Weill Cornell Medicine

Protocol and Statistical Analysis Plan

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

Coordinating Investigator

Ruth Masterson Creber, PhD, RN
Department of Population Health Sciences
425 E 61st St, New York, NY , 10065
rmc2009@med.cornell.edu
+1 646 962 2435

Version 2.0

Finalized 20 October 2020
IRB Protocol #20-01021352-02

Table of Contents

Synopsis 4

1. Background and Objectives 5

1.1. Background 5

1.2. Research Aim 6

1.3. Detailed Description of Scales 6

1.3.1. Definition of Constructs 6

1.3.2. Structure of Scales 6

1.3.3. Response Labels 7

1.3.4. Scoring Algorithm 7

1.3.5. Context of Use 7

1.3.6. Target Population 7

2. Study Design 8

2.1. Methodological Process 8

2.2. Scale Development 8

2.3. Scale Evaluation 8

2.3.1. Participants 8

2.3.2. Baseline Data 9

2.3.3. Survey Administration 10

2.3.4. Controls 11

2.3.5. Content Validity 11

2.3.6. Construct Validity 11

2.3.7. Reliability and Measurement Error 12

2.4. Adherence to Guidelines 12

3. Statistical Analysis 13
3.1. Data Management 13
3.2. Sample Size Calculations 13
3.3. Main Analyses 13
3.4. Secondary Analyses 14

4. References 15

Synopsis

Title	Visual Rating Scales for Anxiety and Anger: A National Study of Measurement Properties
Keywords	Patient Reported Outcome Measures; Health Care Surveys; Patient Participation; Anxiety; Anger
Sponsor	Department of Population Health Sciences, Division of Health Informatics, Weill Cornell Medicine, New York, NY, USA
Sites	Online academic research platform
Design	COSMIN study on measurement properties
Objective	To develop and evaluate the measurement properties of visual rating scales for anxiety and anger.
Outcomes	Content validity • Relevance, comprehensibility, comprehensiveness • Strength-of-association with the measured construct Construct Validity • Convergent validity against visual method (Visual Analogue Scale) • Convergent validity against written method (Patient-Reported Outcome Measurement Information System) Test-retest reliability Measurement error

1. Background and Objectives

1.1. Background

As the healthcare industry enters an era of patient-centered care,[1] providers, researchers, and administrators have increasingly focused on collecting patient-reported outcomes (PROs) alongside more traditional clinical outcomes.[2,3] Patient-centered care requires that patients report information about their symptoms, mood, and the effects of prescribed treatments.[4,5] This information goes beyond traditional clinical outcomes to provide detailed insight into patients' quality-of-life and the effectiveness of their interactions with the healthcare system. As such, federal initiatives in the United States (US) and United Kingdom (UK) have mandated or strongly endorsed PRO collection[6–9] and integration into electronic health records,[10,11] prompting rapid adoption in clinical and research settings.

Today, thousands of research trials[12,13] and hospitals[14] collect PROs, which range from brief self-reported symptom surveys[15] to detailed questionnaires on quality-of-life,[16] functional status,[17] and other constructs. To achieve its goals, PRO measurement must be valid, reliable, and accurately reflect the type, frequency, and intensity of any symptoms being experienced. Validated tools for PRO measurement include the Patient-Reported Outcomes Measurement Information System (PROMIS),[18–20] as well as myriad domain-specific questionnaires. Using these tools, patients answer Likert-response questions to assess symptoms, typically multiple questions per symptom.

Although Likert-response questionnaires can be valid for quantifying PROs, some problems remain unsolved. The *first* and most important problem is that patients with limited literacy struggle to read them. This includes patients with low educational attainment, pediatric patients, non-English speakers, older patients with limited eyesight, and other disadvantaged populations. PROMIS is written at the 6th grade level or below where possible,[21] but 52% of the US adult population cannot read at the 6th grade level,[22] and an estimated 15% of the US population is younger than 6th grade.[23] The *second* problem is that answering written questions can take a long time. Especially for chronic or post-surgical conditions that require frequent monitoring, answering several dozen questions on a daily or weekly basis is a substantial time commitment, one that patients may have difficulty sustaining.[24–26]

