

# Screening for Problematic Worry in Adults With a Single Item From the Penn State Worry Questionnaire

Assessment  
1–11  
© The Author(s) 2017  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1073191117694453  
journals.sagepub.com/home/asm  


Hans S. Schroder<sup>1</sup>, D. Angus Clark<sup>2</sup>, and Jason S. Moser<sup>1</sup>

## Abstract

As the rapid assessment of mental health is a growing need, a quick and valid tool for the early detection of symptoms that can be flexibly deployed across a range of contexts may be especially beneficial. This is particularly true of anxiety problems, which when undetected contribute to health care costs and lost work productivity. Data from more than 10,000 respondents (primarily female undergraduates) were used to test whether a single item from the popular Penn State Worry Questionnaire could serve as a screening tool in settings where administration of the full scale is undesirable. Items were evaluated by examining item response theory models, screening capabilities, stability over time, convergence with other anxiety and depression measures, and a response time analysis that assessed how quickly participants responded to each item. Item 15 (“I worry all the time”) emerged as the strongest item: It was the most discriminating and reliable item, had sensitivity and specificity similar to the full scale, had the highest 1-month and 1-year retest coefficients, the highest convergent correlations with measures of anxiety and depression, and was responded to significantly faster than any other item. We suggest that in time-limited contexts, this item is suitable for screening.

## Keywords

mental health screening, item response theory, worry, Penn State Worry Questionnaire

Undetected and undertreated mental health problems are a leading cause of poor health and morbidity, and can exact a severe psychological toll on sufferers and their families (Prince et al., 2007). Despite this, mental health problems are typically not routinely assessed in clinical settings alongside other universally collected markers of health and well-being, such as blood pressure (Klinkman, 1997; Ozer et al., 2009; Tedeschi et al., 2016; Zenlea, Milliren, Mednick, & Rhodes, 2014; Zimmerman, 2016; Zimmerman & Wienckowski, 1991). The lack of quality, routine mental health screening represents an unfortunate instance of the research-to-practice gap in modern times: Although myriad high-quality assessments of personality traits and clinical symptoms exist and are widely used in research, their actual use in routine clinical settings is rare (Kroenke, Spitzer, & Williams, 2001; McHorney & Tarlov, 1995; Zimmerman, 2016).

This is likely partly due to the fact that many psychological assessments are rather lengthy, and may involve complicated scoring algorithms (in contrast to the general swiftness and ease of the aforementioned blood pressure assessments; Zimmerman & McGlinchey, 2008). Given these practical considerations, exceptionally brief, specifically *single-item*, mental health screening measures are potentially valuable tools that could be easily incorporated across a diverse range of settings to serve as a “first line of defense” for

detecting possible mental health issues (e.g., Smith, Schmidt, Allensworth-Davies, & Saitz, 2010). To be sure, ultrabrief measures of psychological distress are particularly appealing to practitioners in primary care settings who see many patients/clients per day (Klinkman, 1997; Mitchell, Kaar, Coggan, & Herdman, 2008). Indeed, it was to this end that the World Health Organization recently launched an interdisciplinary initiative to develop rapid and precise measurement tools for assessing mental health problems across a wide array of contexts (e.g., primary care, schools; World Health Organization & World Organization of Family Doctors, 2008).

In order to contribute to this effort, the present study sought to provide evidence for a single-item measure that could potentially serve as an early screening tool for *problematic worry*, a key transdiagnostic component of many anxiety and depressive disorders (Newman, Llera, Erickson, Przeworski, & Castonguay, 2013). When untreated, these

<sup>1</sup>Michigan State University, East Lansing, MI, USA

<sup>2</sup>Texas A&M University, College Station, TX, USA

## Corresponding Author:

Hans S. Schroder, Department of Psychology, Michigan State University, 316 Physics Road, Room 69E, East Lansing, MI 48823, USA.  
Email: hans.stuart.schroder@gmail.com

disorders—which are among the most common of all mental health disorders—contribute to billions of dollars each year in lost worker productivity and other health care costs (Baxter, Scott, & Whiteford, 2013; Hoffman, Dukes, & Wittchen, 2008). Thus, quick, routine assessment of problematic worry in primary care and related settings could prove immensely helpful for identifying cases that may benefit from more specialized mental health services. That is, the implementation of a single-item worry screening questionnaire alongside other typical clinical assessments would be a low-cost investment for both patients and practitioners that could help many individuals receive services they may otherwise not have.

Although many questionnaires for measuring anxiety and worry have been developed (e.g., Gladstone et al., 2005; Stober & Joormann, 2001), they were developed primarily for research purposes, and not routine clinical assessment. The most popular inventory of problematic worry used in research is the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990).<sup>1</sup> On the PSWQ, respondents use a 1- to 5-point Likert-type scale (1 = *not at all like me*, 5 = *very much like me*) to rate the extent to which 16 statements regarding problematic worry apply to themselves, 11 of which are positively scored (e.g., “Once I start worrying, I cannot stop”) and five of which are negatively scored (e.g., “I do not tend to worry about things”). The PSWQ has undergone extensive psychometric evaluation and is generally accepted to be a reliable and valid measure of problematic worry (e.g., Behar, Alcaine, Zuellig, & Borkovec, 2003; Brown, 2003).

Importantly, the PSWQ treats problematic worry as a single (i.e., unidimensional, pending certain “method factors”; Brown, 2003) continuous construct, rather than a set of symptoms that are tied to a specific diagnostic nosology (e.g., GAD-7; Spitzer, Kroenker, Williams, & Lowe, 2006). However, total PSWQ scores that exceed an empirically based clinical cutoff score (score of 45 or 61 for outpatients and college students, respectively; Behar et al., 2003; Fresco et al., 2002) indicate that individuals are *at risk* for having an anxiety disorder (see Kanuri, Taylor, Cohen, & Newman, 2015), which makes the PSWQ a useful screening measure from which additional assessments could be administered if warranted. Given the PSWQ’s popularity, psychometric quality, and clearly defined risk threshold, in this article, we specifically investigated whether a single item from this scale could be used to adequately and quickly screen for problematic worry in situations where a lengthier psychological screening is either not desired or possible.

At 16 items, the PSWQ may seem relatively brief; however, the full questionnaire is still likely impractical for routine clinical use. Beyond the fact that 16 items may be too lengthy for routine use in clinical settings, the PSWQ also includes several reverse-scored items (e.g., “I do not tend to worry about things”), which, although useful for detecting

inattentive responding (Schmitt & Stults, 1985), may confuse respondents and bias estimates. Reversed items may also introduce more scoring errors, especially in contexts when rapid scoring is necessary. Furthermore, many of the items on the PSWQ are redundant (e.g., “I worry all the time” and “I am always worrying about something”). Including highly similar items will increase internal consistency, but can be needlessly time consuming, and often frustrate respondents (Robins, Hendin, & Trzesniewski, 2001). In busy, real-world clinical settings, time management is paramount, as is minimizing patient burden and frustration. As such, it is ideal to reduce redundancy as much as possible in screening situations.

