Published in final edited form as:
Psychiatry Res. 2013 March 30; 206(1): 88–97. doi:10.1016/j.psychres.2012.09.034.

Assessment of self-reported negative affect in the NIH Toolbox

Paul A. Pilkonis[a,*], **Seung W. Choi**[b], **John Salsman**[b,c], **Zeeshan Butt**[b,c,d,e], **Tara L. Moore**[a], **Suzanne M. Lawrence**[a], **Nicholas Zill**[f], **Jill M. Cyranowski**[a], **Morgen A. R. Kelly**[a,g], **Sarah S. Knox**[h], and **David Cella**[b,c,e]

[a]Department of Psychiatry, University of Pittsburgh Medical Center, Pittsburgh, PA

[b]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL

[c]Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, IL

[d]Comprehensive Transplant Center, Northwestern University Feinberg School of Medicine, Chicago, IL

[e]Institute for Healthcare Studies, Northwestern University Feinberg School of Medicine, Chicago, IL

[f]Westat, Inc., Rockville, MD

[g]Veterans Affairs Pittsburgh Healthcare System, Pittsburgh, PA

[h]Department of Community Medicine, West Virginia University School of Medicine, Morgantown, WV

Abstract

We report on the selection of self-report measures for inclusion in the NIH Toolbox that are suitable for assessing the full range of negative affect including sadness, fear, and anger. The Toolbox is intended to serve as a “core battery” of assessment tools for cognition, sensation, motor function, and emotional health that will help to overcome the lack of consistency in measures used across epidemiological, observational, and intervention studies. A secondary goal of the NIH Toolbox is the identification of measures that are flexible, efficient, and precise, an agenda best fulfilled by the use of item banks calibrated with models from item response theory (IRT) and suitable for adaptive testing. Results from a sample of 1,763 respondents supported use of the adult and pediatric item banks for emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®) as a starting point for capturing the full range of negative affect in healthy individuals. Content coverage for the adult Toolbox was also enhanced by the development of a scale for somatic arousal using items from the Mood and Anxiety Symptom Questionnaire (MASQ) and scales for hostility and physical aggression using items from the Buss-Perry Aggression Questionnaire (BPAQ).

[*]Corresponding Author: Paul A. Pilkonis, Mailing address: Western Psychiatric Institute and Clinic, 3811 O’Hara Street, Pittsburgh, PA 15213, Telephone: 412.246.5833, Fax: 412.246.5840, pilkonispa@upmc.edu.

Keywords

sadness; fear; anger; item response theory; measurement

1. Introduction

The goal of the NIH Blueprint for Neuroscience Research is to accelerate discoveries regarding brain function in health, aging, and disease. In 2006, the NIH Blueprint funded a contract to develop a "core battery" of assessment tools for four key domains of neurological and behavioral functioning—cognition, sensation, motor function, and emotional health. This NIH Toolbox is intended to be relevant for all patient-oriented research, including epidemiological research, observational studies, and studies of interventions (see www.nihtoolbox.org for more details). In addition, the Toolbox is designed to include constructs relevant to health and aging across the lifespan from ages 3 to 85.

An important general aim is to overcome the lack of consistency in measures used to capture these domains across diverse studies and samples, making it difficult to aggregate and compare results. Traditional assessments of negative affect, for example, have relied on methods derived from classical test theory (CTT), and comparing results across different measures of the same construct (e.g., sadness, fear) presents two major problems. First is the problem of test dependence—total scores are dependent on the particular choice of test items, regardless of whether those items accurately describe a participant's experience or provide meaningful information. The second problem is group dependence—samples that differ in ways unrelated to the construct being measured often yield different scores (Hambleton and Jones, 1993). Such difficulties in comparing data obtained from different measures and samples have been noted by several authors, leading to calls for common, standard assessments (Sanders et al., 1998; Garratt et al.,2002; Baughman et al., 2006; Pan et al., 2012).

1.1. Use of modern psychometric methods

A second aim of the NIH Toolbox is the identification of core measures that are flexible, efficient, and precise in assessing cognition, sensation, motor functioning, and emotional health and, as a result, can be administered in about two hours, allotting approximately 30 minutes to each of the four domains. To achieve this goal without compromising coverage of important content, we emphasized the application of modern psychometric methods (e.g., techniques derived from item response theory, IRT), which create the possibility of computerized adaptive testing (CAT) or the development of short forms that can be tailored to specific samples.

In IRT, the unit of information is the individual item as opposed to the total test score. The assumption is that each respondent will have some amount of the underlying trait or ability being assessed and that their level of the trait determines the probability that they will answer an item in a specific way (Embretson and Reise, 2000). Item statistics are independent of the groups from which they are derived, an obvious benefit when attempting to compare results across studies and samples (Hambleton and Jones, 1993). These statistics indicate the level of the trait at which each item will provide the most information (discrimination) and at which it is most likely to be endorsed (difficulty). With fear, for example, an item asking about panic attacks will yield more precise information among respondents with high levels of fear than will a question about general arousal. Item and test statistics are standardized on the same latent trait scale, resulting in information that is comparable across both test forms and groups.

IRT-calibrated item banks underlie the use of CAT (Weiss, 2004; Smits et al., 2011) in which the presentation of items is tailored individually to respondents and their levels of the latent construct being assessed. The result is an efficient procedure for reducing both the total number of items administered and measurement error following the administration of each successive item (Gibbons et al., 2008). Simulation and empirical studies indicate that CAT using as few as 4-6 polytomous items can achieve excellent precision and that scores derived from CAT correlate strongly with a conventional total score (Gardner et al., 2002; Gardner et al., 2004; Bjorner et al., 2007; Choi and Swartz, 2009; Choi et al., 2010). CAT has been successfully validated in a variety of domains, including depression (Fliege et al., 2005), anxiety (Becker et al., 2008), personality assessment (Forbey et al., 2011), and cognitive impairment (Wouters et al., 2001). Due to shorter testing time and greater precision, CAT may also lead to reduced sample size requirements for expensive clinical trials (Chakravarty et al., 2007).

Two studies assessing depression specifically have demonstrated the benefits of CAT over conventional measures. Choi et al.(2010) evaluated the efficiency of static short forms and CAT and found that while both produced scores that correlated highly with scores from more extensive item banks, CAT outperformed static short forms in almost all criteria. Gardner et al. (2004) tested a CAT version of the Beck Depression Inventory (BDI) and found its predictive validity to be equal to the full BDI. The adaptive BDI required an average of 5.6 items versus the 21 items in the full BDI and showed a correlation of 0.92 with the BDI total score.

1.2. Subdomains for emotional health

Within the domain of emotional health, we used responses to a request for information from NIH-funded investigators with broad experience in neurological and behavioral research, as well as expert review by Toolbox investigators, to converge on four subdomains for inclusion: negative affect, psychological well-being, stress and self-efficacy, and social relationships (Salsman et al., 2011). There is good evidence that negative and positive emotions are best understood as orthogonal (rather than bipolar) constructs (Watson and Tellegen, 1985), and this understanding leads to separate assessment of negative and positive psychological functioning in order to investigate their impact on health. Perceived stress and perceived self-efficacy for coping with stressors also have important links with health (Goldman et al., 2005; Lucas et al., 2007; Stewart and Yuen, 2011). Finally, in a social species, it is important to consider emotion in an interpersonal context and to assess its reciprocal relationships with interpersonal functioning and behavior (Lazarus, 2006; Reis and Collins, 2004).

Given the private nature of emotional experience, self-report is considered the standard for assessment. We focused on self-report in the development of the Toolbox (except for the very youngest respondents, i.e., children ages 3 - 7, where the proxy report of a parent or guardian is required). In this paper, we describe the development of self-report measures of negative affect suitable for respondents from ages 8 to 85. There has been a trend toward increased use of self-report (and less reliance on parent proxy report) in the assessment of children, with more evidence that school-age children are indeed able to accurately report their experiences (Varni et al., 2007; Riley et al, 2004). Several studies have shown that parents and children often disagree, especially when the phenomena in question are internal and not directly observable, i.e., sadness, fear (Kolko and Kazdin, 1993; Yeh and Weisz, 2001; Salbach-Andrae et al., 2009). Additionally, parent proxy reports can be influenced by the parent's own symptoms (Fergusson et al., 1993; Renouf and Kovacs, 1994; Garber et al., 1998).