One potential solution is *visual rating scales*, which use a series of images to illustrate a spectrum of symptom severity. Visual rating scales are used worldwide for pain assessment, including facial expression drawings ('faces scales') such as

the Wong-Baker FACES Pain Rating Scale.[27] Developing visual rating scales for PROs such as anxiety is critical to solve the two problems described above. Patients who have difficulty reading can use visual rating scales (*problem 1*) and will likely complete them more quickly (*problem 2*). In this study, we will develop and evaluate the measurement properties of visual rating scales for monitoring anxiety and anger. Our main focus is to ensure broad accessibility regardless of age, gender, race, or level of education. The **measurement goal** is high-level monitoring of anxiety and anger, especially when subsidiary rather than primary outcomes of treatment.

1.2. Research Aim

Our objective is to develop and evaluate the measurement properties of visual rating scales for anxiety and anger, in accordance with the COSMIN study design checklist version July 2019.[28] The **name and version** will be: (1) Anxiety Rating Scale, 6-Item, Version 1.0, (2) Anger Rating Scale, 6-Item, Version 1.0. The item number is included in the event that a 5-Item scale is created at a later date.

The **target population** will be U.S. adults, without restriction to a specific illness condition at this time. After the initial understanding of the measurement properties in U.S. adults has been established, the measurement properties in specific clinical contexts or populations such as pediatric populations can be evaluated. The **measurement properties** of interest will include content validity, construct validity, test-retest reliability, and measurement error.

1.3. Detailed Description of Scales

1.3.1. Definition of Constructs

The constructs to be measured will include anxiety and anger. In defining these constructs, we will use the definition of the associated symptoms rather than the definition of the associated disorders. This relates to the overall purpose of the visual rating scales as high-level monitoring tools or screening tools rather than diagnostic tools. We therefore define anxiety as "worry and fear about everyday situations" and anger as "a strong feeling of irritation, impatience, or hostility" per the Mayo Clinic symptom definitions. It is notable that "fear" is considered a sub-construct of anxiety in this definition, as it would be difficult to visually distinguish fear from worry. Accordingly, we will not attempt to make this distinction.

1.3.2. Structure of Scales

The visual rating scales will be structured as follows. Each scale will have 6 items, so that it is consistent with pre-existing visual rating scales, most of which have 6 items. Each scale will measure symptom experience within the past week. Instructions will be phrased as follows: "In the past week, how anxious (worried or afraid) have you been?" and "In the

past week, how angry have you been?" The clarification of anxiety as worry or fear is necessary because the term anxiety may not be understood at lower reading levels.

1.3.3. Response Labels

Importantly, numbers (e.g., 0, 2, 4, 6, 8, 10) and verbal descriptions (e.g., no anxiety, worst anxiety, etc.) will not be included alongside the response options at this time. This is important to the assessment of measurement properties. We want the visuals to be assessed, not the numbers or verbal descriptions. If the numbers and verbal descriptions were included, it is possible that the participants might describe their symptoms based on those features rather than the visuals themselves. Accordingly, we will not include numbers or verbal descriptions in this evaluation. Dependent on the outcome, these features can be added at later time to further enhance the visual rating scales.

1.3.4. Scoring Algorithm

Consistent with scoring algorithms for pre-existing visual rating scales, the 6 image categories will be scored using the numbers 0, 2, 4, 6, 8, and 10. The scoring scheme of 0, 1, 2, 3, 4, and 5 will be avoided. This is consistent with previous research suggesting that base-10 scores are more easily understood.[29] Additionally, some researchers have suggested that using base-10 scores is more representative of the level of clinical change between categories, as the gap between them may be wider compared to more sensitive written questionnaires used as diagnostic tools.

1.3.5. Context of Use

Mental health symptoms are increasingly recognized as highly influential on health outcomes and satisfaction in many clinical contexts.[30,31] Symptoms are prevalent but often unmonitored in many illness conditions. For example, anxiety and anger impact myocardial infarction, but symptoms are often unrecognized, unmonitored, and unaddressed in this population.[32,33] Therefore, we anticipate use in multiple clinical contexts. The potential platforms include measurement in the waiting room, at hospital admission, and through digital health technologies such as patient portals and third-party mobile applications.

