Faces Scales for Anxiety and Anger:
A National Study of Measurement Properties

Lisa Grossman Liu, MD, PhD[a,*]; David Russell, PhD[b]; Meghan Reading Turchioe, PhD, RN[c]; Annie C. Myers, MA[c]; Connie M. Baker, MS[d]; Jyotishman Pathak, PhD[c]; and Ruth M. Masterson Creber, PhD, RN[c]

ABSTRACT

Importance: Faces scales are used worldwide to assess pain, but robust faces scales for anxiety and anger do not exist. These scales are urgently needed, because an estimated two-thirds of patients have difficulty reading written questionnaires.

Objective: To develop and evaluate measurement properties of faces scales to monitor two mental health symptoms in US adults (anxiety and anger) in accordance with the COnsensus-based Standards for health Measurement INstruments (COSMIN).

Methods: The development process included population identification, scale generation, and pretesting. The evaluation process included assessment of content validity, construct validity, criterion validity, test-retest reliability, and measurement error using 5 order-randomized, positively controlled online survey studies conducted between April and June 2020. We recruited national purposive samples of US adults representative on age, gender, and race. For each faces scale, participants assessed relevance, comprehensibility, and comprehensiveness (study 1, n = 300), strength-of-association (study 2, n = 300), convergent validity against the visual analog scale (VAS; study 3, n = 305), convergent validity against the Patient-Reported Outcomes Measurement Information System (PROMIS) questionnaires (study 4, n = 1,000), and test-retest reliability and measurement error (study 5, n = 853).

Results: The anxiety and anger faces scales showed high relevance (95%–96%), comprehensibility (93%–97%), comprehensiveness (94%–97%), and strength-of-association (74%–96%). We found very high agreement with the VAS (ρ = 0.94–0.95) and high agreement with PROMIS questionnaires (ρ = 0.74–0.79). Scales showed adequate test-retest reliability (intraclass correlation = 0.70–0.78) and measurement error (standard error of measurement = 1.14–1.22).

Conclusions: Faces scales to monitor anxiety and anger show adequate measurement properties, including content validity, construct validity, criterion validity, test-retest reliability, and measurement error. The recommended use is non-diagnostic monitoring of anxiety and anger, particularly when mental health is an ancillary but important outcome of treatment.

J Clin Psychiatry 2022;83(5):21m14202

To cite: Grossman Liu L, Russell D, Turchioe MR, et al. Faces scales for anxiety and anger: a national study of measurement properties. *J Clin Psychiatry.* 2022;83(5):21m14202.

To share: https://doi.org/10.4088/JCP.21m14202

[a]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, New York

[b]Department of Sociology, Appalachian State University, Boone, North Carolina

[c]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York

[d]Wong-Baker FACES Foundation, Oklahoma City, Oklahoma

**Corresponding author:* Lisa Grossman Liu, MD, PhD, College of Physicians and Surgeons, Department of Biomedical Informatics, Columbia University, 622 W 168th St PH20, New York, NY 10032 (lvg2104@cumc.columbia.edu).

Mental health symptoms, including anxiety and anger, substantially burden patients and health care systems through reduced quality of life, treatment nonadherence, and increased health care utilization.[1–3] Today, thousands of research trials[4,5] and hospitals[6] assess mental health symptoms using written questionnaires. Unfortunately, an estimated two-thirds of patients struggle to read written questionnaires.[7,8] This group may include patients with low literacy, pediatric patients, older patients, and non–English speakers. Although many questionnaires have been written at the 6th grade level,[9] the fact that 52% of US adults cannot read at the 6th grade level[7] and that 15% of the US population is younger than 6th grade[8] demonstrates the ineffectiveness and inaccessibility of written questionnaires for most US patients. Furthermore, answering several dozen questions on a daily or weekly basis is a substantial time commitment, one that many patients have difficulty sustaining regardless of their literacy level.[10–12] A more accessible and efficient means to monitor mental health symptoms is needed, especially when mental health is an ancillary rather than primary outcome.

One potential solution is visual rating scales, which use series of images to illustrate a spectrum of symptom severity. Clinicians worldwide use visual rating scales for pain assessment, especially facial expression drawings ("faces scales") such as the Wong-Baker FACES Pain Rating Scale.[13] Research suggests that patients who have difficulty reading can more easily learn to use faces scales, that patients prefer faces scales over other methods for reporting pain, and that faces scales are more sensitive than comparable methods such as the visual analog scale (VAS) or numeric rating scale, especially in populations with lower literacy.[14–19] Faces scales have been created for mental health symptoms such as anxiety and anger, but there is no robust or well-validated standard, and existing scales have not achieved widespread use.[20–23] Given the high prevalence of mental health symptoms in many illness conditions,[24–27] the lack of robust, accessible scales such as faces scales may prevent sensitive assessment and impair timely treatment.

In this study, we developed and evaluated the measurement properties of faces scales for monitoring two common mental health symptoms in US adults: anxiety and anger (Figures 1A and 1B, respectively). The measurement goal of the proposed faces scales was to conduct non-diagnostic monitoring of anxiety

Clinical Points

- In the United States, thousands of research trials and hospitals assess mental health symptoms using written questionnaires. Unfortunately, an estimated two-thirds of patients struggle to read these questionnaires.
- Faces scales may be the most effective currently available alternative for accessible, sensitive, and efficient symptom assessment, especially when mental health is a subsidiary but important outcome of treatment.
- The development and evaluation of faces scales is an important step toward achieving equitable access to symptom monitoring for patients with limited literacy, which may reduce health care disparities.

and anger in multiple clinical contexts, particularly when these symptoms are ancillary but important outcomes. The measurement properties of interest included content validity, construct validity, criterion validity, test-retest reliability, and measurement error.

METHODS

Study Design

We conducted and reported this development and evaluation in accordance with the COnsensus-based Standards for health Measurement Instruments (COSMIN) study design checklist version July 2019 (Supplementary Table 1).[28] The detailed protocol is available at https://bit.ly/vrs-protocol. The methodological process for development included population identification, scale generation, and pretesting. The methodological process for evaluation included assessment of content validity, construct validity, criterion validity, reliability, and measurement error.