1.3. Definition of negative affect

Emotion "can be manifested as brief *states*, as longer but still transitory *moods*, and as *traits* or patterns of expression that characterize an individual over significant periods of the lifespan" (Goldsmith, 1993). Given the purposes of the Toolbox—to develop measures that are sensitive to change, suitable for monitoring health status over time, and useful for evaluating the effectiveness of interventions—our emphasis is on the measurement of *mood* over brief to intermediate time spans (i.e., days to weeks). Assessment of states that may fluctuate from hour to hour, on the one hand, or assessment of temperamental "set points," on the other hand, are less responsive to the current agenda.

"Negative affect" (NA) is a phrase used to describe unpleasant moods or emotions. These unpleasant states are typically comprised of distinct but often correlated aspects of sadness, fear, and anger. High-NA individuals tend to experience higher levels of these negative emotions, react more negatively to stressful situations, and focus more on subjective experiences (Watson and Clark, 1984).

The overall plan for the Toolbox was to make initial choices about the best candidate instruments for calibration testing, to assess their psychometric properties using a large internet panel, and to make final decisions about inclusion of measures in the Toolbox based on these psychometric outcomes. A parallel NIH Roadmap initiative, the Patient-Reported Outcomes Measurement Information System (PROMIS®), had implemented such methods for a similar agenda—to develop generic measures of physical, mental, and social health that could be used in the assessment of any chronic disease (Cella et al., 2010). PROMIS had developed IRT-calibrated item banks for assessing depression, anxiety, and anger in both adult and pediatric (ages 8 to 17) samples (Irwin et al., 2010a; Irwin et al., 2010b; Pilkonis et al., 2011). Work is underway to link the PROMIS adult and pediatric item banks, thereby ensuring that they are calibrated along a single metric across the lifespan, consistent with the Toolbox agenda. A central question is whether these PROMIS item banks, developed previously with a target population of individuals with chronic health conditions, could be applicable to the Toolbox target of all people in the general population, spanning the full continuum of health. Therefore, we incorporated adapted versions of these item banks (see below) in our Toolbox calibration battery.

At the same time, there are a number of frequently used and commonly recognized measures in the area of negative affect (e.g., PHQ-9, CES-D), and we also wanted to include some of these as potential candidates for the Toolbox battery because of their history, visibility, and research legacy. In most cases, these measures had been developed using CTT. By including them in the Toolbox calibration battery, we were able to see how well they aligned with IRT-calibrated measures of sadness, fear, and anger. Whenever possible, the goal was to make recommendations of measures for inclusion in the Toolbox that had been calibrated using models from IRT and that would be suitable for state-of-the-art adaptive testing.

2. Methods

2.1. Selection of measures

2.1.1. Comprehensive literature searches and expert review—We performed extensive literature searches and received recommendations from consultants and experts in the field for measures assessing sadness, fear, and anger to ensure content validity. Comprehensive literature searches were performed using the PubMed, PsycINFO, Buros Institute Test Reviews Online, Educational Testing Service, Patient-Reported Quality of Life Instrument Database (PROQOLID), Tests and Measures in the Social Sciences, and Health and Psychosocial Instruments (HAPI) databases. Cited reference searches were run on the primary reference for each measure in order to determine its acceptance and use by

the scientific community. For the Toolbox emotional health domain, 554 measures were identified—148 for negative affect.

Our goal was to assess the full spectrum of severity of sadness, fear, and anger. The Toolbox team for negative affect reviewed all measures using several criteria: concept coverage; goal of assessment (continuous measurement of severity vs. diagnostic screening); source of information (self-report vs. proxy report vs. clinical rating); respondent burden (including number of items, reading level, and cognitive complexity); suitability for assessment across the lifespan; and supporting psychometric data. In addition, a careful review of intellectual property issues was done for all measures (Berzon et al., 1994; Ware, 2003; Revicki and Schwartz, 2009). Proprietary measures not available to the general public were excluded.

2.1.2. Instruments selected for Toolbox calibration testing—Table 1 identifies the instruments chosen for Toolbox calibration testing and summarizes their important features (e.g., number of items, time frame, response options). One adaptation was made to the PROMIS adult item banks for depression, anxiety, and anger, which we describe here.

2.1.2.1. PROMIS adult item banks: Depression, anxiety, and anger: The PROMIS item banks for depression, anxiety, and anger include 28, 29, and 29 items, respectively (Pilkonis et al., 2011). However, given the high priority placed on brevity for the Toolbox, we limited the item banks to 20 items for Toolbox calibration testing. The 20 items were selected in two steps. First, we examined incremental correlations with the total IRT-scale score from the full item bank—that is, we started by identifying the single item with the highest correlation with the total score and then added items (in stepwise fashion) that increased the magnitude of the correlation the most. In all three cases (sadness, fear, and anger), the correlation between the score from the best subset of 20 items and the full bank score exceeded 0.99. Second, we ensured that the subsets of 20 items included all items from the static short forms of 8, 7, and 8 items developed for sadness, fear, and anger, respectively, in PROMIS (Pilkonis et al., 2011). This step required one substitution each for sadness and fear, but none for anger.

2.2. Sampling strategy

Participants were drawn from the general population of the United States. They were identified and recruited by Toluna (previously Greenfield Online; www.greenfield.com), an internet survey company. All participants who completed a survey were eligible for prizes or incentive-based compensation. Participants completed a rigorous screening process, their internet protocol (IP) addresses were confirmed to ensure that they were not participating fraudulently in surveys, and they passed a variety of validity checks (e.g., “red herring” questions). Procedures for survey data quality control are described in detail at www.greenfield-ciaosurveys.com/html/qualityassurance.htm. Toluna also timed respondents and removed those who took less than two seconds to answer any 10 items, assuming that those data were likely to be invalid.

2.2.1. Sample—Data were collected from 1,763 participants (58% children and adolescents aged 8-17 years, $n = 1,015$; 42% adults aged 18 and older, $n = 748$). Toluna sent emails to potential candidates in their database to invite them to participate in the study. Potential respondents were screened via the internet to ensure their eligibility, after which they completed a survey of demographic information and the Toolbox negative affect items. Children (aged 8-12 years) and adolescents (aged 13-17 years) completed a survey of 153 items, and adults (18 years and older) completed a survey of 169-187 items. In the child sample, 33% were 8-9 years old and 67% were 10-12 years old. The adolescent sample was comprised of 64% in the age range 13-15 and 36% in the range 16-17. In the adult sample,

three groups (18-39 years, 40-59 years, and 60-92 years) were represented, constituting 33%, 44%, and 24% of the sample, respectively. See Table 2 for a summary of demographic characteristics.

2.3. Data analysis

Measures (and items within measures) were chosen using multiple criteria. We pooled all items from all measures for initial analyses. The primary goal of the initial analyses was to evaluate the dimensional structure of the response data and to develop item pools with sufficient unidimensionality for confirmatory factor analysis (CFA) and IRT analyses. Unidimensionality is a key assumption that underlies the credibility of the model parameters derived from the IRT analyses. Efforts were made to ensure that each measure or subset of items was suitable for unidimensional scaling while still preserving important aspects of item content.

We began by inspecting frequency distributions of individual items for sparse cells (i.e., less than five endorsements in a cell, usually in the most severe response option) and combined such cells with adjoining cells if necessary prior to further analysis. We also examined adjusted item-total correlations, setting a threshold of 0.30 for inclusion in further analysis. We calculated internal consistency reliability, using Cronbach's alpha coefficient. We did careful examinations of dimensionality, including exploratory factor analyses (EFA), eliminating items with modest EFA factor loadings (< 0.40) for single-factor solutions. In addition to EFA, nonparametric multidimensional scaling (MDS) solutions were obtained for each of the three negative affects using polychoric correlations as indices of proximity between items. Two-dimensional plots of MDS solutions provided a useful graphical tool for examining the relationships between items (Roth and Roychoudhury, 1991). A high positive correlation was represented by close proximity in two-dimensional space, and unidimensionality was represented by an elliptical cluster of points (items) arranged along a single line.