1.3.6. Target Population

Because the potential context of use is broad, we identified our **target population** as U.S. adults, without restriction to a patient group with a specific diagnosis at this time. Accordingly, we will recruit U.S. adults nationally representative on age, gender, and race, to be as representative of the target population as possible. The eligibility criteria and sampling methodology is described in the next section. We will not recruit on any illness characteristics at this time, although we anticipate future adaptation of specific contexts and specific populations, such as the pediatric population and populations with limited literacy.

2. Study Design

2.1. Methodological Process

The methodological process for development will include: (*step 1*) population identification, (*step 2*) scale generation, and (*step 3*) pretesting. The methodological process for evaluation will include assessment of: (*step 4*) content validity, (*step 5*) construct validity, and (*step 6*) test-retest reliability and measurement error. The evaluation will involve five separate studies using an online academic research platform.[34] In *study #1*, a national sample of U.S. adults representative on age, gender, and race will be recruited to test relevance, comprehensibility, and comprehensiveness (content validity). In *study #2*, a second representative sample will be recruited to test strength-of-association (content validity). In *study #3*, a third representative sample will be recruited for validation against the visual analogue scale (construct validity). In *study #4*, a fourth representative sample will be recruited for validation against the PROMIS questionnaires (construct validity). In *study #5*, the fourth representative sample will be invited to complete each visual rating scale a second time (test-retest reliability and measurement error).

2.2. Scale Development

A research team of PRO experts, physician experts, patient advocates, and a professional designer will conduct a literature review of existing faces scales. Based on this literature review, the team will develop the new anxiety and anger scales. Then, we will conduct iterative rounds of revision based on free-text commentary from online samples of U.S. adults representative on age, gender, and race. We anticipate conducting up to three rounds of revision.

2.3. Scale Evaluation

2.3.1. Participants

Table 2.1 describes the eligibility criteria for participation. We will recruit representative samples on age, sex, and race as per the United States Census Bureau national adult population demographics on 1 July 2019, publicly available on the website. We will include adult participants (18 or older) who reside in the United States.

Table 2.1. Eligibility Criteria

Inclusion Criteria
Adult (18 years or older)
Residing in the United States
Member of the online research platform
Willing and able to provide informed consent
Can read English proficiently

2.3.2. Baseline Data

We will collect baseline demographic and socioeconomic characteristics from all participants, and assessed their health literacy. **Table 2.2** describes the baseline data that will be collected. Demographic characteristics will include age, gender, race, ethnicity, and primary language, and socioeconomic characteristics will include financial resources, educational attainment, and disability status. To estimate health literacy, we will use the 3-item Brief Health Literacy Screener.[35] We will also collect information about device type, which may impact how the participants view the visual rating scales.

Table 2.2. Baseline Data

Data Field	Question Text	Coded Responses
age	What is your age?	Text (number, minimum 18, maximum 110)
gender	What gender do you identify as?	1 Male 2 Female 3 Non-binary
race	What race do you identify as?	1 White 2 Black or African American 3 Native American or Alaska Native 4 Asian 5 Native Hawaiian or Pacific Islander 6 Multi-race 7 Other
ethnicity	Are you of Hispanic, Latino, or Spanish origin?	1 Yes 0 No
primary_language	What was the primary language spoken in your childhood home?	1 English 2 Mandarin 3 Spanish 4 Other
financial_resources	Do you feel you have enough financial resources to make ends meet?	1 More than enough 2 Enough 3 Not enough

education	What is your highest level of education?	1 Never went to school 2 Less than a high school diploma 4 High school diploma or equivalent (e.g. GED) 5 Some college, no degree 6 Technical, occupational, or vocational school 7 Associate degree (e.g. AA, AS) 8 Bachelor's degree (e.g. BA, BS) 9 Postgraduate degree (e.g. Master's, Professional, or Doctorate)
disabled	Do you have a disability?	1 Yes 0 No
disability_type	Which of the following describes your disability?	1 Problems with physical mobility 2 Problems with seeing 3 Problems with hearing 4 Other
healthlit_1	How often do you have problems learning about your medical condition because of difficulty understanding written information?	1 Never 2 Occasionally 3 Sometimes 4 Often 5 Always
healthlit_2	How confident are you with filling out medical forms by yourself?	1 Extremely 2 Quite a bit 3 Somewhat 4 A little bit 5 Not at all
healthlit_3	How often do you have someone help you read hospital materials?	1 Never 2 Occasionally 3 Sometimes 4 Often 5 Always
device	What device are you using right now?	1 Desktop computer 2 Laptop computer 3 Tablet (e.g. iPad) 4 Mobile phone