Previous studies have attempted to alleviate these issues by constructing briefer versions of the PSWQ. To date, shortened forms of the PSWQ consisting of 10 (Yao et al., 2016), 8 (Hopko et al., 2003), 5 (Topper, Emmelkamp, Watkins, & Ehring, 2014), and even 3 (Berle et al., 2011) items have been developed. Although all these brief forms eliminated the reverse-scored items, they did not eliminate redundancy. For example, the three-item version (Berle et al., 2011) contains the two items: “I worry all the time” and “Many situations make me worry.” Indeed, given that the PSWQ represents a unidimensional measure of worry, there is inherent redundancy so long as multiple items are included. Furthermore, these earlier evaluations of abbreviated versions of the PSWQ mostly relied on samples of patients already enrolled in anxiety disorder clinics. Although clinical samples are helpful in assessing correspondence between screening instruments and formal clinical diagnoses, they are restricted in their range of symptom severity, and are thus not ideally suited for the purpose of calibrating the large-scale *screening* tools that are needed for various settings such as primary care clinics or schools (but see Topper et al., 2014). That is, clinical samples have limited potential in distinguishing individuals with lower versus higher risk of developing an anxiety or depressive disorder.

Previous efforts to pare down the PSWQ (see also Kertz, Lee, & Bjorgvinsson, 2014) have also generally failed to take advantage of more powerful, modern psychometric techniques, such as those represented by item response theory (IRT). IRT represents a collection of latent variable measurement models that provide numerous details regarding psychometric functioning at both the item and test level. IRT offers many advantages over the more widely used classical test theory (CTT), which despite its popularity is limited and problematic both practically and theoretically (Borsboom, Cramer, Kievit, Scholten, & Franic, 2009). For example, unlike in CTT, IRT measurement parameters are generally sample invariant (de Ayala, 2009; Hambleton & Swaminathan, 1985; Markus & Borsboom, 2013). Furthermore, IRT provides a more comprehensive approach to the issue of reliability

such that test and item precision is conditional on the attribute being assessed (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). That is, measurement precision varies across levels/severity of the construct of interest, such that an anxiety questionnaire may be more reliable at measuring high-anxiety individuals than low-anxiety individuals. One of the most important conceptual advances from IRT is the acknowledgment that more items do not necessarily beget better, more reliable measurement (a foundational principal behind computer adaptive testing; de Ayala, 2009; Hambleton & Swaminathan, 1985). Lengthy measures may contain several items that contribute virtually nothing to the measurement of the target construct (as a consequence of either general weakness or redundancy), whereas a small collection of items, or even single items, may provide an adequately precise assessment. Given all of this, IRT models are especially well suited for developing and evaluating short forms, and identifying single items that could serve as adequate indicators in situations where more lengthy assessment is not undertaken.

In this study, we sought to determine whether there was one item of the PSWQ that could serve as a reasonable screening tool for problematic worry in the general population for situations where timing is of the essence (e.g., primary care, school settings). We emphasize that our intention here is not to suggest a single item should replace the full scale in all contexts (e.g., research), or that a single item can capture all aspects of anxiety and other internalizing pathologies; rather, we simply aim to provide a tool for practitioners and researchers across diverse settings to quickly and easily gain some insight into individuals' propensity for problematic worry for the purposes of initial screening and early detection.

To be sure, although historically single-item assessments have been cautioned against, there is increasing recognition of their benefits and capabilities, thanks in part to ongoing psychometric advances, both theoretical and technological (Cheung & Lucas, 2014; Robins et al., 2001). Though single-item measures must be thoroughly vetted, they are often adequate stand-ins for full questionnaires when brevity is required, and there are in fact instances of single-item measures performing *at least* as well in many respects (e.g., predictive power) as their full, parent measure (e.g., Cheung & Lucas, 2014; Robins et al., 2001). Notably, Robins et al. (2001) discuss scenarios in which the distillation of a full questionnaire down to a single item is appropriate. Specifically, they mention that single-item measures are appropriate when the target constructs are highly schematized, unidimensional, and largely reflect subjective experience. As the PSWQ generally fulfills these requirements, it is both reasonable and appropriate to ask whether a single item can potentially serve as a suitable stand-in for the full form in certain circumstances.

In the present study, the identification and evaluation of a single-item measure of problematic worry from the PSWQ consisted of five major steps. First, we used the graded response model (an IRT model for polytomous items; Samejima, 1969) to analyze the PSWQ in a large calibration sample ( $N = 9,040$ ) to identify possible candidate items. At this stage, we also considered how responses to notable items related to the PSWQ clinical cutoff score. Second, we attempted to confirm the IRT analyses with evaluations of the items' diagnostic screening capabilities (e.g., sensitivity, specificity) in an independent validation sample ( $N = 1,191$ ). Third, we examined test-retest and stability coefficients, for individual items and the full-scale PSWQ, at both 1-month and 1-year intervals. Fourth, we evaluated convergent correlations between PSWQ items and other measures of worry, anxiety, and depression. Finally, we considered a more practical aspect of assessment by conducting a response time (RT) analysis of the individual items of the PSWQ. Ideally, a single-item measure of problematic worry would be (a) discriminative and informative, (b) specific and sensitive in screening, (c) stable across time, (d) relatively predictive of other measures of worry, anxiety, and depression compared with other items and the full scale, and (e) the quickest to respond to.

## Method

### Participants

The calibration sample comprised respondents who completed the full PSWQ between 2009 and 2016, at a large Midwestern university ( $N = 9,040$ ). Participants were undergraduates and community members. Most of these data came from online mass screening sessions that did not include additional demographic data. The portion of the sample that did include this information ( $N = 1,893$ ) indicated that 62% of the sample was female.

### Item Response Theory Evaluation of PSWQ

Responses on the PSWQ were initially analyzed with the graded response model (Samejima, 1969), an IRT model for polytomous items, using *Mplus* version 7.4 (Muthén & Muthén, 1998-2012). IRT models provide a number of useful metrics and parameters for assessing the properties of individual items. Most useful here are the discrimination parameters and the item information curves. Discrimination parameters are analogous to factor loadings. They denote how strongly an item is related to the latent trait; higher values signify that the item is better at differentiating between individuals at different levels of the latent trait. Typically, a discrimination value of .80 is considered the minimum acceptable value (a discrimination value of .80 roughly corresponds to a standardized factor loading of

around .43; de Ayala, 2009). Information curves on the other hand provide insight into how precisely an item measures individuals at different levels of the latent trait. Raw information values are in difficult-to-interpret logit units, but they can easily be converted into more meaningful estimates of standard error ( $SE$ ) and reliability. For example, an information value of 3 at some point along on latent trait continuum corresponds to a  $SE$  of .58 ( $1/\sqrt{3}$ ), which in turn can be converted into a rough conditional reliability estimate of .66 ( $1 - .34$ ; alternatively  $1 - (1/3)$ ; Thissen & Orlando, 2001). Again, these three complimentary metrics of measurement precision are conditional on the latent trait such that precision differs along the spectrum of worry.