2.3.1. Confirmatory factor analysis—Single-factor confirmatory models were fit using Mplus 4.21 (Muthén and Muthén, 2006) in order to document unidimensionality. The items were treated as categorical variables. The robust weighted least squares (WLSMV) estimator was used. Error variances and residual correlations were examined to identify items that performed poorly (i.e., items with large errors, pairs of items demonstrating local dependency—significant residual correlations even after accounting for the single underlying latent variable).

2.3.2. Calibration with IRT models

2.3.2.1. Graded response model: Items remaining in the pool for each domain were calibrated with the two-parameter graded response model (GRM) using MULTILOG 7.03 (Thissen et al., 2003). The convergence criterion for the EM cycles was set to 0.0001, with the number of cycles set to 100. IRT model fit was examined for each item using the IRTFIT macro program (Bjorner et al., 2006) and the option for the sum-score-based method (Orlando and Thissen, 2003), which uses the sum score instead of theta for computing the predicted and observed frequencies.

2.3.2.2. Analysis of differential item functioning (DIF): Differential item functioning (DIF) occurs when characteristics such as age, gender, or ethnicity, which may seem extraneous to the assessment of cognitive and psychological functioning, actually do have an effect on measurement. An item is identified as functioning differentially if the item is more (or less) difficult to endorse or more (or less) discriminating in some focal group (compared to a reference group) when the different subgroups have been matched on the latent trait

under investigation. Demographic and health-related variables, for example, have been found to affect response patterns in depression scales such as the CES-D, the Beck Depression Inventory, and the Geriatric Depression Scale (Mui et al., 2001; Kim et al., 2002; Pedersen et al., 2002). We conducted DIF analyses (for both uniform and non-uniform DIF) on the basis of gender (in all groups) and age (in the adult sample, with two age groups, ages 18-45 and 46 and older, which divided the sample approximately in half). Multiple DIF procedures were employed—the IRT likelihood ratio method (Thissen et al., 1993), an ordinal logistic regression procedure (Zumbo, 1999), and a second logistic regression procedure (Choi, et al., 2010) which generates effect size estimates (pseudo-R squares) that are helpful in assessing the clinical significance of DIF results that may be statistically significant (in large samples) but have little effect on IRT trait scores. Items were removed if they showed significant DIF by multiple methods, large effect sizes for DIF, or both.

3. Results

3.1. Sadness

3.1.1. Adult self-report (age 18 and older)—We created a pool of 49 items that included the 20 items from the PROMIS depression item bank, the 20 items from the CES-D, and the 9 items from the PHQ-9. For the total pool of items, Cronbach's alpha was 0.98, with a mean adjusted item-total correlation of 0.72 (in a range from 0.41 to 0.87). No items were flagged for sparse cells or eliminated on the basis of low (<0.30) adjusted item-total correlations. Correlations between the three instruments (based on raw total scores) were PROMIS depression and PHQ-9, 0.84; PROMIS depression and CES-D, 0.88; and PHQ-9 and CES-D, 0.88.

We examined one-, two-, and three-factor solutions with EFAs. The one-factor solution provided the best fit to the data, with loadings for all 49 items being 0.50 or higher. Two- and three-factor solutions produced "splinter" factors with small numbers of items (2-5) that were not readily interpretable and appeared to represent arbitrary indicators of "unique" variance not captured by the major underlying factor. We also examined a two-dimensional plot from a multidimensional scaling analysis. The plot reflected an essentially unidimensional array, with all items included except four outliers from the CES-D—the four items that are worded positively and that segregated for this reason. In general, these results demonstrated that the three instruments are measuring the same latent construct, and they encourage efforts to link or co-calibrate the measures with confidence that it is valid to do so.

In order to examine the PROMIS 20-item depression bank, the PHQ-9, and the CES-D on the same metric, we calibrated the two legacy measures concurrently with the 20-item bank (using the GRM) by fixing the PROMIS item parameters at their bank values. We have created an 8-item short form from the PROMIS depression item bank that includes many of the most informative items (Pilkonis et al., 2011), and we also examined the performance of the short form in this context. Figure 1 displays the test information curves for the 20-item bank, the short form, and the two legacy measures. The alpha coefficient for the short form was 0.97, and the correlation with the 20-item bank was 0.98. It is clear that the full item bank provides the most information, but the figure also illustrates that the PROMIS short form performs better than the two legacy measures, with fewer items. By way of reference, "information" of 10 or greater approximates internal consistency of .90 or greater.

Because of these results and the Toolbox agenda aimed at promoting brief, efficient, and flexible assessment (best operationalized through the use of IRT-calibrated item banks, computerized adaptive testing, or the creation of short forms that can be customized for

different samples), we recommend the PROMIS depression item bank as the primary resource for the Toolbox Sadness Test. The items from the 20-item bank (and the general short form, which gives good coverage across the entire range of measurement) are included in Table 3. In the context of the Toolbox, which aims to measure emotional health in the normal range, the 8-item short form produces reliable scores (as indicated by information > 10) for people nearly one standard deviation less sad than the average person in the calibration sample.

3.1.2. Pediatric self-report (age 8-17)—We created a pool of 47 items that included the 14 items from the PROMIS pediatric depression item bank, the 13 items from the SMFQ, and the 20 items from the CES-DC. For the total pool of items, Cronbach's alpha was 0.97, with a mean adjusted item-total correlation of 0.64 (in a range from 0.31 to 0.80). No items were eliminated on the basis of low (<0.30) adjusted item-total correlations. One item was flagged for a sparse cell: "I felt too sad to eat," an item from the PROMIS bank, for the response option of "almost always." Correlations between the three instruments (based on raw total scores) were PROMIS depression and SMFQ, 0.79; PROMIS depression and CES-DC, 0.83; and SMFQ and CES-DC, 0.81.

We examined one-, two-, and three-factor solutions with EFAs. Consistent with the adult results for depression, a one-factor solution provided the best fit to the data, with loadings for all but 2 of the 47 items being 0.50 or higher. Inspection of the two-dimensional plot from a multidimensional scaling analysis also revealed an essentially unidimensional array, incorporating all items except the four positively worded outliers from the CES-DC. These results were very similar to the results for the adult measures, also demonstrating that the three pediatric instruments are measuring the same latent construct. Again, because of these results and the advantages of IRT-calibrated item banks, we recommend the PROMIS pediatric depression item bank (and its 8-item short form) as the primary resources for the Toolbox Sadness Test for children and adolescents. The alpha coefficient for the short form was .95, and its correlation with the full bank was 0.98. The items from the 14-item PROMIS pediatric depression item bank (and the short form) are also included in Table 3.

3.2. Fear

3.2.1. Adult self-report (age 18 and older)—We created a pool of 55 items that included the 20 items from the PROMIS anxiety item bank, the 7 items from the GAD-7, and the 28 items from the general distress/anxiety (11 items) and the anxious arousal (17 items) subscales of the MASQ. For the total pool of items, Cronbach's alpha was 0.98, with a mean adjusted item-total correlation of 0.70 (in a range from 0.46 to 0.86). No items were eliminated on the basis of low (<0.30) adjusted item-total correlations. However, two items from the MASQ were flagged for sparse cells ("had a lump in my throat" and "felt I was choking"), and the two most severe response options ("often" and "always") were collapsed for these items. Correlations between the three instruments (based on raw total scores) were PROMIS anxiety and MASQ, 0.82; PROMIS anxiety and GAD-7, 0.86; and MASQ and GAD-7, 0.76.