Population Identification

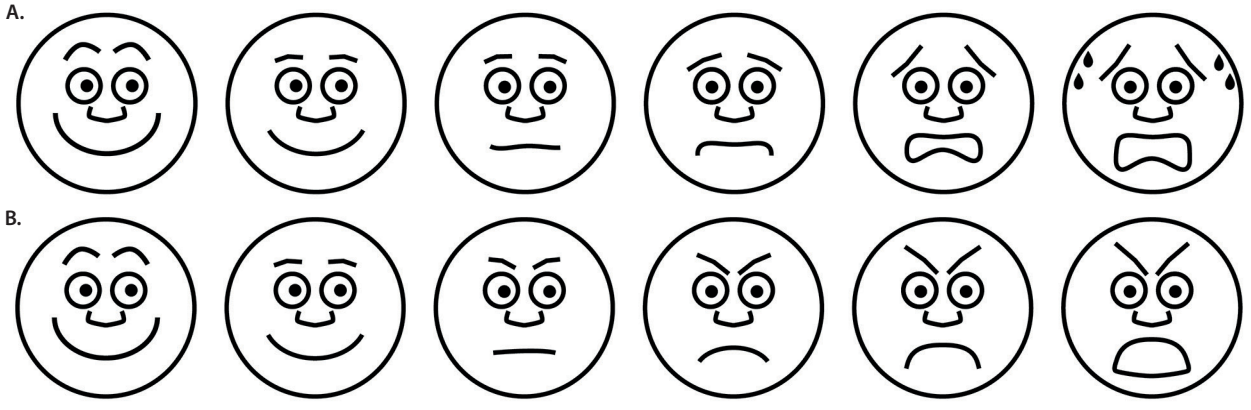
Mental health is increasingly recognized as highly influential on health outcomes and satisfaction in many clinical contexts.[29,30] Symptoms are prevalent but often unmonitored in many illness conditions.[24–27] For example, anxiety and anger greatly impact recovery after myocardial infarction, but symptoms are often unrecognized, unmonitored, and unaddressed in this at-risk population.[31,32] Therefore, the measurement goal was to conduct monitoring of anxiety and anger in multiple clinical contexts. Potential contexts may include measurement in the waiting room, at hospital admission, and through digital health technologies such as patient portals and mobile applications. Because the potential application is broad, we identified our target population as US adults, without restriction to a patient group with a specific diagnosis at this time. We anticipate future adaptation to specific populations such as pediatric patients.

Scale Generation

We assembled an expert team of two patient-reported outcome (PRO) experts, two physician experts, two patient advocacy experts, and a design expert. On the basis of a literature review and the subsequent decisions by the research team, the designer created the faces scales. The expert team searched PubMed for evaluation studies of faces scales for mental health symptoms published before March 30, 2020, using the search terms (*face* OR *fac** OR *visual**) AND (*scale* OR *rating* OR *analog**) AND (*psych** OR *mood* OR *mental*). On the basis of this literature search, the team identified potentially effective visual concepts and human facial expressions, such as those previously associated with anxiety and anger in visual association studies.[33,34] We found that faces scales have been previously developed for anxiety[20,21] and anger,[22] but these scales may not reflect real-world emotions, may have limited acceptability, may fail to capture gradual changes in symptom burden, and, most importantly, lack evaluation of measurement properties.

The team decided to improve on previous faces scales by balancing cartoon-like aspects, which may increase friendliness and generalizability, with additional details reflective of composite photographs of human facial expressions (Figures 1A and 1B). To improve learnability,

Figure 1. The (A) Anxiety and (B) Anger Rating Scales Created for This Study[a]



[a]For the Anxiety Rating Scale and the Anger Rating Scale, © 2021 Cornell University, Columbia University, and the Wong-Baker FACES Foundation. All rights reserved.

Table 1. Participant Characteristics[a]

Variable	Study 1 (n = 300)	Study 2 (n = 300)	Study 3 (n = 305)	Study 4 (n = 1,000)	Study 5[b] (n = 853)
Demographic					
Age, mean ± SD, y	46 ± 16	41 ± 15	45 ± 16	45 ± 16	46 ± 18
Female or non-binary gender	151 (50)	152 (51)	156 (51)	514 (51)	455 (53)
Race					
Asian	26 (9)	25 (8)	27 (9)	75 (8)	64 (8)
Black	45 (15)	45 (15)	45 (15)	139 (14)	114 (13)
Mixed	10 (3)	13 (4)	9 (3)	26 (3)	22 (3)
Native	8 (3)	9 (3)	6 (2)	21 (2)	17 (2)
White	211 (70)	208 (69)	218 (71)	739 (74)	636 (75)
Hispanic or Latino origin	20 (7)	28 (9)	23 (8)	57 (6)	51 (6)
Primary language is not English	19 (6)	19 (6)	19 (6)	50 (5)	42 (5)
Socioeconomic					
Education					
High school graduate or less	10 (3)	32 (11)	23 (8)	111 (11)	91 (11)
Some post–high school (eg, technical)	94 (31)	97 (32)	106 (35)	350 (35)	305 (36)
Bachelor's degree	141 (47)	102 (34)	118 (39)	356 (36)	300 (35)
Graduate degree	55 (18)	69 (23)	58 (19)	183 (18)	157 (18)
Financial resources					
Not enough	73 (24)	81 (27)	107 (35)	294 (29)	246 (29)
Enough	183 (61)	182 (61)	164 (54)	599 (60)	514 (60)
More than enough	44 (15)	37 (12)	34 (11)	107 (11)	93 (11)
Disability status					
Physical disability	20 (7)	11 (4)	27 (9)	73 (7)	65 (8)
Hearing-related disability	1 (0)	1 (0)	4 (1)	13 (1)	10 (1)
Eyesight-related disability	4 (1)	3 (1)	3 (1)	8 (1)	6 (1)
Other disability	21 (7)	22 (7)	28 (9)	86 (9)	75 (9)
No disability	261 (87)	267 (89)	253 (83)	855 (86)	725 (85)
Other					
Inadequate health literacy	89 (30)	101 (34)	89 (29)	285 (29)	224 (26)
Device type					
Desktop or laptop	246 (82)	244 (81)	232 (76)	839 (84)	716 (84)
Tablet	5 (2)	12 (4)	11 (4)	46 (5)	39 (5)
Mobile phone	49 (16)	44 (15)	62 (20)	115 (12)	98 (11)

[a]Values are shown as n (%) unless otherwise noted.
[b]Study 5 is a subsample of study 4.

visual consistency, and acceptability in clinical practice, we used the Wong-Baker visual style and lowest anchor point. The Wong-Baker scale is well-validated, and the Wong-Baker FACES Foundation website had visitors from 170 countries last year.[13] One critique of the Wong-Baker scale is that the lowest anchor point may cause overestimation of pain due to a broadly smiling expression.[15,35,36] As such, we softened the smiling expression to reduce potential overestimation. We deliberately retained the smiling expression, rather than using a neutral one, because lower literacy populations may benefit from it.[37] Additional critiques of the Wong-Baker scale include facial asymmetry and limited visibility in populations with eyesight-related disability. Therefore, we applied visual standardization techniques to improve facial symmetry and increased line thickness and color contrast to improve visibility.

