Received: 26 October 2016 | Revised: 8 February 2017 | Accepted: 10 February 2017

DOI 10.1002/mpr.1561

WILEY

ORIGINAL ARTICLE

The development and validation of static and adaptive screeners to measure the severity of panic disorder, social anxiety disorder, and obsessive compulsive disorder

Matthew Sunderland[1] | Philip J. Batterham[2] | Alison L. Calear[2] | Natacha Carragher[3]

[1] NHMRC Centre for Research Excellence in Mental Health and Substance Use, National Drug and Alcohol Research Centre, UNSW Australia, Sydney, Australia

[2] Centre for Mental Health Research, The Australian National University, Canberra, Australia

[3] Office of Medical Education, Faculty of Medicine, UNSW Australia, Sydney, Australia

Correspondence
Matthew Sunderland, National Drug and Alcohol Research Centre, UNSW Australia, NSW 2031, Australia.
Email: matthews@unsw.edu.au

Funding information
National Health and Medical Research Council, Grant/Award Number: Fellowships 1052327, 1083311, 1013199, 1043952

Abstract

A series of static and adaptive screeners for panic disorder, social anxiety disorder (SAD), and obsessive compulsive disorder (OCD) were developed and compared using data-driven methods to facilitate the measurement of each disorder in community samples. Data comprised 3175 respondents for the development sample and 3755 respondents for the validation sample, recruited independently using Facebook advertising. Item Response Theory (IRT) was utilized to develop static continuous screeners and to simulate computerized adaptive algorithms. The screeners consisted of a small subset of items from each bank (79% reduction in items for panic disorder, 85% reduction in items for SAD, and 84% reduction in items for OCD) that provided similar scores (r = 0.88–0.96). Both static and adaptive screeners were valid with respect to existing scales that purportedly measure similar constructs (r > 0.70 for panic disorder, r > 0.76 for SAD, and r > 0.68 for OCD). The adaptive scales were able to maintain a higher level of precision in comparison to the static scales and evidenced slightly higher concordance with scores generated by the full item banks. The screeners for panic disorder, SAD, and OCD could be used as a flexible approach to measure and monitor the severity of psychopathology in tailored treatment protocols.

KEYWORDS

adaptive testing, IRT, obsessive compulsive disorder, panic disorder, screening, social anxiety

1 | INTRODUCTION

Panic disorder, social anxiety disorder (SAD), and obsessive compulsive disorder (OCD) are characterized by a profound sense of fear, anxiety, and distress related to either the possibility of re-experiencing panic attacks, engaging in social or performance situation, or impairment associated with not being able to engage in unwanted obsessions and compulsions. This fear and anxiety is often highly debilitating and permeates multiple areas of psychosocial functioning and quality of life (Mendlowicz & Stein, 2000). These disorders affect a sizable proportion of the population with lifetime prevalence as high as 10% for SAD, followed by panic disorder (ranging between 2 and 5%) and then OCD (2–3%) (Kessler, Ruscio, Shear, & Wittchen, 2010). Importantly, SAD and OCD have an earlier age of onset in comparison to other mental disorders (e.g. depression, psychosis, and substance use disorders) whereas panic disorder has a similar age of onset to mood disorders and all three are comparatively persistent with high 12-month to lifetime prevalence ratios (Kessler, Petukhova, Sampson, Zaslavsky, & Wittchen, 2012; Kessler et al., 2007).

Despite the heavy burden associated with panic disorder, SAD, and OCD, the time to treatment is often lengthy with the majority of people not seeking treatment at all (Burgess et al., 2009; Thompson, Issakidis, & Hunt, 2008). For those who engage with treatment there is often low adherence rates and high dropout, particularly for open access e-mental health therapies, which can result in poor treatment outcomes (Christensen, Griffiths, & Farrer, 2009; Hilvert-Bruce, Rossouw, Wong, Sunderland, & Andrews, 2012). To address these issues, there have been calls for standardized screening initiatives that identify individuals in the population with heightened disorder severity and establish links to appropriate treatments (Siu et al., 2016). Once in treatment there is a need for continual monitoring of patient outcomes and tailoring the treatment protocol to better suit each individual's needs (Fortney et al., 2017). However, these goals require precise yet highly efficient instruments so that screening and continual

patient outcome monitoring can be widely accepted and successful (Rose et al., 2012).