We examined one-, two-, and three-factor solutions with EFAs. A two-factor solution provided the best fit to the data. The first factor reflected primarily affective and cognitive features. This factor captured all the items from the GAD-7 and the PROMIS item bank and six of the more "psychological" items from the MASQ, e.g., feelings of fear, nervousness, and unease. The second factor reflected primarily somatic features and captured 22 items from the MASQ. We also examined a two-dimensional plot from a multidimensional scaling analysis. The plot also generated two clusters of items reflecting the same distinction between affective and cognitive versus somatic indicators.

Figure 2 displays the test information curves for the measures assessing the affective and cognitive markers of fear—the PROMIS anxiety item bank, the PROMIS 7-item short form, and the GAD-7. Again, it is clear that the item bank provides the most information, but Figure 2 also illustrates that the PROMIS short form performs better than the GAD-7 with the same number of items. Because of these results (comparable to our findings for the adult measures of depression), we recommend the PROMIS anxiety item bank (and its 7-item short form) as the primary resources for the Toolbox Fear Test. The alpha coefficient for the short form was 0.95, and the correlation with the adapted 20-item bank was 0.98. The items from the 20-item anxiety bank (and the short form) are included in Table 4.

Given the presence of an additional factor for somatic arousal, however, we developed a second short form for this construct, using the relevant items from the MASQ. We began with the 22 items identified in the EFA. We eliminated one item (“startled easily”) with a marginal factor loading (0.40), with all other factor loadings ranging from 0.54 to 0.92. We performed a CFA and eliminated two items for local dependency (“had diarrhea” and “had an upset stomach,” whereas “felt nauseous” remained in the pool). CFA fit indices for the 19-item scale were CFI = 0.90 and TLI = 0.99. We calibrated the remaining 19 items in a 2-parameter GRM and eliminated two more items at this stage, one for local dependency (“hands were shaky,” which showed a residual correlation with “was trembling and shaky”) and one for DIF on the basis of gender (“had hot or cold spells,” which was more easily endorsed by women). We performed CAT simulations with the remaining 17 items (using the calibration data) to identify the items that would have been selected most frequently in a CAT environment. We chose the 6 most frequently selected items for a somatic arousal short form (with this 6-item subset correlating 0.95 with the total score from the full 17-item pool). We performed a CFA on these remaining items, and they showed good fit with a unidimensional model: CFI = 0.96 and TLI = 0.97. The alpha coefficient was 0.85. This new 6-item scale is also included in Table 4.

In summary, the Fear Test for the adult Toolbox includes two measures: the PROMIS anxiety item bank (and its 7-item short form) and the new 6-item short form for somatic arousal (derived from the MASQ).

3.2.2. Pediatric self-report (age 8-17)—We created a pool of 56 items that included the 15 items from the PROMIS pediatric anxiety item bank and the 41 items from the SCARED. For the total pool of items, Cronbach’s alpha was 0.97, with a mean adjusted item-total correlation of 0.58 (in a range from 0.27 to 0.74). No items were flagged for sparse cells, but one item from the SCARED (“I follow my mother or father wherever they go”) had an adjusted item-total correlation (0.27) below our threshold of 0.30. The correlation between the two instruments (based on raw total scores) was .68.

We examined various solutions with EFAs. The SCARED was designed to be multidimensional, and an EFA of its 41 items alone produced 6 factors with eigenvalues greater than 1. By contrast, the PROMIS items could be grouped into a single factor reflecting the internal experience of fear and worry. We also examined a two-dimensional plot from a multidimensional scaling analysis. The plot generated three clusters of items. The largest cluster captured all the PROMIS items and the SCARED items for generalized anxiety, panic, somatic symptoms, and school avoidance. Two smaller clusters contained the SCARED items for separation anxiety and social anxiety, respectively.

For a measure of the internalized, psychological experience of fear in children and adolescents, we recommend the PROMIS pediatric anxiety item bank (and its 8-item short form) as the primary resource for the Toolbox Fear Test. This item pool (and short form) are consistent with the content of the generalized anxiety subscale from the SCARED but have

clearer psychometric support, given the more varied content of the SCARED. The alpha coefficient for the short form was 0.93, and its correlation with the full bank was 0.98. The items from the 15-item PROMIS pediatric anxiety bank (and the short form) are also included in Table 4.

3.3. Anger

3.3.1. Adult self-report (age 18 and older)—We created a pool of 49 items that included the 20 items from the PROMIS anger item bank and the 29 items from the BPAQ. For the total pool of items, Cronbach's alpha was 0.96, with a mean adjusted item-total correlation of 0.59 (in a range from 0.17 to 0.76). Three items from the BPAQ were eliminated on the basis of low (<0.30) adjusted item-total correlations, but no items were flagged for sparse cells. The correlation between the two instruments (based on raw total scores) was 0.68.

We examined one-, two-, and three-factor solutions with EFAs of the remaining 46 items. A three-factor solution provided the best fit to the data. The first factor captured the 20 items from the PROMIS item bank, which focuses primarily on the affective and cognitive features of anger. The second factor captured 18 items from the BPAQ that were a mix of physical and verbal aggression. The third factor captured the 8 items from the original hostility subscale of the BPAQ. We also examined a two-dimensional plot from a multidimensional scaling analysis. The plot also generated three clusters of items reflecting the same tripartite structure.

Given these results, we recommend the PROMIS anger item bank (and its 8-item short form) as the primary resource for the Toolbox for a measure of the emotional experience of anger. The alpha coefficient for the short form was 0.94, and the correlation with the 20-item bank was 0.97. The items from the 20-item anger bank (and the short form) are included in Table 5. Note that one item from the anger bank ("I felt like I was ready to explode") exhibited uniform DIF for age, being more easily endorsed among younger adults (age 45 or less) in the Toolbox calibration sample. This, however, was the only item (among the 60 PROMIS items) that displayed age-related DIF, and given that such DIF was neither substantive nor pervasive, we would not propose altering the banks without more convincing empirical evidence. Given the presence of two additional factors for hostility and aggression, however, we did choose to develop two other short forms for these constructs, using the relevant items from the BPAQ.

For hostility, we began with the 8 items from the original subscale and did an iterative series of CFAs and IRT analyses using the 2-parameter GRM. We eliminated two items on the basis of CFA misfit reflecting local dependency ("I know that friends talk about me behind my back" and "I am suspicious of overly friendly strangers"), and one item on the basis of non-uniform gender DIF ("When people are especially nice, I wonder what they want"). This item was easier for men to endorse at lower levels of anger, whereas the opposite was true at higher levels of anger. We performed a CFA on the final five items, and they showed good fit with a unidimensional model: CFI = 0.98 and TLI = 0.98. The alpha coefficient was 0.85. This new 5-item scale is also included in Table 5.

For aggression, we chose to create a measure that would cover the higher end of the spectrum, i.e., physical aggression, and provide more complete content coverage of the domain for Toolbox purposes. Therefore, we began with the 8 items remaining in the pool from the original 9-item physical aggression subscale of the BPAQ. We did an iterative series of CFAs and IRT analyses using the 2-parameter GRM. We eliminated one item on the basis of CFA misfit reflecting local dependency ("Once in a while, I can't control the urge to strike another person") and two items on the basis of uniform gender DIF ("If

somebody hits me, I hit back" and "If I have to resort to violence to protect my rights, I will"). Both of these latter items were easier for men to endorse. We performed a CFA on the final five items, and they showed good fit with a unidimensional model: CFI = 0.98 and TLI = 0.99. The alpha coefficient was 0.83. This new 5-item scale is also included in Table 5.

In summary, the Toolbox Anger Test includes three measures: the PROMIS anger item bank (and its 8-item short form), the new 5-item short form for hostility (derived from the BPAQ), and the new 5-item short form for physical aggression (also derived from the BPAQ) and serving as a marker for externalizing behaviors in adults.

3.3.2. Pediatric self-report (age 8-17)—We created a pool of 18 items that included the 6 items from the PROMIS pediatric anger scale and 12 items from the AESC. For the total pool of items, Cronbach's alpha was 0.94, with a mean adjusted item-total correlation of 0.67 (in a range from 0.46 to 0.78). No items were eliminated on the basis of low (<0.30) adjusted item-total correlations or flagged for sparse cells. The correlation between the two instruments (based on raw total scores) was 0.73.