Survey Procedures

Between April and June 2020, we evaluated each faces scale using multiple online surveys, available at https://bit.ly/vrs-surveys. The Wong-Baker scale was included as a control. We recruited US adults representative of the general population on age, gender, and race using an online academic research platform.[38] Each survey was reviewed by the expert team, pretested by 2 members of the general public, and revised in multiple rounds. The surveys collected information about demographic characteristics, socioeconomic characteristics, and health literacy. Demographic characteristics included age, gender, race, ethnicity, and primary language. Socioeconomic characteristics included education, financial status, and disability status. We screened for inadequate health literacy using the 3-item brief tool.[39] We recorded participants' device type (desktop, laptop, tablet, or mobile phone), because the mobile device displayed a vertical faces scale, while the other devices displayed a horizontal one. This display setting was necessary because the horizontal scale was too small to see on mobile. The research platform recruited participants purposively and monitored participants using proprietary quality checks to ensure each was human, attentive, and naive to research.[38] Participants provided informed consent and received $15 per hour in compensation. All studies had Weill Cornell Medicine Institutional Review Board approval. We computed all statistics in R Version 3.6.3.[40] We did not allow participants to skip questions, so we had no missing data.

Pretesting

We pretested each faces scale in 100 US adults. We asked that each respondent provide free-text commentary on

Table 2. Validation of the Faces Scales Against the VAS

Variable	Image (0)	Image (2)	Image (4)	Image (6)	Image (8)	Image (10)
Anxiety						
Participants, n (%)	61 (20)	41 (13)	51 (17)	49 (16)	57 (19)	46 (15)
VAS score, median (IQR)	0 (0–1)	7 (4–14)	37 (25–49)	59 (50–70)	83 (78–88)	97 (94–100)
Pseudomedian (95% CI)	5.5 (2.5–9.0)	10.0 (7.5–12.5)	37.5 (33.0–42.5)	59.5 (55.0–64.0)	82.5 (80.0–84.5)	96.5 (94.5–98.0)
Anger						
Participants, n (%)	43 (14)	51 (17)	58 (19)	39 (13)	61 (20)	53 (17)
VAS score, median (IQR)	0 (0–1)	11 (6–16)	34 (23–49)	62 (46–71)	87 (79–92)	100 (97–100)
Pseudomedian (95% CI)	1.5 (1.0–3.0)	12.0 (10.0–15.0)	35.5 (31.0–40.5)	58.5 (52.0–65.0)	85.5 (82.0–88.0)	99.0 (97.0–100.0)
Pain [Wong-Baker]						
Participants, n (%)	60 (20)	41 (13)	54 (18)	59 (19)	58 (19)	33 (11)
VAS score, median (IQR)	0 (0–1)	8 (6–14)	32 (25–45)	60 (50–69)	84 (78–89)	100 (96–100)
Pseudomedian (95% CI)	3.0 (2.0–4.0)	10.0 (7.5–14.5)	34.5 (30.0–38.5)	59.5 (55.0–63.5)	83.0 (81.0–85.5)	98.0 (97.0–100.0)

Abbreviations: IQR = interquartile range, VAS = visual analog scale.

Table 3. Validation of the Faces Scales Against the PROMIS Questionnaires

Variable	Image (0)	Image (2)	Image (4)	Image (6)	Image (8)	Image (10)
Anxiety (PROMIS raw score range, 4–20)						
Participants, n (%)	78 (8)	206 (21)	299 (30)	253 (25)	125 (13)	39 (4)
PROMIS raw score, median (IQR)	4 (4–5)	6 (5–7)	8 (7–10)	12 (10–13)	14 (12–16)	17 (14–19)
Pseudomedian (95% CI)	4.5 (4.0–4.5)	6.0 (5.5–6.0)	8.0 (8.0–8.5)	11.5 (11.0–11.5)	14.0 (13.5–14.5)	16.5 (15.0–17.0)
Anger (PROMIS raw score range, 5–25)						
Participants, n (%)	177 (18)	312 (31)	320 (32)	132 (13)	52 (5)	7 (1)
PROMIS raw score, median (IQR)	6 (5–8)	9 (7–11)	13 (10–15)	15 (14–17)	17 (15–20)	20 (18–25)
Pseudomedian (95% CI)	6.5 (6.5–7.0)	9.0 (9.0–9.5)	12.5 (12.0–13.0)	15.5 (14.5–16.0)	17.0 (16.0–18.0)	21.5 (16.5–25.0)
Pain (Wong-Baker) (PROMIS raw score range, 3–15)						
Participants, n (%)	314 (31)	276 (28)	224 (22)	132 (13)	44 (4)	10 (1)
PROMIS raw score, median (IQR)	3 (3–3)	5 (3–6)	6 (5–7)	8 (7–9)	10 (8–10)	11 (10–12)
Pseudomedian (95% CI)	3.0 (3.0–3.0)	5.0 (4.5–5.0)	6.0 (6.0–6.5)	8.0 (8.0–8.5)	9.5 (8.5–10.0)	11.0 (9.0–12.5)

Abbreviations: IQR = interquartile range, PROMIS = Patient-Reported Outcome Measurement Information System.

potential areas of improvement. The expert team reviewed all free-text commentary together. The commentary was strongly positive, and no negative themes emerged; therefore the team did not make revisions.

Content Validity

Study 1. We assessed (1) relevance of each item for the patient's experience, (2) comprehensibility of the instructions and response options, and (3) comprehensiveness of each scale for the range of possible severity. For each scale, we recruited a unique sample of 100 participants to prevent carryover effects between scales; therefore, we recruited 300 participants total. For the analysis, we calculated the percentage of participants who agreed or strongly agreed that the scale was relevant, comprehensible, and comprehensive.