2.3.3. Survey Administration

Surveys will be administered using the online research platform *Prolific Academic*.[36] A content expert will use an iterative process to develop each survey. Specifically, each survey will be revised in multiple rounds, with varied input from the research team and members of the general public. Each survey will incorporate formal attention checks to ensure quality. As per recent literature on the potential downsides of incorporating lengthy or intentionally tricky attention checks, attention checks will be short, clear, and consistent with the main content of the surveys. Participants will receive $15 per hour in compensation, which is consistent with the minimum wage where our institution is located (New York City). All surveys will use "force response," meaning participants will not be allowed to leave questions unanswered.

2.3.4. Controls

For comparison with the anxiety and anger scales, the Wong-Baker scale will serve as a positive control. The presentation order of symptoms (anxiety, anger, pain) and presentation order of survey type (visual rating scale, visual analogue scale, or written questionnaire) will be independently computer-randomized to control for potential order effects. Separate samples will be recruited for each study to prevent crossover effects (participants cannot participate more than once).

2.3.5. Content Validity

Table 2.3 displays the questions that will be used to assess each measurement property. We will assess (1) *relevance* of each item for the patients' experience, (2) *comprehensibility* of the instructions and response options, (3) *comprehensiveness* of each scale for the range of possible severity, (4) *strength-of-association* with the construct (e.g., anxiety, anger, pain). To prevent crossover effects, we will use separate studies. Study #1 will assess relevance, comprehensibility, and comprehensiveness, whereas study #2 will assess strength-of-association.

Table 2.3. Content Validity

Property	Question Text	Coded Responses
Relevance	In your opinion, do these faces represent anxiety?	1 Very easy 2 Somewhat easy 3 Neither easy nor difficult 4 Somewhat difficult 5 Very difficult
Comprehensibility	Is it easy or difficult to understand that these faces represent anxiety?	1 Extremely representative 2 Very representative 3 Moderately representative 4 Slightly representative 5 Not at all representative
Comprehensiveness	Does the first face represent the least possible anxiety?	1 Yes 0 No
	Does this last face represent the most possible anxiety?	1 Yes 0 No
Strength-of-Association	What feelings could the first image represent?	Free-text response
	What feelings could the last image represent?	Free-text response

2.3.6. Construct Validity

We will examine the expected relationship of the anxiety, anger, and pain scales with one visual measure (visual analogue scale), and one written measure (Patient-Reported Outcome Measurement Information System or *PROMIS*), in study #3 and study #4, respectively. A standard horizontal visual analogue scale with both numeric labels (0 to 100) and verbal labels (e.g., no anxiety to worst imaginable anxiety) will be used. We will use the following PROMIS questionnaires for comparison: Anxiety 4a v1.0, Anger 5a v1.1, and Pain Intensity 3a v1.0.

2.3.7. Reliability and Measurement Error

We will examine test-retest reliability in participants who complete study #4. Generally, the time interval between test and retest is two weeks, however, we are concerned that visual information may be more easily recalled, which could bias our test-retest results. As such, we will use a time interval of two months, however, PROMIS t-scores will be used to control for clinical stability across time points. The PROMIS measures (i.e., study #4) were selected to control for clinical stability due to greater precision and validity than the visual analogue scale. The exact same survey will be used at both time points.

2.4. Adherence to Guidelines

These studies will be performed in accordance with World Medical Association Declaration of Helsinki. The Weill Cornell Medicine Institutional Review Board will approve the studies. We will conduct and report on the development and evaluation process in accordance with COSMIN.[37]

3. Statistical Analysis

3.1. Data Management

The participant will enter baseline data and outcome data electronically into Qualtrics Survey Software at the time of collection. Data transfer between Qualtrics and computers for statistical analysis will be HIPAA-compliant and the data will be stored in a secure server environment.