### Screening Potential

To further evaluate the screening potential of the PSWQ items, an additional validation sample was collected after the initial IRT analyses were conducted. It consisted of 1,191 undergraduates (73% female) who completed the full PSWQ and the GAD-7, a seven-item self-report measure of GAD symptom severity (Spitzer et al., 2006). The full PSWQ scale and selected items from the IRT analyses were used to predict likely diagnostic status for GAD—operationalized as a score of 10 or higher on the GAD-7 (Spitzer et al., 2006). We examined the receiver operating characteristic (ROC) curve, sensitivity, specificity, positive predictive power (PPP), and negative predictive power (NPP). Sensitivity measures the proportion of positives that are correctly identified as such, whereas specificity measures the proportion of negatives that are correctly identified as such. PPP is the probability that an individual is disordered, given that the test has identified that individual as disordered (i.e., the number of true positive cases divided by the total number of cases the test indicated as being positive), whereas NPP is the probability that an individual is nondisordered, given that the test has identified that individual as such. Together, these indices provide complimentary information on the screening utility of PSWQ items (see also Behar et al., 2003).

### Temporal Stability

The remainder of the analyses was conducted primarily within subsamples of the calibration data set. To evaluate temporal stability, test-retest coefficients were computed for individual items and the full PSWQ using both a subset of the participants that completed two PSWQs approximately 1 month apart ( $N = 165$ ,  $M = 27.77$  days,  $SD = 1.41$ , range: 27-39 days), and another subset of participants ( $N = 34$ ) who completed two PSWQs about 1-year apart ( $M = 13.51$  months,  $SD = 2.38$ , range: 8-18 months). Previous studies have found 1-month test-retest coefficients of .87 (Stober, 1998) and .93 (Meyer et al., 1990); we are not

aware of previous work examining retest coefficients over 1 year.

### Convergent Validity

The degree to which individual PSWQ items, the full scale, and the three-item scale related to other well-validated measures of worry, anxiety, and depression was examined next in subsamples with available data (subsample sizes are also provided in the note of Table 3). We examined the three-item scale as well because this is the briefest PSWQ iteration currently in the literature. There were five-criterion measures included for this part of the analysis.

The GAD-7 (Spitzer et al., 2006) is a seven-item scale of GAD symptom severity (e.g., “feeling nervous, anxious, or on edge”) scored on a 0 “not at all” to 3 “nearly every day” regarding the past 2 weeks. Data for the GAD-7 were available from 1,191 participants from the validation data set ( $M = 4.43$ ,  $SD = 4.44$ , current  $\alpha = .91$ ).

The Worry Domains Questionnaire–Short Form (WDQ-SF; Stober & Joormann, 2001) assesses worry in five domains: relationships, lack of confidence, aimless future, financial issues, and work (e.g., “I worry that I will not keep my workload up to date”). Data for the WDQ-SF were available from 400 to 405 participants ( $M = 15.20$ ,  $SD = 9.63$ ; current  $\alpha = .92$ ).

The State–Trait Anxiety Inventory–Trait Version (STAI; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983), which consists of 20 items (e.g., “I am tense”), is one of the most commonly used measures of trait anxiety. Data for the STAI were available for 1,137 to 1,114 participants ( $M = 41.86$ ,  $SD = 9.77$ ; current  $\alpha = .92$ ).

The 17-item Mood and Anxiety Symptom Questionnaire–Anxious Arousal subscale (MASQ-AA; Watson & Clark, 1991) asks participants to rate how much they experienced symptoms of somatic anxiety (e.g., “hands were cold or sweaty”) over the past week, including today. Data for the MASQ-AA were available from 1,949 to 1,960 participants ( $M = 28.65$ ,  $SD = 10.96$ ; current  $\alpha = .92$ ).

The 21-item MASQ–Anhedonic Depression subscale (MASQ-AD; Watson & Clark, 1991) is similar in format to the MASQ-AA but asks about anhedonia (e.g., “felt really happy,” reverse scored). Data for the MASQ-AD were available from 1,949 to 1,960 participants ( $M = 53.17$ ,  $SD = 14.02$ ; current  $\alpha = .92$ ).

### Response Time Analysis

Finally, an RT analysis was conducted on individual PSWQ items in order to evaluate the total time it takes individuals to complete the PSWQ. This analysis included a small subsample of the participants ( $N = 154$ )<sup>2</sup> who completed the PSWQ using a computer program that presented items individually while RTs were measured (using E-Prime 2.0

**Table 1.** Item Discrimination Values and Information Along the Worry Continuum.

Item	$\alpha$	Level of worry in standard deviation units										
		-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5
1	0.94	0.27	0.28	0.28	0.28	0.28	0.27	0.26	0.24	0.21	0.17	0.12
2	2.32	0.57	1.16	1.49	1.52	1.57	1.56	1.48	1.44	1.32	0.77	0.31
3	1.74	0.83	0.86	0.90	0.92	0.91	0.86	0.85	0.76	0.54	0.30	0.14
4	2.64	0.45	1.19	1.83	1.74	1.93	1.96	1.95	1.79	1.82	1.12	0.41
5	2.88	0.44	1.33	2.26	2.28	2.44	2.38	2.12	2.16	1.21	0.38	0.10
6	1.97	0.93	1.18	1.1	1.15	1.16	1.1	1.05	1.04	0.79	0.43	0.19
7	2.97	0.13	0.52	1.56	2.35	2.28	2.55	2.56	2.4	2.2	1.03	0.29
8	1.70	0.76	0.80	0.81	0.85	0.86	0.85	0.81	0.79	0.75	0.57	0.34
9	1.80	0.48	0.76	0.89	0.91	0.95	0.98	0.97	0.94	0.90	0.73	0.45
10	1.67	0.79	0.83	0.86	0.86	0.83	0.74	0.54	0.32	0.16	0.07	0.03
11	1.10	0.32	0.35	0.36	0.37	0.37	0.37	0.37	0.36	0.35	0.33	0.30
12	2.15	0.24	0.58	1.06	1.32	1.38	1.41	1.4	1.35	1.18	0.73	0.32
13	2.50	0.84	1.54	1.69	1.76	1.83	1.74	1.5	1.62	1.38	0.62	0.21
14	2.42	0.26	0.71	1.37	1.57	1.59	1.71	1.71	1.63	1.6	1.17	0.53
<b>15</b>	<b>3.40</b>	<b>0.03</b>	<b>0.16</b>	<b>0.78</b>	<b>2.46</b>	<b>3.05</b>	<b>3.24</b>	<b>3.22</b>	<b>3.04</b>	<b>3.06</b>	<b>1.7</b>	<b>0.42</b>
16	1.66	0.72	0.79	0.81	0.84	0.84	0.82	0.79	0.76	0.61	0.39	0.21

Note.  $\alpha$  = discrimination values. Information presented in 0.5 standard deviation increments around the mean of the latent worry factor. Item 15 is bolded for emphasis.

software). The purpose here was simply to provide a metric of assessment time that practitioners might find useful, given the clinical relevance of client burden.