Study 2. To supplement the perspectives on relevance, comprehensibility, and comprehensiveness, we tested strength-of-association. The strength-of-association test quantified how much each scale represents its intended symptom by determining what percentage of participants associate it with that symptom without instruction, which indicates what level of instruction may be required. Participants viewed each scale and provided free-text commentary on potential symptoms associated with the least severe (lowest) and most severe (highest) images. A temperature scale served as an attention check and teaching task, which primed participants to the desired content and granularity of free-text comments (eg, "cold" or "hot"). We computer-randomized the presentation order to control for potential order effects.

Table 4. Criterion Validity of the Faces Scales

Construct	AUC (95% CI)	Sensitivity[a] (95% CI)	Specificity[a] (95% CI)
Anxiety	0.877 (0.857–0.898)	0.814	0.823
Anger	0.898 (0.868–0.929)	0.821	0.846
Pain (Wong-Baker)	0.951 (0.900–1.000)	0.909	0.822

[a]Computed using the optimal threshold of 6 or higher. The 6 image categories correspond to the numbers 0, 2, 4, 6, 8, 10 per standard (see protocol at https://bit.ly/vrs-protocol).
Abbreviation: AUC = area under the curve.

We calculated that 289 participants were needed to determine an expected proportion of 0.75 of participants associating each image with the correct symptom with a precision of 0.05 and a confidence level of 95%. We increased to 300 participants because our survey platform required it to ensure sufficient randomness and representativeness. For the analysis, two researchers independently coded free-text responses as correct or incorrect (Cohen's κ: 0.85; agreement: 93%). We accepted synonyms and subconstructs as correct (eg, for *anxiety*, we accepted *nervous*, *worry*, *fear*, *terror*, *panic*, and more; see Supplementary Table 2). Conflicts were resolved by discussion. We calculated the proportion

Table 5. Test-Retest Reliability and Measurement Error of the Faces Scales

Construct	Test Mean (SD)[a,b]	Retest Mean (SD)	Cohen *d* Statistic	ICC (95% CI)	SEM[a] (95% CI)	Size[c] (n)
Anxiety	4.47 (2.46)[d]	3.67 (2.51)	0.198	0.78 (0.71–0.83)	1.14 (1.01–1.27)	159
Anger	3.13 (2.23)	3.02 (2.22)	0.081	0.70 (0.61–0.78)	1.22 (1.07–1.38)	153
Pain (Wong-Baker)	2.68 (2.48)	3.04 (2.47)	0.032	0.78 (0.74–0.82)	1.09 (1.00–1.19)	346

[a]The 6 image categories correspond to the numbers 0, 2, 4, 6, 8, 10 per standard (see protocol at https://bit.ly/vrs-protocol).
[b]Mean and standard deviation provided to aid interpretation of SEM. Please interpret with caution as these data are ordinal.
[c]Number of clinically stable participants available for calculation of ICC and SEM.
[d]Please note that the novel coronavirus (COVID-19) pandemic arrived in the United States concurrently with the test (April 2020), which may have contributed to the elevated mean anxiety score seen here.
Abbreviations: ICC = intraclass correlation, SEM = standard error of measurement.

of participants who associated each image with the intended construct.

Construct and Criterion Validity

Studies 3 and 4. We examined the expected relationship of each faces scale with one visual measure and one written measure in two respective studies. The VAS was our visual measures.[41,42] The VAS displayed two axis ticks labeled "0 (No Symptom)" and "100 (Worst Imaginable Symptom)" connected by a long horizontal x-axis (Supplementary Figure 1). The standardized questionnaires from the National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS) were our written measures.[43–45] The following PROMIS questionnaires were included for comparison: Anxiety 4a v1.0, Anger 5a v1.1, and Pain Intensity 3a v1.0. We presented the faces scales and the VAS or PROMIS on separate pages to prevent participants from comparing them, and we used computer randomization to control for potential order effects. In study 3, participants rated someone else's symptom experience to balance responses so that potential differences in distribution by category could be examined.[46] In study 4, participants rated their own symptom experience.[46]

We calculated that 305 participants each were needed to detect an expected correlation of 0.75 or higher between each faces scale and its corresponding VAS or PROMIS raw scores with a margin of error of 0.05 and a confidence level of 95%. For the analysis, we used the Spearman correlation (ρ) to measure agreement between VAS scores (study 3) or PROMIS raw scores (study 4) and each faces scale.[46] We calculated corresponding median VAS or PROMIS raw scores for each faces scale category. We also calculated the pseudomedian, which accounts for asymmetry. Additionally, we conducted bivariate subgroup analyses using Wilcoxon rank sum tests to determine whether the median VAS and PROMIS raw scores for each faces scale category differed by demographic, socioeconomic, and other characteristics. To generate receiver operating characteristic (ROC) curves, we first dichotomized PROMIS T-scores using established cutoffs for clinically significant anxiety (T-score > 60), anger (T-score > 65), and pain (T-score > 60). Using the dichotomized T-scores as our gold standard, we generated one curve for each faces scale and calculated the respective area under the curve (AUC). Additionally, we computed sensitivity and specificity for each threshold and determined the optimal threshold for each faces scale by equivalency.

Reliability and Measurement Error

Study 5. We examined test-retest reliability after 2 months in adults who previously completed the initial validation against PROMIS. Since visual information is recalled more easily than written information,[47] we chose a lengthier time interval between test and retest to prevent recall of previous answers. Participants completed the same survey at both time points. PROMIS T-scores were used to determine clinical stability on the construct measured. For each construct, participants with a retest PROMIS T-score within a standard minimally important difference of 5 of their test PROMIS T-score were considered clinically stable and included in the analysis.[48]

We calculated that 79 participants were needed to detect an expected intraclass correlation (ICC) of 0.70 with a null hypothesis (ICC) value of 0.50 and a confidence level of 95%. To account for potential clinical instability, we increased the sample size of study 4 and therefore study 5 to 1,000 participants. For the analysis, we used ICC (2-way random-effects model, consistency, single rater) to assess test-retest reliability.[49] Measurement error was expressed as the standard error of measurement (SEM).[50]

RESULTS

We recruited 300 participants in study 1, 300 in study 2, 305 in study 3, and 1,000 in study 4. Of the 1,000 potential participants, 853 (85%) completed study 5 to assess test-retest reliability and measurement error. No statistically significant differences in baseline characteristics existed between participants who did and did not complete study 5. Table 1 presents participants' characteristics by study.