We examined one-, two-, and three-factor solutions with EFAs. A one-factor solution provided the best fit to the data, with loadings on 15 of the 18 items being 0.70 or higher (and no loadings less than 0.50). We also examined a two-dimensional plot from a multidimensional scaling analysis, which largely confirmed the unidimensional structure. Two positively worded items from the AESC were outliers on the basis of their linguistic difference, but all other items were grouped into a single cluster. In order to assess this factor reflecting the internalized, psychological experience of anger, we recommend the PROMIS pediatric anger scale as the primary resource for the Toolbox. The alpha coefficient was 0.92. The PROMIS pediatric anger scale is also included in Table 5.

3.4. Correlations among the measures

Validation of the Toolbox measures remains an important goal for future research, but some initial results on the convergent and discriminant validity of the measures is available from the calibration sample. For the adult subsample, Table 6 displays the intercorrelations among the six measures recommended for the Toolbox Sadness, Fear, and Anger Tests. The 15 correlations ranged from 0.37 to 0.81, with a median of 0.57, demonstrating good convergent validity among the measures of negative affect as expected. The smallest correlation (0.37, greatest discrimination) occurred between BPAQ physical aggression and PROMIS anxiety, and the largest correlation (0.81, greatest convergence) occurred between PROMIS depression and PROMIS anxiety. In general, the BPAQ physical aggression scale was the most distinctive measure, correlating in a range from 0.37 to 0.52 with the other five measures, whereas the other measures correlated among themselves in a range from 0.50 to 0.81. For the pediatric measures, the correlations were PROMIS depression and PROMIS anxiety, 0.71; PROMIS depression and PROMIS anger, 0.71; and PROMIS anxiety and PROMIS anger, 0.63.

4. Discussion

The NIH Toolbox for the measurement of emotional health includes negative affect as one of its four subdomains. Toolbox negative affect is divided into three components: sadness, fear, and anger. Because they were designed to cover the full range of negative affect, the adapted PROMIS item banks and their respective short forms performed very well for Toolbox purposes and can be recommended for inclusion. Their content is consistent with the subdomain of negative affect, and their calibration with IRT models promotes the Toolbox agenda focused on creating more flexible and efficient methods of clinical

assessment. These item banks can be used for computerized adaptive testing or can be adapted as short forms (either in the standard version or in versions customized for particular samples or disease states depending on the expected base rates of negative affect).

In addition to adapting PROMIS item banks for use in the Toolbox, we were able to enrich the content of the adult Toolbox by creating three new measures from existing items in legacy instruments. These new measures provide improved assessment of somatic arousal, hostility, and physical aggression. The symptoms that best differentiate fear from near neighbors (e.g., sadness) reflect autonomic arousal and other symptoms common to the human fear response. This factor has typically been labeled “anxious arousal” or “physiological reactivity” (Clark and Watson, 1991; Brown et al., 1998; Chorpita et al., 1998). Our new scale for somatic arousal (using items from the MASQ) reflects this constellation of indicators and provides a useful complement to the Toolbox item bank, which has many affective and cognitive markers. In a similar way, the new measures for anger complement the Toolbox anger item bank, which focuses on the emotional experience of anger. The two new measures capture, on one hand, hostility, which has been linked most clearly to cardiovascular disease (Chida and Steptoe, 2009), and, on the other hand, overt markers of aggression, which provide assessment of externalizing behaviors not otherwise assessed in the Toolbox battery.

Further validation research on Toolbox item banks and associated measures is ongoing. For example, they are being evaluated for their sensitivity to change, their ability to detect differences between distinct clinical conditions, and their concurrent validity with additional legacy measures in samples of patients with depression, sleep disorders, and chronic pain. Although the primary purpose of the Toolbox item banks is to create a continuous metric of the severity of emotional distress, such validation work in various clinical (and non-clinical comparison) groups will also allow us to assess the utility of the banks in screening for diagnosable disorders.

As mentioned above, work is also underway to link the Toolbox and PROMIS adult and pediatric item banks, thereby ensuring that they are calibrated along a single metric across the lifespan, consistent with the Toolbox agenda. In this regard, our analysis of DIF by age in the adult sample alone (dividing the group at age 45) is reassuring, demonstrating the equivalence of the Toolbox measures in young and middle adulthood versus later adulthood. No items in the newly derived Toolbox measures (MASQ somatic arousal, BPAQ hostility, BPAQ physical aggression) displayed age-related DIF, and only 1 of 60 PROMIS items demonstrated such DIF.

In summary, we have a single measure of sadness, two measures for fear, and three measures for anger in the adult Toolbox battery. Our experience with CAT (Choi et al., 2010) is that 4-6 items will provide excellent precision for the assessment of all of these constructs when using adaptive testing. If CAT is not feasible and static short forms are required, then the total number of items for the adult Toolbox profile of negative affect would be 40: 8 for sadness (from the Toolbox short form), 13 for fear (7 from the Toolbox short form and 6 for somatic arousal), and 19 for anger (8 from the Toolbox short form, 5 for hostility, and 5 for physical aggression). The entire battery is likely to require 8-10 minutes for adult respondents. The pediatric Toolbox profile would include 22 items: 8 each for sadness and fear from the Toolbox short forms and 6 from the Toolbox anger scale. Again, the assessment is likely to require only a few minutes for children and adolescents. The brevity, precision, and efficiency of these measures make them good candidates for use not only in clinical research but also in clinical practice. In both venues, the measures will capture more information than many traditional instruments while minimizing respondent burden.

Acknowledgments

This project is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research and the Office of Behavioral and Social Sciences Research, National Institutes of Health, under Contract No. HHS-N-260-2006-00007-C.

We thank our consultants for the negative affect subdomain of emotional health, especially Michael Scheier, PhD, Daniel Shaw, PhD, and David Watson, PhD. We also thank the members of the NIH project team, Pim Brouwers, PhD, Catherine M. Stoney, PhD, and Gitanjali Taneja, PhD, who provided critical thinking and constructive expertise during the development of the emotional health measurement battery. We also acknowledge the contribution of Tracy Podrabsky and Natalie McKinney, who provided support for data analysis.

References

Angold A, Costello EJ, Messer SC, Pickles A. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. International Journal of Methods in Psychiatric Research. 1995; 5(4):237–249.

Baughman RW, Farkas R, Guzman M, Huerta M. The national institutes of health blueprint for neuroscience research. The Journal of Neuroscience. 2006; 26(41):10329–10331. [PubMed: 17035514]

Becker J, Fliege H, Kocalevent RD, Bjorner JB, Rose M, Walter OB, Klapp BF. Functioning and validity of a Computerized Adaptive Test to measure anxiety (CAT). Depression and Anxiety Journal. 2008; 25(12):182–194.

Berzon R, Patrick D, Guyatt G, Conley JM. Intellectual property considerations in the development and use of HRQL measures for clinical trial research. Quality of Life Research. 1994; 3(4):273–277. [PubMed: 7812280]

Birmaher B, Khetarpal S, Brent D, Cully M, Balach L, Kaufman J, Neer SM. The Screen for Child Anxiety Related Emotional Disorders (SCARED): scale construction and psychometric characteristics. Journal of the American Academy of Child and Adolescent Psychiatry. 1997; 36(4): 545–553. [PubMed: 9100430]

Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: Item banking and computerized adaptive assessment. Quality of Life Research. 2007; 16:95–108. [PubMed: 17530450]

Bjorner, JB.; Smith, KJ.; Orlando, M.; Stone, CA.; Thissen, D.; Xiaowa, S. IRTFIT: A macro for item fit and local dependence tests under IRT models. L. L. Thurstone Psychometric Laboratory; Chapel Hill, NC: 2006. Computer software

Brown TA, Chorpita BF, Barlow DH. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. Journal of Abnormal Psychology. 1998; 107:179–192. [PubMed: 9604548]

Buss AH, Perry M. The Aggression Questionnaire. Journal of Personality and Social Psychology. 1992; 63:452–459. [PubMed: 1403624]

Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse DJ, Choi SW, Cook KF, DeVellis R, DeWalt D, Fries JF, Gershon R, Hahn E, Lai J-S, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays RD. 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. 2010; 63:1179–1194. [PubMed: 20685078]

Charkravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. The Journal of Rheumatology. 2007; 34(6):1426–1431. [PubMed: 17552069]

Chida Y, Steptoe A. The association of anger and hostility with future coronary heart disease: a meta-analytic review of prospective evidence. Journal of the American College of Cardiology. 2009; 53(11):936–946. [PubMed: 19281923]

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. Quality of Life Research. 2010; 19:125–136. [PubMed: 19941077]

Choi SW, Swartz JR. Comparison of CAT item selection criteria for polytomous items. Applied Psychological Measurement. 2009; 33:419–440. [PubMed: 20011456]

Chorpita BF, Albano AM, Barlow DH. The structure of negative emotions in a clinical sample of children and adolescents. Journal of Abnormal Psychology. 1998; 107:74–85. [PubMed: 9505040]

Clark LA, Watson D. Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. Journal of Abnormal Psychology. 1991; 100:316–336. [PubMed: 1918611]

Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. Lawrence Erlbaum Associates; Mahwah: 2000.