## Results

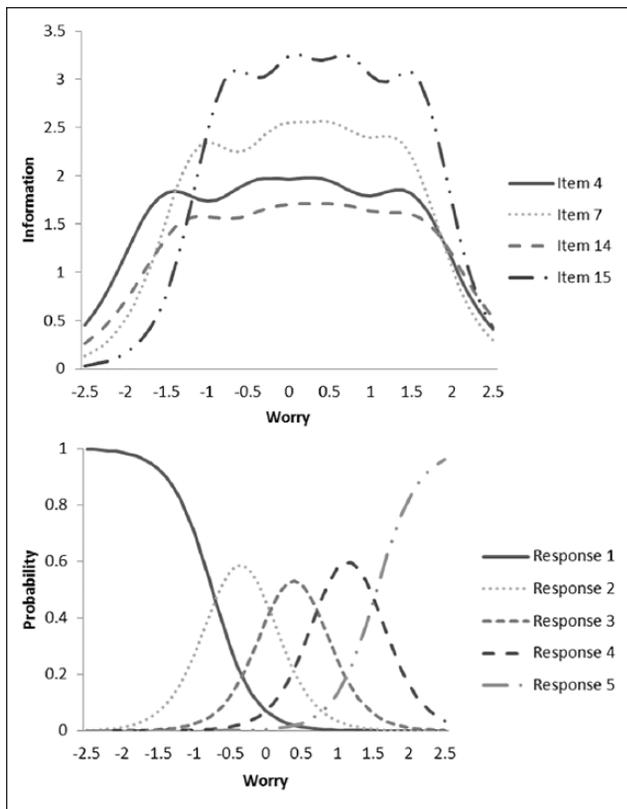
### Graded Response Model

The graded response model assumes a unidimensional construct, and in line with previous work and theory (Brown, 2003), preliminary item factor analyses supported a single-factor solution. Notably, however, there was some evidence of a potential secondary, method factor that the reverse-scored items loaded on. This auxiliary clustering could represent a violation of another model assumption, specifically that of local dependence (i.e., no residual correlations between indicators after conditioning on the latent factor). However, supplemental analyses suggested that this is likely not a major issue for the present analyses. Specifically, item parameters are roughly the same when reverse-scored items are versus are not included in the graded response model (deviations from 0% to 10%;  $M = 3\%$ ). Furthermore, a brief battery of simulations were run in which unidimensional graded response models were fit to 1,000 data sets generated from a population bifactor model (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012) based on the actual data (where there were two method factors, one for positively scored items, and one for negatively scored items). That is, these simulations can illustrate the bias that might result from fitting the graded response model to data with a

degree of local independence analogous to the actual data. The results of these simulations suggest that the degree of local dependence in the current data only introduces a relatively trivial amount of bias, with parameter bias in the estimated parameters ranging from 0% to 10% ( $M = 2\%$ ). The correlation matrix for all of the items in the calibration sample is available in the supplemental material (<http://asm.sagepub.com/content/by/supplemental-data>).

The current model was specified such that all 16 PSWQ items loaded on a single latent “worry” dimension. The mean and variance of the latent dimension were set to 0 and 1, respectively, to identify the model and place “worry” on the standard normal metric. Item discrimination values, and information values at selected levels of the worry dimension, are shown in Table 1. Of all the items, Item 15 (“I worry all the time”) had the highest discrimination value, and was the most informative item from the mean to 2 standard deviations above the mean. Thus, Item 15 is the most strongly related to the dimension of worry, and is able to measure individuals around and above the average level of worry with considerably more precision than other items.

Item 15’s information curve is displayed alongside the second most informative item (Item 7), and the two other items that make up the three-item PSWQ (Items 4 and 14) in Figure 1. Here, it is easy to see how Item 15 surpasses other notable items in terms of measurement precision across most levels of worry considered (between  $-2.5$  and  $2.5$  standard deviations from the mean). Indeed, this figure demonstrates that the reliability of the three-item form is mostly a function of the inclusion of Item 15; Item 15 is



**Figure 1.** Item information and characteristic curves.  
 Note. Top = item information curves for four items; bottom = item characteristic curve for Item 15 (“I worry all the time”). x-Axis represents worry in standardized units. Response 1 corresponds to “not at all like me” and Response 5 corresponds to “very much like me.”

much more informative than the other two items, and Item 15 by itself is at least as informative as the other two items combined across much of the trait.

These findings suggest Item 15 as a potential candidate for a single-item measure of worry. Of course, it is not enough for one item to merely be more informative than the other items; it also has to be adequately precise in the absolute sense. Item 15 has an information value of at least 3 (reliability estimate of .66) between  $-0.5$  and  $1.5$  standard deviations from the mean. This implies that worry can be assessed relatively precisely for a substantial portion of the population using Item 15 alone. Specifically, because worry is assumed to be normally distributed, Item 15 achieves a level of reliability of at least .66 for around 63% of the total population, and 86% of the population that falls above the average level of worry, which is the range of the population most relevant for screening purposes.

The potential utility of Item 15 is further illustrated by its item characteristic curve, depicted in the bottom portion of Figure 1. The item characteristic curve graphs the probability of endorsing a given item category at different levels of the latent trait. As shown in Figure 1, at around 1.2 standard deviations above the mean, individuals are most likely to

endorse either Category 4 or 5. Starting around 2 standard deviations above the mean, individuals are most likely to endorse Category 5. Therefore, individuals who endorse either a Category 4 or 5 on this item are likely to be at least 1 standard deviation above the average level of worry in the general population. Thus, although fine-grained distinctions between levels of worry may be more difficult to make above 1.5 standard deviations based on this single item, if an individual endorses Response Option 4 or 5, it is still possible to discern that they are likely to be substantially more worried than the average person. To be sure, we believe this level of precision is more than sufficient for many settings that would simply seek to identify individuals with problematic levels of worry for potential follow-up (as opposed to actually diagnosing, or researching, them).