Content Validity

Study 1. Results showed high relevance (anxiety: 95%; anger: 96%; Wong-Baker pain: 91%), high comprehensibility (anxiety: 93%; anger: 97%; Wong-Baker pain: 85%), and high comprehensiveness (anxiety: 97%; anger: 94%; Wong-Baker pain: 92%).

Study 2. The anxiety and anger scales had high strength-of-association (anxiety [highest]: 74% [95% CI, 69%–79%];

anxiety [lowest]: 91% [95% CI, 87%–94%]; anger [highest]: 95% [95% CI, 92%–97%]; anger [lowest]: 96% [95% CI, 93%–98%]). High strength-of-association suggests that verbal or written instruction may not be required before use. The Wong-Baker scale had low strength-of association at the highest level (Wong-Baker pain [highest]: 7% [95% CI, 4%–10%], Wong-Baker pain [lowest]: 97% [95% CI, 96%–100%]). Low strength-of-association suggests that verbal or written instruction may be required before use.

Construct and Criterion Validity

Studies 3 and 4. We found very high agreement between each faces scale and the VAS (anxiety: $\rho = 0.95$ [95% CI, 0.94–0.96]; anger: $\rho = 0.94$ [95% CI, 0.93–0.95]; Wong-Baker pain: $\rho = 0.95$ [95% CI, 0.94–0.96]). We found high agreement between each faces scale and the PROMIS questionnaires (anxiety: $\rho = 0.79$ [95% CI, 0.77–0.81]; anger: $\rho = 0.74$ [95% CI, 0.71–0.76]; Wong-Baker pain: $\rho = 0.81$ [95% CI, 0.79–0.83]). Table 2 and Table 3 display the median VAS and PROMIS raw scores, respectively, across categories. Supplementary Figure 2 and Supplementary Figure 3 display boxplots of the VAS and PROMIS raw scores across categories, respectively.

No statistically significant differences in the VAS or PROMIS raw scores by category existed based on age, gender, race, ethnicity, primary language, education, financial resources, disability status, health literacy, or device type. Lack of difference by education suggests that the faces scales are valid at varied levels of education. Lack of difference by device type suggests that faces scales can be displayed horizontally or vertically.

Faces scales had good-to-excellent accuracy for predicting moderate-to-severe anxiety, anger, and pain (anxiety: AUC = 0.877 [95% CI, 0.857–0.898]; anger: AUC = 0.898 [95% CI, 0.868–0.929]; Wong-Baker pain: AUC = 0.951 [95% CI, 0.900–1.000]). Supplementary Figure 4 displays the ROC curves. For each faces scale, the optimal threshold was 6 or higher for predicting moderate-to-severe anxiety, anger, and pain. In regard to interpreting the threshold, the faces scales use base-10 scoring per standard (see protocol: https://bit.ly/vrs-protocol); therefore, the 6 image categories correspond to the numbers 0, 2, 4, 6, 8, and 10. At the optimal threshold, the faces scales demonstrated fair sensitivity and specificity (> 80%) for detecting moderate-to-severe anxiety, anger, and pain (Table 4).

Reliability and Measurement Error

Study 5. Intraclass correlation and SEM are shown in Table 5. The anxiety and Wong-Baker pain scales had good ICC (anxiety: ICC = 0.78 [95% CI, 0.71–0.83]; Wong-Baker pain: ICC = 0.78 [95% CI, 0.74–0.82]), whereas the anger scale had moderate ICC (anger: ICC = 0.70 [95% CI, 0.61–0.78]). Faces scales showed adequate standard error of measurement (anxiety: SEM = 1.14 [95% CI, 1.01–1.27]; anger: SEM = 1.22 [95% CI, 1.07–1.38]; Wong-Baker pain: SEM = 1.09 [95% CI, 1.00–1.19]). In regard to interpreting the SEM, the faces scales use base-10 scoring per standard (see protocol: https://bit.ly/vrs-protocol); therefore, the 6 image categories correspond to the numbers 0, 2, 4, 6, 8, and 10.

DISCUSSION

An estimated two-thirds of US patients struggle to read written questionnaires, and faces scales may be an accessible and sensitive alternative.[7,8,14–19] In this study, faces scales showed adequate measurement properties for assessment of anxiety and anger in US adults. The findings support use of faces scales for non-diagnostic monitoring of anxiety and anger to ensure that more patients can easily and efficiently report their symptoms. The anxiety and anger scales demonstrated high strength-of-association, suggesting that they could potentially be used in non-verbal or very young populations. Lack of differences by education suggested that the scales are accessible at varied levels. Further evaluation of the anxiety and anger scales in specific populations, such as children, is needed.[51]

To our knowledge, this study is one of the first to follow the COSMIN study design checklist for single-item visual rating scales. In contrast with previous studies on visual rating scales, this study included patients in the development process, thereby meeting the COSMIN standards and ensuring robust content validity. The scales demonstrated high agreement with the PROMIS written questionnaires, the gold standard developed by the US National Institutes of Health for patient-reported outcome measurement. PROMIS questionnaires have been validated both in the general population and in individuals living with varied chronic conditions[43–45]; therefore, high agreement suggests that the scales could be validated in varied chronic conditions. Because COSMIN has not yet been widely used for single-item visual rating scales, adapting it for this study required some extrapolation, and we did not assess responsiveness or cross-cultural validity. Future consensus regarding adaptation of COSMIN to single-item visual rating scales will help ensure methodological standardization.

The design and evaluation of faces scales are an important step toward achieving equitable access to health care for all Americans. When written questionnaires are inaccessible, it may preclude patients' engagement with PROs[10–12] and cause intervention-generated inequity, a phenomenon in which well-intentioned interventions worsen existing health disparities rather than reduce them.[52–54] Although faces scales do not solve every problem associated with accessibility of PROs for persons with low literacy, they are a promising strategy to improve ease-of-use of PRO measures and increase representation of diverse populations in PRO datasets.

The recommended use of the faces scales is non-diagnostic monitoring of anxiety and anger, which is essential because mental health strongly impacts health outcomes in multiple chronic conditions, yet is frequently unmonitored, unrecognized, and untreated. Faces scales are not suitable for other functions, such as diagnosis or determination of

primary outcomes. Furthermore, faces scales should not substitute for human contact with the health care system or real conversations with a therapist. As with any tool intended to improve communication between patients and the health care system, faces scales should be combined with multiple approaches to ensure the best possible communication and highest level of care.