3.2. Sample Size Calculations

For *study #1*, we will recruit a unique sample of 100 participants to prevent carryover effects between scales (300 participants total). For *study #2*, we calculated the required sample size 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%. These criteria indicated a minimum sample size of 289 participants. We increased to 300 participants because our survey platform required it to ensure sufficient randomness and representativeness. For *study #3* and *study #4*, we calculated the required sample size needed to detect an expected correlation of 0.75 or higher between each visual rating scale and its corresponding VAS or PROMIS raw scores with a margin of error of 0.05 and a confidence level of 95%. These criteria indicated a minimum sample size of 305 participants. For *study #5*, we calculated the required sample size 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%. These criteria indicated a minimum sample size of 79 participants. To account for potential attrition and clinical instability, we increased the sample size of study #4 and therefore study #5 to 1000 participants.

3.3. Main Analyses

We will compute all statistics in R. For *study #1*, we will compute descriptive statistics of Likert responses. For *study #2*, two researchers will independently code free-text responses as correct or incorrect. Conflicts will be resolved by discussion. We will calculate the proportion of participants who associated each image with the correct construct. For *study #3* and *study #4*, we will use Spearman's correlation to measure agreement between the visual analogue scale scores (*study #3*) or PROMIS raw scores (*study #4*) and each visual rating scale. For *study #5*, we will use

intraclass correlation (two-way mixed-effects model) to assess test-retest reliability. Measurement error will be expressed as the standard error of measurement. In every study, the range of likely values for sample statistics will be estimated using 95% confidence intervals.

3.4. Secondary Analyses

For study #3 and study #4, we will conduct bivariate subgroup analyses using Wilcoxon rank-sum tests to determine whether the median visual analogue scale score and PROMIS raw scores for each visual rating scale category differs by age, gender, race, ethnicity, primary language, financial resources, education, disability status, health literacy, and device type.

4. References

1. Berwick, D. Era 3 for Medicine and Health Care. *JAMA* **315**, 1329–30 (2016).
2. Black, N. Patient reported outcome measures could help transform healthcare. *BMJ* **346**, f167 (2013).
3. Weldring, T. & Smith, S. M. S. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Heal. Serv. Insights* **6**, 61–8 (2013).
4. Deshpande, P., Sudeepthi, Bl., Rajan, S. & Abdul Nazir, C. Patient-reported outcomes: A new era in clinical research. *Perspect. Clin. Res.* **2**, 137 (2011).
5. NEJM Catalyst. What Is Patient-Centered Care? *NEJM Catal. Innov. Care Deliv.* **3**, (2017).
6. NHS Information Centre. National Patient Reported Outcomes Programme. (2020). Available at: www.ic.nhs.uk/proms. (Accessed: 8th April 2020)
7. Centers for Medicare & Medicaid Services. Medicare and Medicaid Programs; Electronic Health Record Incentive Program-Stage 3 and Modifications to Meaningful Use in 2015 Through 2017. (2015). Available at: https://www.federalregister.gov/documents/2015/10/16/2015-25595/medicare-and-medicaid-programs-electronic-health-record-incentive-program-stage-3-and-modifications. (Accessed: 8th April 2020)
8. Centers for Medicare & Medicaid Services. *CMS Quality Measure Development Plan: 2018 Annual Report*. (2018).
9. US National Quality Forum. Patient-Reported Outcomes in Performance Measurement. (2012). Available at: https://www.qualityforum.org/Projects/n-r/Patient-Reported_Outcomes/Patient-Reported_Outcomes.aspx. (Accessed: 8th April 2020)
10. Northwestern University. Electronic Health Record Access to Seamless Integration of PROMIS® (EASIPRO). (2020). Available at: https://sites.northwestern.edu/easipro/. (Accessed: 8th April 2020)
11. Wu, A., Jensen, R., Salzberg, C. & Snyder, C. *Advances in the use of patient reported outcome measures in electronic health records*. *PCORI National Workshop* (2013).
12. U.S. Food and Drug Administration. *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims: Guidance for Industry*. (2009).
13. Willke, R. J., Burke, L. B. & Erickson, P. Measuring treatment impact: A review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Control. Clin. Trials* **25**, 535–552 (2004).
14. European Medicines Agency. *Fourth report on the progress of the interaction with patients' and consumers' organisations*. (2011).