Fergusson DM, Lynskey MT, Horwood JL. The effect of maternal depression on maternal ratings of child behavior. Journal of Abnormal Child Psychology. 1993; 21(3):245–269. [PubMed: 8335763]

Fliege H, Becker J, Walter ObB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Quality of Life Research. 2005; 14(10):2277–2291. [PubMed: 16328907]

Forbey JD, Ben-Porath YS, Arbisi PA. The MMPI-2 computerized adaptive version (MMPI-2-CA) in a veterans administration medical outpatient facility. Psychological Assessment. 2011

Garber J, Van Slyke DA, Walker LS. Concordance between mothers' and childrens' reports of somatic and emotional symptoms in patients with recurrent abdominal pain or emotional disorders. Journal of Abnormal Child Psychology. 1998; 26(5):381–391. [PubMed: 9826296]

Gardner W, Kelleher KJ, Pajer KA. Multidimensional adaptive testing for mental health problems in primary care. Medical Care. 2002; 40:812–823. [PubMed: 12218771]

Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, Buysse D, Frank E. Computerized adaptive measurement of depression: A simulation study. BMC Psychiatry. 2004; 4:13. [PubMed: 15132755]

Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. British Medical Journal. 2002; 324:1–5. [PubMed: 11777781]

Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagiolini A, Grochocinski VJ, Bhaumik DK, Stover A, Bock RD, Immekus JC. Using computerized adaptive testing to reduce the burden of mental health assessment. Psychiatric Services. 2008; 59:361–368. [PubMed: 18378832]

Goldman N, Glei DA, Seplaki C, Liu IW, Weinstein M. Perceived stress and physiological dysregulation in older adults. Stress. 2005; 8(2):95–105. [PubMed: 16019601]

Goldsmith, HH. Temperament: Variability in developing emotion systems. In: Bates, JE.; Haviland, JM.; Lewis, M., editors. Handbook of emotion. The Guilford Press; New York: 1993. p. 353-364.

Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice. 1993; 12(3):38–47.

Irwin DE, Stucky B, Langer MM, Thissen D, Dewitt EM, Lai J-S, Varni JW, Yeatts K, DeWalt D. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. Quality of Life Research. 2010a; 19(4):595–607. [PubMed: 20213516]

Irwin DE, Stucky BD, Langer MM, Thissen D, Dewitt EM, Lai J-S, Yeatts K, Varni JW, DeWalt D. PROMIS Pediatric Anger Scale: An Item response Theory Analysis. 2010b Unpublished manuscript.

Jacobs GA, Phelps M, Rohrs B. Assessment of anger expression in children: The pediatric anger expression scale. Personality and Individual Differences. 1989; 10:59–65.

Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF. Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. Psychology and Aging. 2002; 17:379–391. [PubMed: 12243380]

Kolko DJ, Kazdin AE. Emotional/behavioral problems in clinic and nonclinic children: correspondence among child, parent and teacher reports. Journal of Child Psychology and Psychiatry. 1993; 34(6):991–1006. [PubMed: 8408380]

Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine. 2001; 16(9):606–613. [PubMed: 11556941]

Lazarus RS. Emotions and interpersonal relationships: Toward a person-centered conceptualization of emotions and coping. Journal of Personality. 2006; 74(1):9–46. [PubMed: 16451225]

Lucas RM, Ponsonby AL, Dear K. Mid-life stress is associated with both up- and donw-regulation of markers of humoral and cellular immunity. Stress. 2007; 10(4):351–361. [PubMed: 17853062]

Mui AC, Burnette D, Chen LM. Cross-cultural assessment of geriatric depression: A review of the CES-D and GDS. Measurement in Older Ethnically Diverse Populations. Journal of Mental Health and Aging. 2001; 7:137–164.

Muthén, LK.; Muthén, B. Mplus user's guide. 4. Muthén & Muthén; Los Angeles: 2006.

Orlando M, Thissen D. Further invesigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement. 2003; 27:289–298.

Pan YJ, Knapp M, McCrone P. Cost-effectiveness comparisions between antidepressant treatments in depression: Evidence from database analyses and prospective studies. Journal of Affective Disorders. 2012; 139(2):113–125. [PubMed: 21851987]

Pedersen RD, Pallay AG, Rudolph RL. Can improvement in well-being and functioning be distinguished from depression improvement in antidepressant clinical trials? Quality of Life Research. 2002; 11:9–17. [PubMed: 12003058]

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. Assessment. 2011; 18:263–283. [PubMed: 21697139]

Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977:385–401.

Reis HT, Collins WA. Relationships, human behavior, and psychological science. Current Directions in Psychological Science. 2004; 13(6):233–237.

Renouf AG, Kovacs M. Concordance between mothers' reports and children's self-reports of depressive symptoms: a longitudinal study. Journal of the American Academy of Child and Adolescent Psychiatry. 1994; 33(2):208–216. [PubMed: 8150792]

Revicki D, Schwartz C. Intellectual property rights and good research practice. Quality of Life Research. 2009; 18(10):1279–1280. [PubMed: 19885743]

Riley AW. Evidence that school-age children can self-report on their health. Ambulatory Pediatrics. 2004; 4(4):374–376.

Roth WM, Roychoudhury A. Nonmetric multidimensional item analysis in the construction of an anxiety attitude survey. Educational and Psychological Measurement. 1991; 51:931–942.

Salbach-Andrae H, Lenz K, Lehmkuhl U. Patterns of agreement among parent, teacher, and youth ratings in a referred sample. European Psychiatry. 2009; 24(5):345–351. [PubMed: 18789656]

Salsman JM, Butt Z, Pilkonis PA, Cyranowski JM, Zill N, Hendrie HC, Kupst MJ, Kelly MAR, Bode RK, Choi SW, Lai J-S, Griffith JW, Stoney CM, Brouwers P, Knox SS, Cella D. Emotional health and its assessment within the NIH Toolbox. 2011 Unpublished manuscript.

Sanders C, Egger M, Donovan J, Tallon D, Frankel S. Reporting on quality of life in randomised controlled trials: bibliographic study. British Medical Journal. 1998; 317:1191–1194. [PubMed: 9794853]

Smits N, Cuijpers P, van Straten A. Applying computerized adaptive testing to the CES-D scale: A simulation study. Psychiatry Research. 2011; 188:147–155. [PubMed: 21208660]

Spitzer R, Kroenke K, Williams J, Lowe. The GAD 7. A brief measure for assessing generalised anxiety disorder. Archives of Internal Medicine. 2006; 166:1092–1097. [PubMed: 16717171]

Stewart DE, Yuen T. A systematic review of resilience in the physically ill. Psychosomatics. 2011; 52(3):199–209. [PubMed: 21565591]

Thissen, D.; Chen, W-H.; Bock, RD. Multilog (Version 7). Scientific Software International; Lincolnwood, IL: 2003.