We evaluated the faces scales in a national sample of US adults consistent with our target population. We have not yet evaluated them in specific populations, such as populations with low literacy or low educational attainment, pediatric patients, older patients, persons with disabilities, and non–English speakers. We have also not evaluated the scales in clinical contexts. Doing so will be necessary in follow-up studies, and a new experimental design will be required, as administering questionnaires and recruiting online may not reach these populations.

Limitations

This study is an initial evaluation of measurement properties, intended as a foundational exploration of faces scales for mental health symptoms in US adults. Assessment of additional measurement properties is needed, including responsiveness, cross-cultural validity, and discriminative validity. Online recruitment has limitations, such as non-credible reporting, inattention to the research survey, robotic or duplicate responses, and non-naivete to research.[38] We proactively mitigated these limitations by removing participants who completed surveys too quickly (< 1%), requiring non-duplicate IP addresses, and using participants verified by our online academic research platform as human, attentive, and naive to research. Additionally, we did not assess for previous familiarity with the Wong-Baker FACES Pain Rating Scale, which may have favorably biased its evaluation results.

Submitted: August 6, 2021; accepted February 8, 2022.

Published online: August 8, 2022.

Author contributions: L.G.L., M.R.T., and R.M.M.C. conceptualized the studies. L.G.L. conceptualized and designed the scales. All authors collaboratively refined the study design and surveys. A.C.M. managed the Institutional Review Board application. L.G.L. managed recruitment, performed the analyses, and drafted the manuscript. D.R. and A.C.M. participated in the analysis. All authors participated in interpretation of the results and critical revision of the manuscript.

Relevant financial relationships: The authors report no financial or other relationship associated with this work.

Funding/support: This work was supported by the National Institute of Nursing Research (R00NR016275 and R00NR016275-05S1, PI: Dr Masterson Creber) and the National Library of Medicine (F31LM054013, PI: Dr Grossman Liu).

Role of the sponsor: The supporters had no role in the design, analysis, interpretation, or publication of this study.

Acknowledgements: We gratefully acknowledge the support and contributions of the many patients, health care providers, and research experts who participated and without whom this work would not be possible.

Additional information: The complete data are available upon request. Please contact the corresponding author.

Supplementary material: Available at Psychiatrist.com.

REFERENCES

1. Barrera TL, Norton PJ. Quality of life impairment in generalized anxiety disorder, social phobia, and panic disorder. *J Anxiety Disord*. 2009;23(8):1086–1090.
2. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. *Arch Intern Med*. 2000;160(14):2101–2107.
3. Nipp RD, El-Jawahri A, Moran SM, et al. The relationship between physical and psychological symptoms and health care utilization in hospitalized patients with advanced cancer. *Cancer*. 2017;123(23):4720–4727.
4. US Food and Drug Administration. *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims: Guidance for Industry*. US Food and Drug Administration website. https://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf. 2009. Accessed July 1, 2020.
5. Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Control Clin Trials*. 2004;25(6):535–552.
6. European Medicines Agency. *Fourth Report on the Progress of the Interaction with Patients' and Consumers' Organisations*. European Medicines Agency website. https://www.ema.europa.eu/en/documents/report/fourth-report-progress-interaction-patients-consumers-organisations-2010-results/analysis-degree-satisfaction-patients-consumers-involved-european-medicines-agency-act_en.pdf. 2011. Accessed July 1, 2020.
7. Goodman M, Finnegan R, Mohadjer L, et al. *Literacy, Numeracy and Problem Solving in Technology-Rich Environments Among U.S. Adults: Results from the Program for the International Assessment of Adult Competencies 2012*. Washington, DC. https://www.voced.edu.au/content/ngv%3A59624. 2012. Accessed July 1, 2020.
8. United States Census Bureau. 2018 Population Estimates by Age, Sex, Race and Hispanic Origin. US Census Bureau website. https://www.census.gov/newsroom/press-kits/2019/detailed-estimates.html. Published 2019. Accessed April 8, 2020.
9. PROMIS Cooperative Group. PROMIS instrument development and validation scientific standards version 2.0. HealthMeasures website. http://www.nihpromis.org/Documents/PROMISStandards_Vers2.0_Final.pdf?AspxAutoDetectCookieSupport=1. 2013. Accessed July 1, 2020.
10. Ali J, Basch E, Baumhauer J, et al. *Users' Guide to Integrating Patient-Reported Outcomes in Electronic Health Records*. Baltimore, Maryland: Johns Hopkins University; 2017.
11. Palmer MJ, Mercieca-Bebber R, King M, et al. A systematic review and development of a classification framework for factors associated with missing patient-reported outcome data. *Clin Trials*. 2018;15(1):95–106.
12. Mercieca-Bebber R, Palmer MJ, Brundage M, et al. Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (PRO) data: a systematic review. *BMJ Open*. 2016;6(6):e010938.
13. Wong-Baker FACES Foundation. Wong-Baker FACES Pain Rating Scale. Wong-Baker FACES Foundation website. https://wongbakerfaces.org/. Published 2016. Accessed April 8, 2020.
14. Wong DL, Baker CM. Pain in children: comparison of assessment scales. *Pediatr Nurs*. 1988;14(1):9–17. https://www.ncbi.nlm.nih.gov/pubmed/3344163
15. Chambers CT, Hardial J, Craig KD, et al. Faces scales for the measurement of postoperative pain intensity in children following minor surgery. *Clin J Pain*. 2005;21(3):277–285.
16. West N, Oakes L, Hinds PS, et al. Measuring pain in pediatric oncology ICU patients. *J Pediatr Oncol Nurs*. 1994;11(2):64–68, discussion 69–70.
17. Keck JF, Gerkensmeyer JE, Joyce BA, et al. Reliability and validity of the Faces and Word Descriptor Scales to measure procedural pain. *J Pediatr Nurs*. 1996;11(6):368–374.
18. Herr KA, Garand L. Assessment and measurement of pain in older adults. *Clin Geriatr Med*. 2001;17(3):457–478, vi.
19. Halfaker DA, Akeson ST, Hathcock DR, et al. Psychological Aspects of Pain. In: *Pain Procedures in Clinical Practice*. Elsevier; 2011:13–22. 10.1016/B978-1-4160-3779-8.10003-X.
20. Kuttner L, LePage T. Face scales for the assessment of pediatric pain: a critical review. *Can J Behav Sci / Rev Can des Sci du Comport*. 1989;21(2):198–209.
21. McKinley S, Madronio C. Validity of the Faces Anxiety Scale for the assessment of state anxiety in intensive care patients not receiving mechanical ventilation. *J Psychosom Res*. 2008;64(5):503–507.
22. Raccanello D, Brondino M, Pasini M. Achievement emotions in technology enhanced learning: development and validation of self-report instruments in the Italian context. *Interact Des Archit*. 2014;23(1):68–81.
23. Kontou E. Thomas SA, Lincoln NB. Psychometric properties of a revised version of the Visual Analog Mood Scales. *Clin Rehabil*.

