 1957–1963. https://doi.org/10.1212/WNL.0000000000002164.
15. Skolarus, L. E., Sanchez, B. N., Morgenstern, L. B., Garcia, N. M., Smith, M. A., Brown, D. L., et al. (2010). Validity of proxies and correction for proxy use when evaluating social determinants of health in stroke patients. *Stroke, 41*(3), 510–515. https://doi.org/10.1161/STROKEAHA.109.571703.
16. Weinfurt, K. P., Trucco, S. M., Willke, R. J., & Schulman, K. A. (2002). Measuring agreement between patient and proxy responses to multidimensional health-related quality-of-life measures in clinical trials. An application of psychometric profile analysis. *Journal of Clinical Epidemiology, 55*(6), 608–618.
17. 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*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011.
18. Katzan, I. L., Thompson, N. R., Lapin, B., & Uchino, K. (2017). Added value of patient-reported outcome measures in stroke clinical practice. *Journal of the American Heart Association*. https://doi.org/10.1161/JAHA.116.005356.
19. Basch, E. (2014). New frontiers in patient-reported outcomes: Adverse event reporting, comparative effectiveness, and quality assessment. *Annual Review of Medicine, 65*, 307–317. https://doi.org/10.1146/annurev-med-010713-141500.
20. Salinas, J., Sprinkhuizen, S. M., Ackerson, T., Bernhardt, J., Davie, C., George, M. G., et al. (2016). An international standard set of patient-centered outcome measures after stroke. *Stroke, 47*(1), 180–186. https://doi.org/10.1161/STROKEAHA.115.010898.
21. Katzan, I. L., & Lapin, B. (2018). PROMIS GH (Patient-Reported Outcomes Measurement Information System Global Health) scale in stroke: A validation study. *Stroke, 49*(1), 147–154. https://doi.org/10.1161/STROKEAHA.117.018766.
22. Katzan, I., Speck, M., Dopler, C., Urchek, J., Bielawski, K., Dunphy, C., et al. (2011). The Knowledge Program: An innovative, comprehensive electronic data capture system and warehouse. In: AMIA Annual Symposium Proceedings, 2011, p. 683–692.
23. Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research, 18*(7), 873–880. https://doi.org/10.1007/s11136-009-9496-9.
24. 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*(11), 1169–1178. https://doi.org/10.1016/j.jclinepi.2009.11.021.
25. Husted, J. A., Cook, R. J., Farewell, V. T., & Gladman, D. D. (2000). Methods for assessing responsiveness: A critical review

and recommendations. *Journal of Clinical Epidemiology, 53*(5), 459–468.

26. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1–30.
27. Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S115–S123. https://doi.org/10.1097/01.mlr.0000245183.28384.ed.
28. Kline, R. (2011). *Principles and practice of structural equation modeling* (3rd edn.). New York: Guilford Press.
29. 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*(5 Suppl 1), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.
30. Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36.
31. Crins, M. H. P., Terwee, C. B., Ogreden, O., Schuller, W., Dekker, P., Flens, G., et al. (2019). Differential item functioning of the PROMIS physical function, pain interference, and pain behavior item banks across patients with different musculoskeletal disorders and persons from the general population. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-2087-x.
32. Hays, R. D., Calderon, J. L., Spritzer, K. L., Reise, S. P., & Paz, S. H. (2018). Differential item functioning by language on the PROMIS((R)) physical functioning items for children and adolescents. *Quality of Life Research, 27*(1), 235–247. https://doi.org/10.1007/s11136-017-1691-5.
33. Wanders, R. B., Wardenaar, K. J., Kessler, R. C., Penninx, B. W., Meijer, R. R., & de Jonge, P. (2015). Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *Journal of Psychosomatic Research, 78*(2), 130–136. https://doi.org/10.1016/j.jpsychores.2014.08.014.
34. Sneeuw, K. C., Aaronson, N. K., de Haan, R. J., & Limburg, M. (1997). Assessing quality of life after stroke. The value and limitations of proxy ratings. *Stroke, 28*(8), 1541–1549.
35. Ellis, B. H., Bannister, W. M., Cox, J. K., Fowler, B. M., Shannon, E. D., Drachman, D., et al. (2003). Utilization of the propensity score method: An exploratory comparison of proxy-completed to self-completed responses in the Medicare Health Outcomes Survey. *Health and Quality of Life Outcomes, 1*, 47. https://doi.org/10.1186/1477-7525-1-47.
36. Howland, M., Allan, K. C., Carlton, C. E., Tatsuoka, C., Smyth, K. A., & Sajatovic, M. (2017). Patient-rated versus proxy-rated cognitive and functional measures in older adults. *Patient Related Outcome Measures, 8*, 33–42. https://doi.org/10.2147/PROM.S126919.
37. Pickard, A. S., & Knight, S. J. (2005). Proxy evaluation of health-related quality of life: A conceptual framework for understanding multiple proxy perspectives. *Medical Care, 43*(5), 493–499.