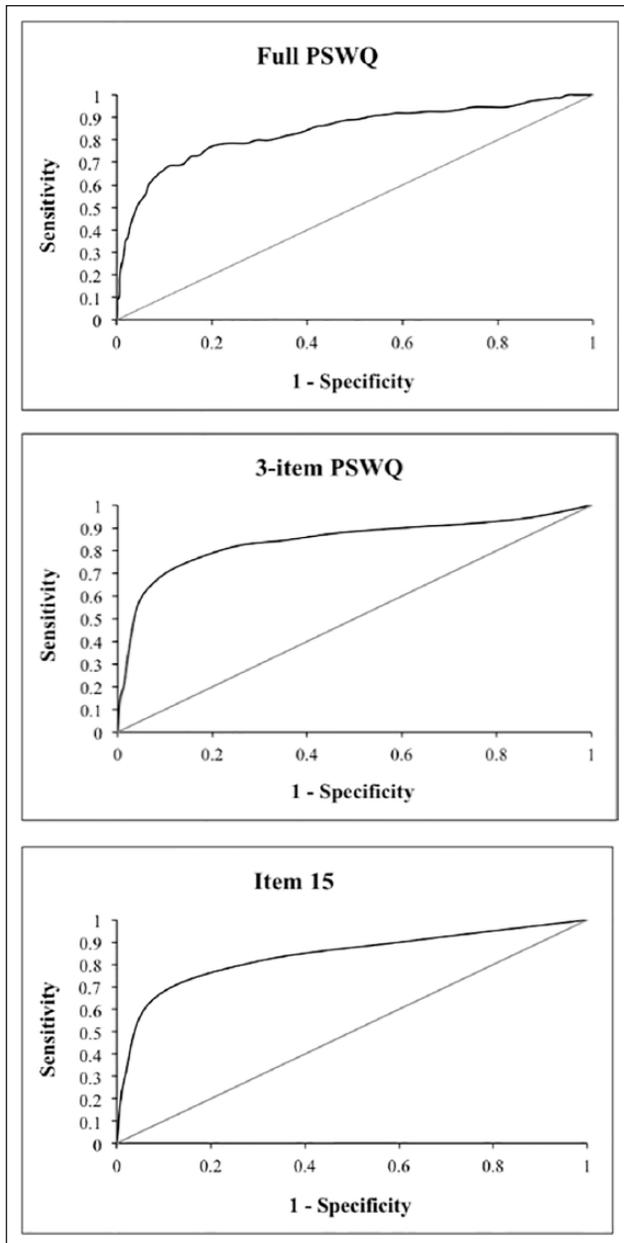
Reinforcing this notion, it is noteworthy that in the calibration sample, 73% of respondents that endorsed Category 4 on Item 15 had total PSWQ scores above the recommended “clinical cutoff” (Behar et al., 2003), and 96% of respondents that endorsed Category 5 had total PSWQ scores above the clinical cutoff (in contrast, only 27% of respondents that endorsed Category 3 had total PSWQ scores above the clinical cutoff). Virtually, identical results were produced in the validation sample: 80% of respondents who endorsed Category 4 on Item 15 and 93% of those who endorsed Category 5 on this item scored above the PSWQ total cutoff, whereas only 23% of those who endorsed Category 3 on Item 15 scored above the PSWQ total cutoff. In other words, responses of 4 or 5 on Item 15 make it substantially more likely that individuals are above the clinical cutoff score on the full measure.

Taken together, the IRT analysis of the PSWQ suggests that Item 15 may be an acceptable single-item gauge of problematic worry. We now turn to additional analyses to examine this possibility further. In these analyses, we compare Item 15 with the full scale (Meyer et al., 1990) and to the three-item PSWQ (Berle et al., 2011).

### Screening and Validation Analyses

Examination of the ROC curve for each PSWQ item in the validation sample revealed that Item 15 had the largest area under the curve ( $AUC = .83$ ,  $SE = .021$ , 95% confidence interval [CI: .79, .87]), which was very similar to the full-scale PSWQ ( $AUC = .84$ ,  $SE = .020$ , 95% CI [.80, .88]) and the three-item PSWQ ( $AUC = .84$ ,  $SE = .022$ , 95% CI [.80, .88]). As shown in Figure 2, ROC curves of the full scale, the three-item version, and of Item 15 were nearly identical.<sup>3</sup>

Screening results (sensitivity, specificity, PPP, and NPP) of the full scale, the three-item version, and Item 15 are presented in Table 2. Replicating previous work in college students (Behar et al., 2003), a score of 62 maximized specificity and sensitivity, and replicating previous analyses of the three-item scale (Kertz et al., 2014), a score of 11 on the three-item version maximized these characteristics. Most important, a score



**Figure 2.** Receiver operating characteristic (ROC) curves for the full PSWQ, three-item version, and Item 15.  
Note. PSWQ = Penn State Worry Questionnaire.

of 4 on Item 15 produced nearly identical specificity, PPP, and NPP, with only a slight decrease in sensitivity (.64, compared with .68 in the other versions). The cutoff of 4 on Item 15 replicates the IRT analyses in the calibration data set.

### Temporal Stability

Table 3 contains both the 1-month and 1-year retest coefficients for the PSWQ, its items, and the three-item brief form. The full-scale PSWQ (.77, .81), three-item PSWQ (.76, .80), and Item 15 (.72, .82), all demonstrated similar

**Table 2.** Screening Utility of the Full PSWQ, the Three-Item Version, and Item 15.

Measure	Cutoff score	Sensitivity	Specificity	PPP	NPP
Full PSWQ	60	.72	.85	.40	.96
Full PSWQ	61	.69	.86	.42	.95
<b>Full PSWQ</b>	<b>62</b>	<b>.68</b>	<b>.89</b>	<b>.47</b>	<b>.95</b>
Full PSWQ	63	.64	.90	.50	.95
Full PSWQ	64	.64	.92	.52	.95
Full PSWQ	65	.60	.93	.56	.94
PSWQ-3	8	.84	.67	.30	.97
PSWQ-3	9	.79	.77	.33	.96
PSWQ-3	10	.74	.86	.43	.96
<b>PSWQ-3</b>	<b>11</b>	<b>.68</b>	<b>.92</b>	<b>.54</b>	<b>.95</b>
PSWQ-3	12	.55	.95	.63	.94
PSWQ-3	13	.41	.98	.78	.92
Item 15	3	.81	.71	.28	.96
<b>Item 15</b>	<b>4</b>	<b>.64</b>	<b>.92</b>	<b>.55</b>	<b>.95</b>
Item 15	5	.30	.98	.72	.91

Note.  $N = 1,191$ . PSWQ = Penn State Worry Questionnaire; PPP = positive predictive power; NPP = negative predictive power; PSWQ-3 = three-item PSWQ. Suggested cutoff scores are bolded.

levels of stability at both time intervals. Compared with the other individual items, Item 15 was the most stable. Its cross-time correlations were 1 standard deviation above the average item retest coefficient at 1 month ( $M r = .54$ ), and nearly 1 standard deviation above the average at 1 year ( $M r = .60$ ). Thus, Item 15 is both more stable over time than all of the other individual items, and about as stable as the full-scale PSWQ and the three-item measure. The 1-year stability findings however should be interpreted cautiously given the relatively small sample size ( $N = 34$ ).

### Convergent Validity

Correlations between the PSWQ (individual items, the full scale, and the three-item scale) and the criterion measures are presented in Table 3. Compared with the other PSWQ items, Item 15 was the most strongly related to all four of the external measures. Notably, the correlations between Item 15 and the criterion measures were also at least as large as the correlations based on the entire PSWQ scale and the three-item scale.