2012;26(12):1133–1140.
24. Turner J, Kelly B. Emotional dimensions of chronic disease. *West J Med*. 2000;172(2):124–128.
25. Schalet BD, Pilkonis PA, Yu L, et al. Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. *J Clin Epidemiol*. 2016;73(10):119–127.
26. Cella D, Lai J-S, Jensen SE, et al. PROMIS fatigue item bank had clinical validity across diverse chronic conditions. *J Clin Epidemiol*. 2016;73(10):128–134.
27. Howland M, Tatsuoka C, Smyth KA, et al. Evaluating PROMIS() applied cognition items in a sample of older adults at risk for cognitive decline. *Psychiatry Res*. 2017;247(1):39–42.
28. Mokkink LB, Prinsen CA, Patrick DL, et al. COSMIN Study Design checklist for patient-reported outcome measurement instruments: Version July 2019. COSMIN website. https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf. 2019. Accessed July 1, 2020.
29. Schenker Y, Stewart A, Na B, et al. Depressive symptoms and perceived doctor-patient communication in the Heart and Soul Study. *J Gen Intern Med*. 2009;24(5):550–556.
30. Read C, Armstrong AW. Association between the mental health of patients with psoriasis and their satisfaction with physicians. *JAMA Dermatol*. 2020;156(7):754–762.
31. Sowden GL, Huffman JC. The impact of mental illness on cardiac outcomes: a review for the cardiologist. *Int J Cardiol*. 2009;132(1):30–37.
32. Bremner JD, Campanella C, Khan Z, et al. Brain correlates of mental stress-induced myocardial ischemia. *Psychosom Med*. 2018;80(6):515–525. https://www.ncbi.nlm.nih.gov/pubmed/29794945.
33. Arcia A, Suero-Tejeda N, Spiegel-Gotsch N, et al. Helping Hispanic family caregivers of persons with dementia "get the picture" about health status through tailored infographics. *Gerontologist*. 2019;59(5):e479–e489.
34. Pollak JP, Adams P, Gay G. *PAM: A Photographic Affect Meter for Frequent, In Situ Measurement of Affect*. Conf Hum Factors Comput Syst - Proc; 2011:725–734. 10.1145/1978942.1979047.
35. Bieri D, Reeve RA, Champion DG, et al. The Faces Pain Scale for the self-assessment of the severity of pain experienced by children: development, initial validation, and preliminary investigation for ratio scale properties. *Pain*. 1990;41(2):139–150.
36. Tomlinson D, von Baeyer CL, Stinson JN, et al. A systematic review of faces scales for the self-report of pain intensity in children. *Pediatrics*. 2010;126(5):e1168–e1198.
37. Stange M, Barry A, Smyth J, et al. Effects of smiley face scales on visual processing of satisfaction questions in web surveys. *Soc Sci Comput Rev*. 2018;36(6):756–766.
38. Peer E, Brandimarte L, Samat S, et al. Beyond the Turk: alternative platforms for crowdsourcing behavioral research. *J Exp Soc Psychol*. 2017;70:153–163.
39. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. *Fam Med*. 2004;36(8):588–594.
40. R Core Team. R: A Language and Environment for Statistical Computing. R-Project website. https://www.r-project.org/. 2015. Accessed July 1, 2020.
41. Gould D, Kelly D, Goldstone L, et al. Examining the validity of pressure ulcer risk assessment scales: developing and using illustrated patient simulations to collect the data: Information Point: Visual Analogue Scale. *J Clin Nurs*. 2001;10(5):706.
42. Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. *Res Nurs Health*. 1990;13(4):227–236.
43. HealthMeasures. PROMIS (Patient-Reported Outcomes Measurement Information System). https://www.healthmeasures.net/explore-measurement-systems/promis. Published 2020. Accessed April 8, 2020.
44. Cella D, Riley W, Stone A, et al; PROMIS Cooperative Group. 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(11):1179–1194.
45. Cella D, Yount S, Rothrock N, et al; PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care*. 2007;45(suppl 1):S3–S11.
46. Kim TK. Practical statistics in pain research. *Korean J Pain*. 2017;30(4):243–249.
47. Paivio A, Csapo K. Picture superiority in free recall: imagery or dual coding? *Cognit Psychol*. 1973;5(2):176–206.
48. HealthMeasures. Interpret Scores: PROMIS: Meaningful Change. HealthMeasures website. https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/meaningful-change. Published 2019. Accessed May 27, 2019.
49. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23–34.
50. Payne RW. Reliability theory and clinical psychology. *J Clin Psychol*. 1989;45(2):351–353.
51. Tkacz J, Brady B. The increasing rate of childhood mental illnesses and associated healthcare costs in the United States: trends over the past decade. *Value Health*. 2019;22(suppl 2):S231.
52. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. *J Am Med Inform Assoc*. 2018;25(8):1080–1088.
53. Lorenc T, Petticrew M, Welch V, et al. What types of interventions generate inequalities? evidence from systematic reviews. *J Epidemiol Community Health*. 2013;67(2):190–193.
54. Hart JT. The inverse care law. *Lancet*. 1971;297(7696):405–412.

See supplementary material for this article at PSYCHIATRIST.COM.

THE JOURNAL OF
CLINICAL PSYCHIATRY
THE OFFICIAL JOURNAL OF THE AMERICAN SOCIETY OF CLINICAL PSYCHOPHARMACOLOGY

Supplementary Material

Article Title: Faces Scales for Anxiety and Anger: A National Study of Measurement Properties

Author(s): Lisa Grossman Liu, PhD; David Russell, PhD; Meghan Reading Turchioe, PhD, RN; Annie C. Myers, MA; Connie M. Baker, MS; Jyotishman Pathak, PhD; and Ruth M. Masterson Creber, PhD, RN

DOI Number: https://doi.org/10.4088/JCP.21m14202

List of Supplementary Material for the article

Disclaimer

This Supplementary Material has been provided by the author(s) as an enhancement to the published article. It has been approved by peer review; however, it has undergone neither editing nor formatting by in-house editorial staff. The material is presented in the manner supplied by the author.