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Differential Item Functioning. Lawrence Erlbaum; Hillsdale, NJ: 1993. p. 67-113.

Varni JW, Limbers CA, Burwinkle TM. How young can children reliably and validly self-report their health-related quality of life? An analysis of 8,591 children across age subgroups with the PedsQL 4.0 Generic Core Scales. Health and Quality of Life Outcomes. 2007; 5(1)10.1186/1477-7525-5-1

Ware JE. Conceptualization and measurement of health-related quality of life: Comments on an evolving field. Archives of Physical Medicine and Rehabilitation. 2003; 84(supplement 2):S43–S51. [PubMed: 12692771]

Watson D, Tellegen A. Toward a consensual structure of mood. Psychological Bulletin. 1985; 98:219–235. [PubMed: 3901060]

Watson D, Clark LA. Negative affectivity: The disposition to experience aversive emotional states. Psychological Bulletin. 1984; 96:465–490. [PubMed: 6393179]

Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA. Testing a tripartate model I: Evaluating the convergent and discriminant validity of anxiety and depression symptoms scales. Journal of Abnormal Psychology. 1995a; 104:3–14. [PubMed: 7897050]

Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA. Testing a tripartite model II: Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. Journal of Abnormal Psychology. 1995b; 10:15–25. [PubMed: 7897037]

Weiss DJ. Computerized adaptive testing for effective and efficient measurement in counseling and education. Measurement and Evaluation in Counseling and Development. 2004; 37:70–84.

Weissman MM, Orvaschel H, Padian N. Children's symptom and social functioning self-report scales. The journal of nervous and mental disease. 1980; 168(12):736–740. [PubMed: 7452212]

Wouters H, van Campen J, Appels B, Lindeboom R, Buiter M, de Haan RJ, Zwinderman AH, van Gool WA, Schmand B. Does adaptive cognitive testing combine efficiency with precision? Prospective findings. Journal of Alzheimer's Disease. 2011; 25(4):595–603.

Yeh M, Weisz JR. Why are we here at the clinic? Parent-child (dis)agreement on referral problems at outpatient treatment entry. Journal of Consulting and Clinical Psychology. 2001; 69(6):1018–1025. [PubMed: 11777105]

Zumbo, BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Directorate of Human Resources Research and Evaluation, Department of National Defense; Ottawa, ON: 1999.

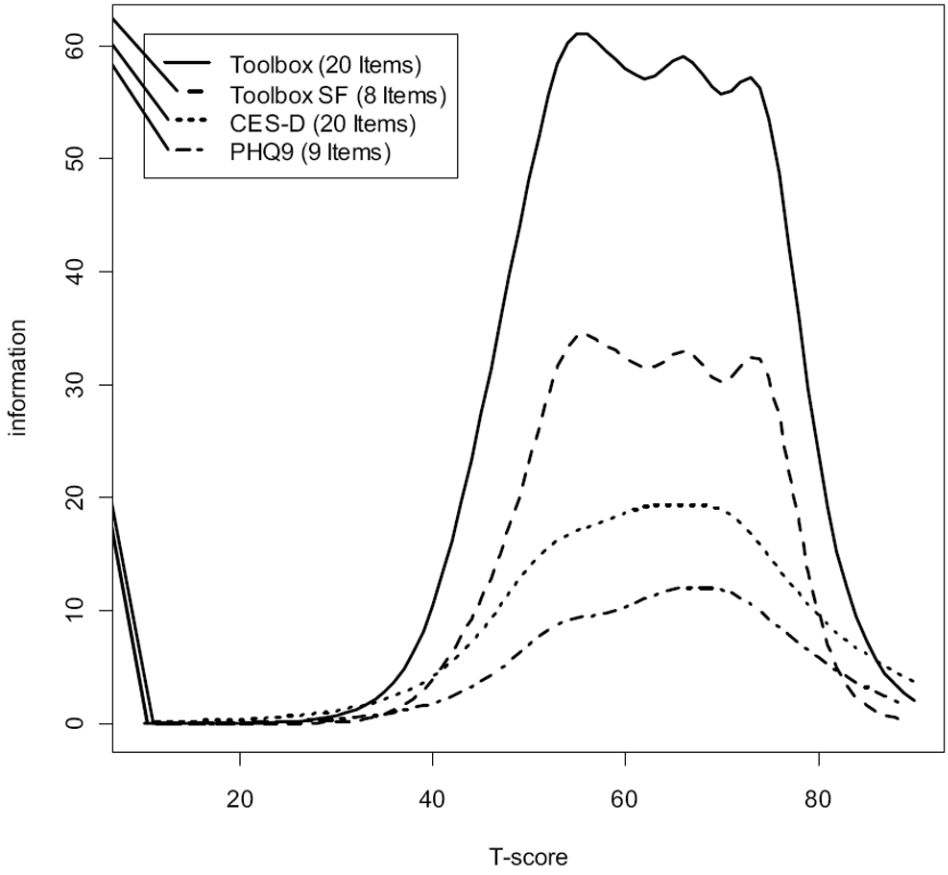

Figure 1.
Test information curves for adult sadness: 20-item bank, 8-item short form, CES-D, and PHQ-9.

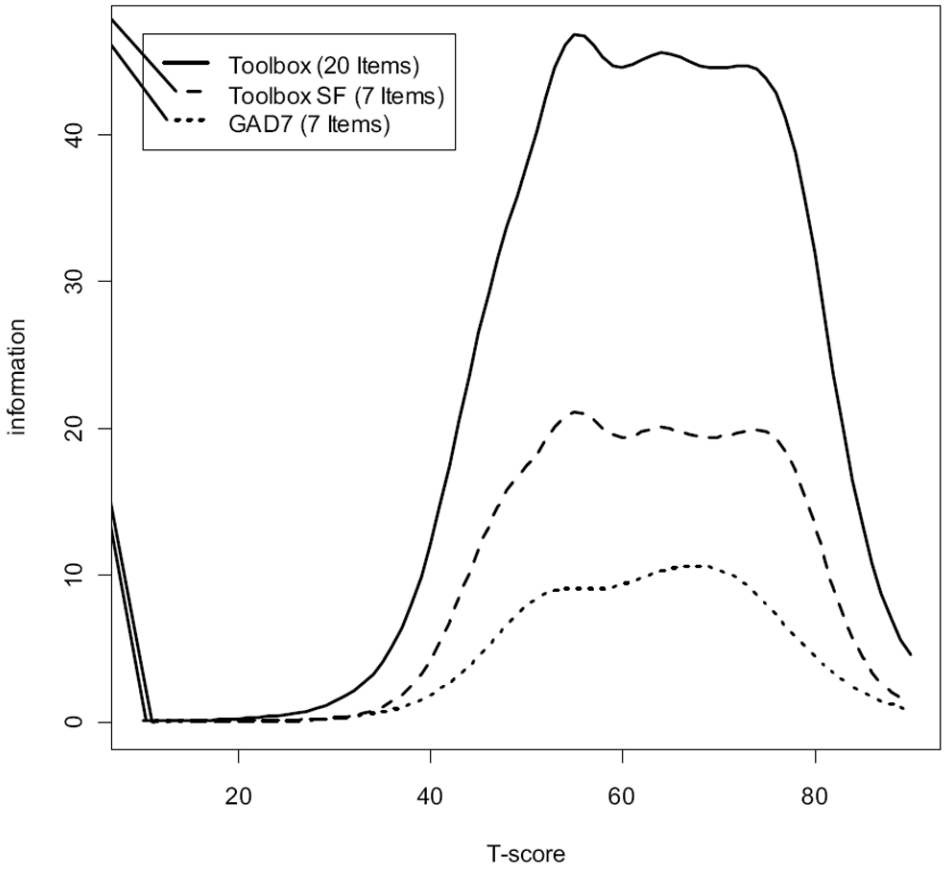

Figure 2.
Test information curves for adult fear: 20-item bank, 7-item short form, and GAD-7.