We formally compared Item 15's correlations with the full scale's correlations using Steiger's (1980)  $Z$  test for dependent correlations. Correlations were not significantly different for the WDQ-SF and STAI ( $Z_s = 1.15$  and  $1.38$ , respectively,  $ps > .17$ ), suggesting that Item 15 relates to these external measures just as well as the full PSWQ scale. Moreover, the correlations with Item 15 were significantly *higher* than those with the full scale for the MASQ-AA ( $Z = 6.08$ ,  $p < .001$ ), and MASQ-AD ( $Z = 2.36$ ,  $p = .02$ ). In contrast, the convergent correlations based on the full PSWQ were significantly higher than

**Table 3.** Item-Total Correlations, Test-Retest Coefficients, Convergent Correlations, and Assessment Time Considerations of PSWQ Items.

PSWQ item	Item- total, <i>r</i>	I-Month retest, <i>r</i>	I-Year retest, <i>r</i>	Correlations with worry, anxiety, and depression (Pearson's <i>r</i> )					Assessment time
				GAD-7	WDQ-SF	STAI	MASQ- AA	MASQ- AD	Response time (s), <i>M</i> ( <i>SD</i> )
1. If I do not have enough time to do everything, I do not worry about it (R)	.49	.26	.25	.20	.23	.18	-.02	.08	8.69 (3.53)
2. My worries overwhelm me	.75	.43	.56	.50	.42	.51	.24	.34	4.27 (2.11)
3. I do not tend to worry about things (R)	.67	.44	.48	.34	.37	.36	.04	.19	4.59 (2.16)
4. Many situations make me worry <sup>a</sup>	.78	.64	.70	.49	.43	.50	.23	.31	3.66 (1.61)
5. I know I should not worry about things, but I just cannot help it	.81	.64	.59	.49	.41	.42	.17	.23	4.88 (2.35)
6. When I am under pressure, I worry a lot	.72	.50	.49	.40	.40	.39	.10	.21	3.86 (2.07)
7. I am always worrying about something	.81	.65	.61	.55	.43	.54	.23	.33	2.85 (1.42)
8. I find it easy to dismiss worrisome thoughts (R)	.67	.52	.62	.26	.31	.39	.09	.24	4.32 (1.81)
9. As soon as I finish one task, I start to worry about everything else I have to do	.70	.40	.55	.41	.37	.42	.18	.27	5.34 (2.44)
10. I never worry about anything (R)	.65	.49	.67	.28	.27	.19	.06	.09	3.19 (1.82)
11. When there is nothing more I can do about a concern, I do not worry about it anymore (R)	.55	.57	.39	.21	.19	.29	.09	.20	6.24 (2.42)
12. I have been a worrier all my life	.74	.66	.78	.41	.32	.35	.14	.19	2.97 (1.18)
13. I notice that I have been worrying about things	.77	.56	.60	.49	.42	.42	.14	.25	4.40 (1.80)
14. Once I start worrying, I cannot stop <sup>a</sup>	.76	.56	.72	.53	.36	.54	.27	.35	2.90 (1.44)
<b>15. I worry all the time<sup>a</sup></b>	<b>.82</b>	<b>.72</b>	<b>.82</b>	<b>.59</b>	<b>.46</b>	<b>.57</b>	<b>.28</b>	<b>.36</b>	<b>2.26 (1.23)</b>
16. I worry about projects until they are all done	.68	.48	.41	.32	.37	.28	.10	.16	3.43 (1.57)
Average	0.72	0.54	0.60	0.41	0.36	0.40	0.15	0.24	4.24
Standard deviation	0.18	0.16	0.23	0.15	0.09	0.14	0.09	0.09	1.57
Full PSWQ	1.00	0.77	0.81	0.59	0.49	0.55	0.20	0.33	67.86 (19.22)
Three-item PSWQ	0.90	0.76	0.80	0.60	0.47	0.59	0.24	0.32	8.82 (3.04)

Note. PSWQ = Penn State Worry Questionnaire. (R) indicates reverse-scored item. For item-total correlations,  $N = 9,040$ .  $N$ s ranged from 164 to 165 for 1-month retest coefficients and  $N = 34$  for 1-year retest coefficients. GAD-7 (Generalized Anxiety Disorder-7):  $N = 1,189$ -1,195; WDQ-SF (Worry Domains Questionnaire-Short Form):  $N = 400$ -405; STAI (State-Trait Anxiety Inventory-Trait Version):  $N = 1,137$ -1,144; MASQ (Mood and Anxiety Symptom Questionnaire; AA: Anxious Arousal, AD: Anhedonic Depression):  $N = 1,949$ -1,960; response time data from  $N = 154$ . Averages and standard deviations of correlations were computed after coefficients were first transformed using Fisher's  $r$ -to- $z$  transform and then those values were retransformed into  $r$  values. Item 15 is bolded as it had the highest item-total correlations, retest coefficients, convergent correlations, and shortest response times of the individual items.

<sup>a</sup>Indicates the item is included on the three-item PSWQ (Berle et al., 2011).

the correlations between the other individual items and the criterion measures (all  $Z$ s  $> 2.99$ ,  $ps < .01$ ).

We similarly compared convergent correlations between Item 15 and the three-item PSWQ. Correlations with the WDQ-SF were not significantly different ( $Z = 0.54$ ,  $p = .54$ ), but the correlation with the STAI was significantly higher for the three-item scale ( $Z = 2.62$ ,  $p = .009$ ). On the other hand, correlations with the MASQ-AA and MASQ-AD were significantly higher with Item 15 than with the three-item scale ( $Z = 2.65$ ,  $p = .008$  and  $Z = 2.93$ ,  $p = .003$ , respectively). In sum, the correlations between Item 15 and the anxiety, worry, and depression measures were similar in magnitude or slightly larger than they were for the full PSWQ, and three-item PSWQ (with the exception of the STAI; though the absolute difference in magnitude between Item 15 and the three-item form was only .02).

### Response Time Analysis

The full PSWQ took participants approximately 1 minute to complete ( $M = 67$  seconds; Table 3), and the three-item scale took approximately 9 seconds to complete ( $M = 8.82$  seconds). Item 15 had the shortest RT of all PSWQ items, with an average RT of just over 2 seconds. Paired-samples  $t$  tests confirmed that RTs were significantly shorter for Item 15 compared with all other items (all  $ps < .0001$ , Cohen's  $d$  ranged from 0.44 to 2.43). This is likely because Item 15 has the fewest number of syllables (6) compared with the other items of the PSWQ ( $M$  number of syllables = 13.53,  $SD = 5.24$ ). Indeed, the number of syllables in each item was correlated with RTs ( $r = .78$ ,  $p = .0003$ ).<sup>4</sup>

### Discussion

The current study demonstrates via various methods, and across two large samples, that Item 15 of the PSWQ

“I worry all the time”) can likely serve as a suitable stand-alone item for the rapid assessment of problematic worry in brief screening contexts where the full PSWQ cannot be administered. Compared with the other PSWQ items, Item 15 measured worry most precisely, was most representative of the full PSWQ, was most stable across 1-month and 1-year intervals, was most predictive of worry, trait anxiety, anxious arousal, and anhedonic depression, and was the fastest item to complete (with an average of just over 2 seconds). Item 15 performed well across these domains in the absolute sense as well, suggesting that, for example, not only is Item 15 more reliable than the other items but it is also likely to be adequately reliable for most screening purposes.