Supplementary Table 1. Overview of Methodological Process

Development

Name	Description
Population Identification	Definition of measurement goal and target population
Scale Generation	Informed by expert team and literature review
Pretesting	Revision based on free-text commentary by target population

Evaluation

Name	Measurement Properties Assessed	Study	Size
Content Validity	Relevance, comprehensibility, comprehensiveness	#1	300
	Strength-of-association with the construct being measured	#2	300
Construct and Criterion Validity	Convergent validity against a visual method [VAS]	#3	305
	Convergent and concurrent validity against a written method [PROMIS]	#4	1000
Reliability and Measurement Error	Test-retest reliability and measurement error	#5	853

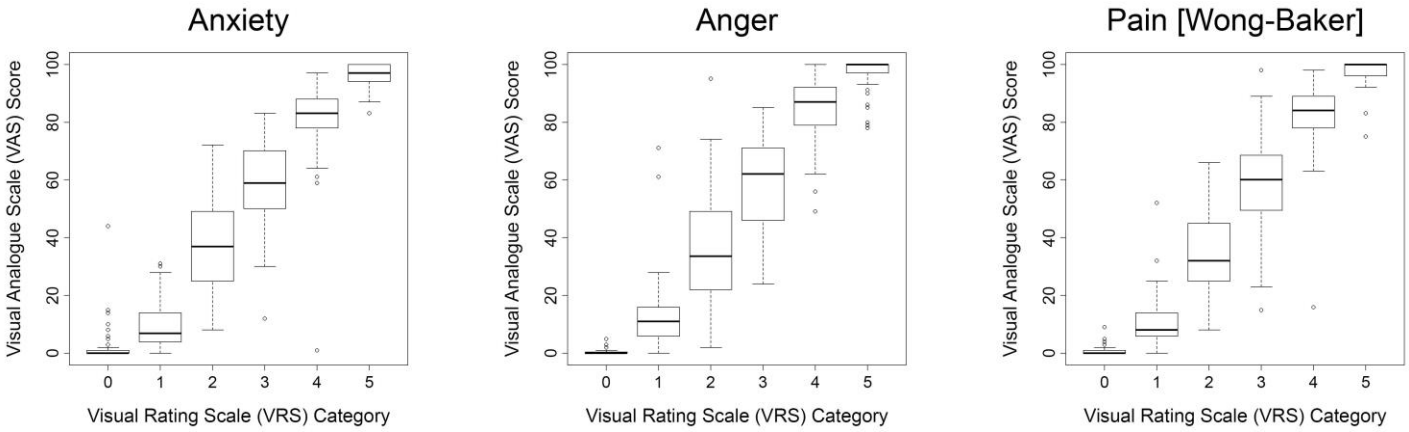
Abbreviations: VAS, visual analogue scale; PROMIS, Patient-Reported Outcome Measurement Information System

Supplementary Table 2. Accepted Participant-Defined Synonyms and Sub-Constructs for Anxiety, Anger, and Pain

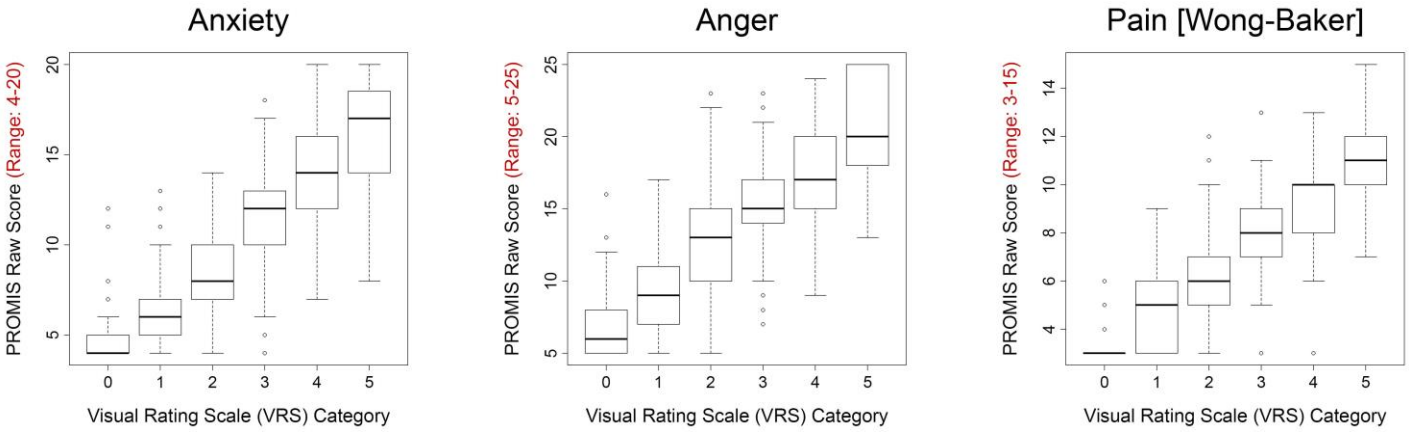
Construct	Synonyms and Sub-Constructs
Anxiety	anxious, very anxious, anxiousness, nervous, nervousness, sweating nervously, worry, worried, very worried, extremely worried, afraid, being afraid, fear, extreme fear, fearful, 10 fear, terror, terrified, fright, frightened, frightening news, scare, scared, extremely scared, alarmed, panic, panicked, panicking, frantic, uneasy, dread, angst, sweat, stress, stressed, distress, distressed, being in distress, severe distress, overwhelmed
Anger	angry, 10 angry, extreme anger, extremely angry, very angry, very angry to the point of yelling, frustrated, infuriated, furious, fury, mad, very mad, upset, outrage, rage, irate
Pain	worst pain, lots of pain, painful, in pain, very painful, a lot of pain, severe pain, physical pain, extreme pain, great pain

Supplementary Figure 1. The Visual Analogue Scale

Supplementary Figure 2. Distribution of VAS Scores by Ordinal Image Category for each Faces Scale



Supplementary Figure 3. Distribution of PROMIS Raw Scores by Ordinal Image Category for each Faces Scale



Supplementary Figure 4. ROC Curves for each Faces Scale

95% confidence interval shown in blue.