Table 1

Summary of Instruments

Measures of Sadness			
Measure	**Key Citations**	**Age Range**	**Key Features**
PROMIS Depression Item Bank	Pilkonis et al., 2011	18+	20 items, “past seven days”, 5-point scale for frequency
Patient Health Questionnaire (PHQ-9)	Kroenke et al., 2001	18+	9 items, “past two weeks”, 4-point scale for duration
Center for Epidemiologic Studies Depression Scale (CES-D)	Radloff, 1977	18+	20 items, “past week”, 4-point scale for duration
PROMIS Pediatric Depression Item Bank	Irwin et al., 2010a	8 - 17	14 items, “past seven days”, 5-point scale for frequency
The Short Mood and Feelings Questionnaire (SMFQ)	Angold et al., 1995	8 - 17	13 items, “past two weeks”, 3-point scale for true/not true
Center for Epidemiologic Studies Depression Scale Child Form (CES-DC)	Weissman et al.,1980	8 - 17	20 items, “past week”, 4-point scale for frequency
Measures of Fear			
Measure	Key Citations	Age Range	Key Features
PROMIS Anxiety Item Bank	Pilkonis et al., 2011	18+	20 items, “past seven days”, 5-point scale for frequency
Generalized Anxiety Disorder 7 item Scale (GAD-7)	Spitzer et al., 2006	18+	7 items, “past two weeks”, 4-point scale for duration
Mood and Anxiety Questionnaire (MASQ)	Watson et al., 1995a, 1995b	18+	28 items from the General Distress (Anxiety) and Anxious Arousal subscales, “past week”, 5-point scale for severity
PROMIS Pediatric Anxiety Item Bank	Irwin et al., 2010a	8 - 17	15 items, “past seven days”, 5-point scale for frequency
Screen for Child Anxiety Related Emotional Disorders (SCARED)	Birmaher et al., 1997	8 -17	41 items, “past three months”, 3-point scale for true/not true
Measures of Anger			
Measure	Key Citations	Age Range	Key Features
PROMIS Anger Item Bank	Pilkonis et al., 2011	18+	20 items, “past seven days”, 5-point scale for frequency
Buss-Perry Aggression Questionnaire (BPAQ)	Buss and Perry, 1992	18+	29 items, no time frame, 7-point scale for characteristic/uncharacteristic
PROMIS Pediatric Anger Scale	Irwin et al., 2010b	8 - 17	6 items, “past seven days”, 5-point scale for frequency
Anger Expression Scale for Children (AESC)	Jacobs et al., 1989	8 - 17	12 items for the subjective experience of anger, no time frame, 4-point scale for frequency

Table 2

Demographic characteristics of the NIH Toolbox sample

	Pediatric Age 8-17 (*N* = 1,015)	Adult Age 18+ (*N* = 748)
	%	%
Sex		
Male	49.4	43.9
Ethnicity		
Hispanic	9.8	15.2
Race		
American Indian/Alaska Native	2.1	2.8
Asian American	2.1	2.8
Black/African American	11.6	9.1
Native Hawaiian/Other Pacific Islander	0.5	0.9
White	82.5	80.1
Other	4.8	6.4
Education		
High school diploma, GED, or vocational/technical training	NA	32.0
Further educational attainment	NA	68.0
Mean Age	12.5	47.2

Note. The percentages for race total more than 100% because 34 pediatric participants (3.3%) endorsed more than one race and 16 adult participants (2.1%) endorsed more than one race.

Table 3

Toolbox Sadness Test

Adult Items
I felt hopeless.*
I felt depressed.*
I felt worthless.*
I felt helpless.*
I felt like a failure.*
I felt that I had nothing to look forward to.*
I felt unhappy.*
I felt sad.*
I felt that my life was empty.
I felt discouraged about the future.
I found that things in my life were overwhelming.
I felt disappointed in myself.
I felt that nothing was interesting.
I withdrew from other people.
I felt emotionally exhausted.
I had trouble making decisions.
I felt lonely.
I had trouble feeling close to people.
I felt pessimistic.
I felt ignored by people.
Pediatric Items
I could not stop feeling sad.*
I felt everything in my life went wrong.*
I felt like I couldn’t do anything right.*
I felt unhappy.*
I felt alone.*
I felt lonely.*
I thought that my life was bad.*
I felt sad.*
Being sad made it hard for me to do things with my friends.
It was hard for me to have fun.
I felt too sad to eat.
I felt stressed.
I didn’t care about anything.
I wanted to be by myself.

Note. Items included in the short form are marked with an asterisk. Items are reprinted with the permission of the PROMIS Health Organization and the PROMIS Cooperative Group.

Table 4

Toolbox Fear Test

Adult Items	
	I found it hard to focus on anything other than my anxiety.*
	My worries overwhelmed me.
	I felt uneasy.*
	I felt fearful.*
	I felt frightened.
	I felt nervous.*
	I felt anxious.*
	I felt tense.*
	Many situations made me worry.
	I felt worried.*
	I had sudden feelings of panic.
	I was concerned about my mental health.
	I felt upset.
	I felt indecisive.
	I had trouble relaxing.
	I had trouble paying attention.
	I felt fidgety.
	I was easily startled.
	I worried about other people's reactions to me.
	I had difficulty sleeping.
	Somatic Arousal
	Felt dizzy or lightheaded.
	Muscles were tense or sore.
	Felt nauseous.
	Heart was racing or pounding.
	Was short of breath.
	Muscles twitched or trembled.
Pediatric Items	
	I felt scared.*
	I worried about what could happen to me.*
	I worried when I went to bed at night.*
	I felt worried.*
	I felt like something awful might happen.*
	I was worried I might die.
	I woke up at night scared.
	I worried when I was at home.
	I felt nervous.*
	I thought about scary things.*
	I got scared really easy.

Adult Items
I was afraid that I would make mistakes.*
It was hard for me to relax.
I worried when I was away from home.
I was afraid of going to school.

Note. Items included in the short form are marked with an asterisk. The general fear items are reprinted with the permission of the PROMIS Health Organization and the PROMIS Cooperative Group.

Table 5

Toolbox Anger Test

Adult Items	
	I was grouchy.*
	I stayed angry for hours.*
	I felt angry.*
	I felt like I was ready to explode.*
	I felt angrier than I thought I should.*
	I felt annoyed.*
	I made myself angry about something just by thinking about it.*
	When I was angry, I sulked.
	I was irritated more than people knew.*
	I felt like yelling at someone.
	I was stubborn with others.
	Even after I expressed my anger, I had trouble forgetting about it.
	I felt bitter about things.
	I felt resentful when I didn't get my way.
	I had trouble controlling my temper.
	I felt that people were trying to anger me.
	I felt guilty about my anger.
	I was angry when something blocked my plans.
	I felt envious of others.
	I disagreed with people.
	Physical Aggression
	Given enough provocation, I may hit another person.
	I get into fights a little more than the average person.
	There are people who pushed me so far that we came to blows.
	I have threatened people I know.
	I have become so mad that I have broken things.
	Hostility
	I am sometimes eaten up with jealousy.
	At times I feel I have gotten a raw deal out of life.
	Other people always seem to get the breaks.
	I wonder why sometimes I feel so bitter about things.
	I sometimes feel that people are laughing at me behind my back.
Pediatric Items	
	I felt mad.
	I was so angry I felt like yelling at somebody.
	I was so angry I felt like throwing something.
	When I got mad, I stayed mad.
	I felt upset.
	I felt fed up.

Note. Items included in the short form are marked with an asterisk. The general anger items are reprinted with the permission of the PROMIS Health Organization and the PROMIS Cooperative Group.

Table 6

Correlations among Toolbox Measures for Adults

	No. Items	1	2	3	4	5	6
1. BPAQ – Hostility	6	1.0					
2. PROMIS – Anger short form	8	0.59	1.0				
3. PROMIS – Anxiety short form	7	0.55	0.70	1.0			
4. PROMIS – Depression short form	8	0.63	0.72	0.81	1.0		
5. BPAQ – Physical aggression	5	0.52	0.48	0.37	0.41	1.0	
6. MASQ – Somatic arousal	6	0.50	0.57	0.68	0.62	0.44	1.0