Notably, previous psychometric analyses of the PSWQ have also found Item 15 to be especially strong (Brown, 2003; Olatunji, Schottenbauer, Rodriguez, Glass, & Arnkoff, 2007; van der Heiden, Muris, Bos, & van der Molen, 2010; van Rijsoort, Emmelkamp, & Vervaeke, 1999). Indeed, the shortest version of the PSWQ to date (three items) also includes Item 15 (Berle et al., 2011). It is worth noting here again though that the IRT analyses revealed that Item 15 was substantially more informative than the other two items of the three-item form (Items 4 and 14; see Figure 1). More significantly, Item 15 also performed at least as well as the full three-item form in many other respects, such as retest reliability, ROC, sensitivity and specificity properties, and criterion correlations. Item 15 was also 75% faster to complete than the three-item form (see Table 3). These results suggest that Item 15 carries most of the potency of the three-item version, and that the other two items may be somewhat superfluous. In other words, Item 15 generally measures problematic worry as well as the three-item scale, and thus is objectively more efficient for assessing problematic worry. Altogether then, if one were interested in rapid screening, Item 15 is likely sufficient without the other two accompanying items of the three-item scale.

In addition to identifying and supporting a single item measure for problematic worry, this study also generally extends previous work on abbreviated forms of the PSWQ in notable ways. The large unselected samples employed here allowed us to examine PSWQ items across the entire spectrum of problematic worry. Furthermore, the IRT model allowed us to critically evaluate the functioning of individual items along different levels of trait worry in a way that the CTT approach of previous studies could not. This was important because a single-item screening measure should be able to measure worry with adequate precision at the higher end of the worry spectrum so as to identify those at potential risk.

Having a psychometrically adequate single-item assessor of problematic worry is useful for several reasons. First, it would allow practitioners who do not routinely screen for

problematic worry to do so with minimal burden. Second, a single item allows for simultaneous assessment and scoring (i.e., no math required) to immediately begin clinical decision making regarding problematic worry. Third, the single item is less burdensome on respondents as it eliminates the reverse wording and redundancy of the full scale while retaining the excellent psychometric properties of the full scale (retest reliability, convergent validity). The RT data indicate that using only Item 15 could cut the total assessment time of problematic worry by 97%. Indeed, using a single item is in the spirit of current large-scale efforts to develop standardized “toolkits” to rapidly assess symptoms across diverse settings (e.g., PROMIS and PhenX; Fries, Bruce, & Cella, 2005; Hamilton et al., 2011).

The primary limitation of the current study was the reliance on a predominantly undergraduate population. Yet undergraduates are not immune to psychopathology (Hunt & Eisenberg, 2010), and the range of scores in the sample suggests it is likely that many individuals would have met diagnostic criteria for anxiety disorders. Without a clinical sample, we were unable to provide definitive cut-scores for this item, although the IRT and screening analyses suggest that a score  $\geq 4$  makes it much more probable that participants would score above the clinical cutoff value on the full PSWQ. Further research is necessary to validate this proposed single-item cutoff. However, our use of a nonclinical sample was purposeful in that we wanted to assess this single item of worry among a large sample of individuals who were *not* already in care, as a screening item would likely be used by the broader population, not just those already in clinical care. That is, we were interested in screening for problematic worry in the general population, which makes clinical samples somewhat less appropriate. Additionally, our study cannot speak to the utility of the screening item’s sensitivity to change with treatment; we cannot yet recommend clinicians treating problematic worry to track changes across time with this item. Finally, the majority of data came from anonymous online surveys and further research is needed to verify findings with interview and paper-and-pencil methods.

In conclusion, multiple strains of evidence suggest that Item 15 may be used as a rapid screening tool for problematic worry in contexts where the full PSWQ cannot be administered. We believe it would be beneficial for other research teams to critically evaluate their scales to home in on single items that provide adequate coverage of a construct. A brief “core battery” of nonredundant and rigorously studied items of mental health screens would be highly useful and efficient in a diverse range of clinical settings. The hope is that such a screening instrument would allow for rapid and reliable detection of problematic mental health concerns that can facilitate conversation between patients and providers to lead to faster detection and intervention.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Notes

1. A literature search (www.scopus.com) on February 26, 2016, found that 2,255 documents had referenced the PSWQ, whereas 1,462 referenced the Generalized Anxiety Disorder-7 (GAD-7; Spitzer et al., 2006), and 258 referenced the Worry Domains Questionnaire (Stober & Joormann, 2001).
2. Data analyzed here only considered participants who responded to items in a reasonable time. Participants who provided outlying RTs ( $Z$  score  $\geq 3$ ) were removed from the full sample (full sample  $N = 197$ ) prior to analysis. However, results were virtually identical when we considered the entire sample.
3. AUC for each item is presented in Table S2 of the supplemental material.
4. Like nearly all studies using the PSWQ, items were presented in chronological order and were not randomized. One potential possibility is that participants were faster to respond to Item 15 than to other items simply because it was one of the last items in the "PSWQ response set." This is an unlikely explanation, however, because the correlation between item number and RT was modest ( $r = -.52$ ,  $p = .04$ ) and nonsignificant ( $r = -.34$ ,  $p = .22$ ) after we removed Item 1 from the analysis to reduce the impact of the restart effect (prolonged RTs on the first item of a response set; Altmann & Gray, 2008).