Researchers have utilized several advances in item banking and psychometric models to improve the validity, precision, and efficiency associated with mental health screening (Gibbons, Weiss, Frank, & Kupfer, 2016). Most notably, methods from Item Response Theory (IRT) have made it possible to maximize the level of precision exhibited by the full item bank whilst only administering a fraction of the total bank to potential respondents. IRT provides a statistical framework that calibrates the observed responses from different items to an examinee's location on a single latent construct of interest, e.g. disorder severity (Embretson & Reise, 2000). Having prior information regarding the performance of each test item at varying levels of disorder severity allows developers to select a reduced set of items (a short screener) that optimally discriminate between respondents at varying degrees of severity on the trait continuum. Importantly, different types of short screeners can be defined depending on the goal of the instrument and the technique used to select and administer items. Specifically, screeners can be administered in a static manner (i.e. respondents are administered the same fixed subset of optimal items) or an adaptive manner (i.e. respondents are administered a smaller subset of items tailored to their previous responses). The primary goal of adaptive testing is to maximize the level of precision associated with each individual's severity score while minimizing the number of items that are administered. The precision can be controlled at a more consistent level across the range of severity in comparison to brief static screeners. That being said, both static and adaptive approaches result in different degrees of complexity, efficiency, and precision, which require empirical testing and comparisons in order to justify their use (Cella, Gershon, Lai, & Choi, 2007).

Previously, the National Institutes of Health (NIH)-funded PROMIS (patient reported outcomes measurement information system) initiative developed and evaluated item banks and screeners based on IRT and adaptive testing to measure the severity of mental health constructs, including depression, anxiety, and anger (Pilkonis et al., 2011). Their results indicated that the item banks and screeners provided relatively more test information (e.g. greater precision) than previously well-established scales for depression, anxiety and anger. Furthermore, the item banks displayed sufficient convergent and divergent validity against the legacy scales while the screeners substantially reduced the number of items administered while maintaining very high correlations with the item bank (r = 0.96). Choi, Reise, Pilkonis, Hays, and Cella (2010) further compared static and adaptive screeners using the PROMIS depression item bank and found that adaptive tests performed marginally better than static tests when estimating scores using as few items as possible. Similarly, Gibbons et al. (2012, 2014) examined the efficiency of two adaptive tests developed using item banks (developed independently from PROMIS) for depression and generalized anxiety and found substantial savings in average items administered with little reduction in precision and good convergent validity with legacy scales and DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition) diagnoses. However, no comparisons were made to equivalent length static versions.

Despite the favourable results of data-driven approaches to screener development using item banks, much of this research has focused on a relatively small number of mental health constructs, namely depression and generalized anxiety. Batterham, Sunderland, Carragher, and Calear (2016b), recently developed several item banks that specifically measure panic disorder, SAD, and OCD. However, data-driven screeners based on these item banks have yet to be developed or validated in the population. Similarly, studies have yet to examine differences in the performance of static versus adaptive screeners when estimating underlying severity scores for these disorders. The efficiency associated with adaptive algorithms can vary across item banks depending on the size, nature, and quality of the items. Comparisons are required to determine whether the added complexity associated with adaptive testing is warranted over and above static scales.

The current study aimed to develop and validate data-driven (e.g. based on IRT) static and adaptive screeners for panic disorder, SAD, and OCD using three newly constructed item banks in a community sample of Australian adults. The results of the newly developed screeners were validated against legacy screening scales for each disorder as well as DSM-5 diagnoses. Similarly, the results were replicated using a sample of adults that were recruited separately from the development sample. Finally, comparisons were drawn between the performance of the static and adaptive screeners to determine if the added complexity associated with adaptively administered items was warranted in the context of assessing panic disorder, SAD, and OCD.

2 | METHODS

2.1 | Samples

2.1.1 | Development sample

Respondents were recruited from the online social media website Facebook using a series of advertisements targeted to all Australian adults aged 18 years or older during August–December 2014. A total of 39,945 users clicked on the advertisement with 10,082 adults consenting to participate in the survey and 5011 (49.7%) completing the survey. To reduce the questionnaire length, consenting participants were given the option of completing a brief form of the survey, which comprised one of three versions covering a different combination of mental disorders, or one of three full forms that covered all disorders of interest but were presented in a different order. Only the full version collected information that encompassed all item banks and DSM-5 criteria, hence the current study included participants who opted to complete the full version of the survey, comprising 3175 respondents (63.4% of completers). The survey was approved by the Australian National University Human Research Ethics Committee (protocol #2013/509).