15. Locklear, T. *et al.* Patient-Reported Outcomes. in *NIH Collaboratory Living Textbook of Pragmatic Clinical Trials* (2014).
16. Centers for Disease Control and Prevention. Health-Related Quality of Life. (2018). Available at: https://www.cdc.gov/hrqol/. (Accessed: 8th April 2020)
17. Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A. & Jaffe, M. W. Studies of Illness in the Aged The Index of ADL: A Standardized Measure of Biological and Psychosocial Function. *JAMA* **185**, 914–9 (1963).
18. HealthMeasures. PROMIS® (Patient-Reported Outcomes Measurement Information System). (2020). Available at: http://www.healthmeasures.net/explore-measurement-systems/promis. (Accessed: 8th April 2020)
19. Cella, D. *et al.* The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J. Clin. Epidemiol.* **63**, 1179–1194 (2010).
20. Cella, D. *et al.* The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med. Care* **45**, S3–S11 (2007).
21. PROMIS Cooperative Group. PROMIS® instrument development and validation scientific standards version 2.0. (2013).
22. Goodman, M., Finnegan, R., Mohadjer, L., Krenzke, T. & Hogan, J. *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*. *U.S. Department of Education* (2012). doi:10.1787/9789264128859-en
23. United States Census Bureau. 2018 Population Estimates by Age, Sex, Race and Hispanic Origin. (2019). Available at: https://www.census.gov/newsroom/press-kits/2019/detailed-estimates.html. (Accessed: 8th April 2020)
24. Ali, J. *et al. Users' Guide to Integrating Patient-Reported Outcomes in Electronic Health Records*. (2017).
25. Palmer, M. J. *et al.* A systematic review and development of a classification framework for factors associated with missing patient-reported outcome data. *Clin. Trials* **15**, 95–106 (2018).
26. Mercieca-Bebber, R. *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* **6**, (2016).
27. Wong-Baker FACES Foundation. Wong-Baker FACES Pain Rating Scale. (2016). Available at: https://wongbakerfaces.org/. (Accessed: 8th April 2020)
28. Mokkink, L. B. *et al.* COSMIN Study Design checklist for Patient-reported outcome measurement instruments (Version July 2019). Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf.
29. Garcia-Retamero, R. & Cokely, E. T. Designing Visual AIDS That Promote Risk Literacy: A Systematic Review of Health Research and Evidence-Based Design Heuristics. *Hum. Factors* **59**, 582–627 (2017).
30. Schenker, Y., Stewart, A., Na, B. & Whooley, M. A. Depressive symptoms and perceived doctor-patient communication in the heart and soul study. *J. Gen. Intern. Med.* **24**, 550–556 (2009).

31. Read, C. & Armstrong, A. W. Association between the Mental Health of Patients with Psoriasis and Their Satisfaction with Physicians. *JAMA Dermatology* **156**, 754–762 (2020).
32. Sowden, G. L. & Huffman, J. C. The impact of mental illness on cardiac outcomes: A review for the cardiologist. *Int. J. Cardiol.* **132**, 30–37 (2009).
33. Bremner, J. D. *et al.* Brain Correlates of Mental Stress-Induced Myocardial Ischemia. *Psychosom. Med.* **80**, 515–525 (2017).
34. Peer, E., Brandimarte, L., Samat, S. & Acquisti, A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *J. Exp. Soc. Psychol.* **70**, 153–163 (2017).
35. Chew, L. D., Bradley, K. A. & Boyko, E. J. Brief questions to identify patients with inadequate health literacy. *Fam. Med.* **36**, 588–94 (2004).
36. Palan, S. & Schitter, C. Prolific.ac—A subject pool for online experiments. *J. Behav. Exp. Financ.* **17**, 22–27 (2018).
37. Schulz, K. F., Altman, D. G., Moher, D. & CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann. Intern. Med.* **152**, 726–32 (2010).