## References

- Altmann, E. M., & Gray, W. D. (2008). An integrated model of cognitive control in task switching. *Psychological Review*, *115*, 602-639.
- Baxter, A. J., Scott, K. M., & Whiteford, H. A. (2013). Global prevalence of anxiety disorders: A systematic review and meta-regression. *Psychological Medicine*, *43*, 897-910.
- Behar, E., Alcaïne, O., Zuellig, A. R., & Borkovec, T. D. (2003). Screening for generalized anxiety disorder using the Penn State Worry Questionnaire: A receiver operating characteristic analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, *34*, 25-43.
- Berle, D., Starcevic, V., Moses, K., Hannan, A., Milicevic, D., & Sammut, P. (2011). Preliminary validation of an ultra-brief version of the Penn State Worry Questionnaire. *Clinical Psychology & Psychotherapy*, *18*, 339-346.
- Borsboom, D., Cramer, A. O., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135-170). Charlotte, NC: Information Age.
- Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors of method effects? *Behaviour Research and Therapy*, *41*, 1411-1426.
- Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*, 220-250.
- Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. *Quality Life Research*, *23*, 2809-2818.
- de Ayala, R. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fresco, D. M., Heimberg, R. G., Mennin, D. S., & Turk, C. L. (2002). Confirmatory factor analysis of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *40*, 313-323.
- Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient reported outcomes. *Clinical and Experimental Rheumatology*, *23*(Suppl. 39), S53-S57.
- Gladstone, G. L., Parker, G. B., Mitchell, P. B., Malhi, G. S., Wilhelm, K. A., & Austin, M. P. (2005). A Brief Measure of Worry Severity (BMWS): Personality and clinical correlates of severe worriers. *Journal of Anxiety Disorders*, *19*, 877-892.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Issues and applications*. Boston, MA: Kluwer Nijhoff.
- Hamilton, C. M., Strader, L. C., Pratt, J. G., Maiese, D., Hendershot, T., Kwok, R. K., . . . Haines, J. (2011). The PhenX toolkit: Get the most from your measures. *American Journal of Epidemiology*, *174*, 253-260.
- Hoffman, D. L., Dukes, E. M., & Wittchen, H. (2008). Human and economic burden of generalized anxiety disorder. *Depression and Anxiety*, *25*, 72-90.
- Hopko, D. R., Stanelly, M. A., Reas, D. L., Wetherell, J. L., Beck, J. G., Novey, D. M., & Averill, P. M. (2003). Assessing worry in older adults: Confirmatory factor analysis of the Penn State Worry Questionnaire and psychometric properties of an abbreviated model. *Psychological Assessment*, *15*, 173-183.
- Hunt, J., & Eisenberg, D. (2010). Mental health problems and help-seeking behavior among college students. *Journal of Adolescent Health*, *46*, 3-10.
- Kanuri, N., Taylor, C. B., Cohen, J. M., & Newman, M. G. (2015). Classification models for subthreshold generalized anxiety disorder in a college population: Implications for prevention. *Journal of Anxiety Disorders*, *34*, 43-52.
- Kertz, S. J., Lee, J., Bjogvinsson, T. (2014). Psychometric properties of abbreviated and ultra-brief versions of the Penn State Worry Questionnaire. *Psychological Assessment*, *26*, 1146-1154.
- Klinkman, M. S. (1997). Competing demands in psychosocial care: A model for the identification and treatment of depressive disorders in primary care. *General Hospital Psychiatry*, *19*, 98-111.
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*, 606-613.
- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

- McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research*, *4*, 293-307.
- Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *28*, 487-495.
- Mitchell, A. J., Kaar, S., Coggan, C., & Herdman, J. (2008). Acceptability of common screening methods used to detect distress and related mood disorders: Preferences of cancer specialists and non-specialists. *Psycho-Oncology*, *17*, 226-236.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Newman, M. G., Llera, S. J., Erickson, T. M., Przeworski, A., & Castonguay, L. G. (2013). Worry and generalized anxiety disorder: A review and theoretical synthesis of evidence on nature, etiology, mechanisms, and treatment. *Annual Review of Clinical Psychology*, *9*, 275-297.
- Olatunji, B. O., Schottenbauer, M. A., Rodriguez, B. F., Glass, C. R., & Arnkoff, D. B. (2007). The structure of worry: Relations between positive/negative personality characteristics and the Penn State Worry Questionnaire. *Journal of Anxiety Disorders*, *21*, 540-553.
- Ozer, E. M., Zahnd, E. G., Adams, S. H., Husting, S. R., Wibbelsman, C. J., Norman, K. P., & Smiga, S. M. (2009). Are adolescents being screened for emotional distress in primary care? *Journal of Adolescent Health*, *44*, 520-527.
- Prince, M., Patel, V., Saxena, S., Maj, M., Maserko, J., Phillips, M. R., & Rahman, A. (2007). No health without mental health. *Lancet*, *370*, 859-876.
- Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*, 151-161.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*, 100-114.
- Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*, 367-373.
- Smith, P. C., Schmidt, S. M., Allensworth-Davies, D., & Saitz, R. (2010). A single-question screening test for drug use in primary care. *Archives of Internal Medicine*, *170*, 1155-1160.
- Spielberger, C. D., Gorssuch, R. L., Lushene, P. R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Spitzer, R. L., Kroenker, K., Williams, J. B., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*, 1092-1097.
- Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.
- Stober, J. (1998). Reliability and validity of two widely-used worry questionnaires: Self-report and self-peer convergence. *Personality and Individual Differences*, *24*, 887-890.
- Stober, J., & Joormann, J. (2001). A short form of the Worry Domains Questionnaire: Construction and factorial validation. *Personality and Individual Differences*, *31*, 119-126.
- Yao, B., Sripada, R. K., Klumpp, H., Abelson, J. L., Muzik, M., Zhao, Z., . . . Warren, R. (2016). Penn State Worry Questionnaire - 10: A new tool for measurement-based care. *Psychiatry Research*, *239*, 62-67.
- Tedeschi, G. J., Cummins, S. E., Anderson, C. M., Anthenelli, R. M., Zhuang, Y-L., & Zhu, S-H. (2016). Smokers with self-reported mental health conditions: A case for screening in the context of tobacco cessation services. *PLoS ONE*, *11*(7), e0159127. doi:10.1371/journal.pone.0159127
- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum.
- Topper, M., Emmelkamp, P. M. G., Watkins, E., & Ehring, T. (2014). Development and assessment of brief versions of the Penn State Worry Questionnaire and the Ruminative Response Scale. *British Journal of Clinical Psychology*, *53*, 402-421.
- van der Heiden, C., Muris, P., Bos, A. E. R., & van der Molen, H. T. (2010). Factor structure of the Dutch version of the Penn State Worry Questionnaire. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*, 304-309.
- van Rijsoort, S., Emmelkamp, P., & Vervaeke, G. (1999). The Penn State Worry Questionnaire and the Worry Domains Questionnaire: Structure, reliability and validity. *Clinical Psychology & Psychotherapy*, *6*, 297-307.
- Watson, D., & Clark, L. A. (1991). *The Mood and Anxiety Symptom Questionnaire*. Unpublished manuscript, University of Iowa, Iowa City.
- World Health Organization & World Organization of Family Doctors. (2008). *Integrating mental health into primary care: A global perspective*. Retrieved from [http://www.who.int/mental\\_health/resources/mentalhealth\\_PHC\\_2008.pdf](http://www.who.int/mental_health/resources/mentalhealth_PHC_2008.pdf)
- Zenlea, I. S., Milliren, C. E., Mednick, L., & Rhodes, E. T. (2014). Depression screening in adolescents in the United States: A national study of ambulatory office-based practice. *Academic Pediatrics*, *14*, 186-191.
- Zimmerman, M. (2016). *Using scales to monitor symptoms and treat depression (measurement based care)*. Retrieved from <http://www.uptodate.com/contents/using-scales-to-monitor-symptoms-and-treat-depression-measurement-based-care>
- Zimmerman, M., & McGlinchey, J. B. (2008). Why don't psychiatrists use scales to measure outcome when treating depressed patients? *Journal of Clinical Psychiatry*, *69*, 1916-1919.
- Zimmerman, M. A., & Wienckowski, L. A. (1991). Revisiting health and mental health linkages: A policy whose time has come . . . again. *Journal of Public Health Policy*, *12*, 510-524.