2.1.2 | Validation sample

Respondents for the validation sample were recruited using a temporally distinct recruitment process that also used Facebook advertising targeting Australians aged 18 years or older during January–February 2016. A total of 7174 users clicked the advertisements with 5379 (75%) consenting to complete the survey. Of those consenting respondents, a total of 3577 (66%) completed the survey and were included

in the analysis. The survey was approved by the Australian National University Human Research Ethics Committee (protocol #2015/717).

2.1.3 | Weighting scheme

The two samples were not representative of the general population with comparisons indicating that our samples had higher levels of psychopathology and over-representation of females. As such, a weighting scheme was applied to the data when reporting all results in the current study. This scheme was designed to make the sample more representative of the general population in terms of age, gender and psychopathology distributions, to account for these potential imbalances (Batterham et al., 2016b). Age and gender distributions were obtained from the Australian Bureau of Statistics (http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0). Representative data on mental disorders was obtained from the 2007 Australian National Survey of Mental Health and Well-being (Slade, Johnston, Oakley Browne, Andrews, & Whiteford, 2009). The sample characteristics using unweighted and weighted data are provided in Table 1.

2.2 | Measures

2.2.1 | Item banks

The item banks for panic disorder, SAD, and OCD were constructed as part of a larger project with the aim of developing comprehensive and efficient assessment tools for multiple mental and substance use disorders. A multi-stage process was used to select items, which included: systematic literature searches, feedback by consumers and expert researchers/clinicians, and reduction based on relevance and item content overlap (more details provided in Batterham et al., 2015). Initial testing of the item banks was conducted to determine the presence of a unidimensional structure, local independence, and measurement invariance across gender, age, and education (see Batterham et al., 2016b). The items were consistent with the item banks developed by the PROMIS initiative in terms of response scale, past tense, first person, and past 30-day time frame. The final item banks included 19 items for panic disorder, 44 items for OCD, and 26 items for SAD; items were rated on a 5-point categorical response scale from "0 = never" to "4 = always". The individual items from each bank are provided in the Supporting Information along with the unweighted frequencies associated with each response category from the development sample. Similarly, the histograms and descriptive statistics for the total raw sum scores of each item bank are provide in the Supporting Information and demonstrate the sample captures a range of psychopathology from no symptoms to high rates of symptomatology (albeit the distributions were skewed with higher rates in the low severity range as indicative of community samples as opposed to clinical samples).

2.2.2 | Legacy scales

The Panic Disorder Screener (PADIS) is a four-item questionnaire designed to fill the need for a brief severity rating scale of panic disorder for use in research and clinical settings. The PADIS was recently found to be more accurate than the Patient Health

TABLE 1 Socio-demographic characteristics of the weighted and unweighted development and validation data

		Development (*n* = 3175)		Validation (*n* = 3577)	
		%	Weighted %	%	Weighted %
Age	18–25	11.7	13.9	19.3	14.0
	26–35	8.7	18.4	12.2	18.6
	36–45	15.9	17.6	15.0	17.4
	46–55	26.1	17.3	20.6	17.3
	56–65	26.5	16.5	22.8	16.0
	66+	11.0	16.4	10.2	16.8
Gender	Male	20.4	49.0	19.3	48.6
	Female	79.6	51.0	80.7	51.4
Education	Primary School	0.2	0.3	0.1	0.1
	Some secondary school	2.9	2.6	2.3	2.2
	Year 10 or equivalent	9.0	7.8	4.6	5.4
	Year 12 or equivalent	13.9	16.0	14.5	12.2
	Certificate level I–IV	15.2	12.6	12.1	12.6
	Diploma/Associate diploma	15.1	13.8	10.1	11.2
	Bachelor degree	19.8	22.2	25.0	24.5
	Graduate diploma/certificate	12.2	11.3	14.1	14.2
	Master's Degree	8.7	10.2	12.9	13.9
	Doctoral degree	2.6	3.1	4.1	3.8
	Prefer not to say	0.4	0.5	0.2	0.1
Language	English only	93.3	91.0	92.4	92.5
	English and another language	6.3	8.4	7.2	6.8
	Another language only	0.4	0.6	0.4	0.8
Location	Metropolitan area (capital city)	44.8	48.9	52.3	52.6
	Regional area (other city / town)	40.3	37.9	36.0	36.4
	Rural or remote area	15.0	13.2	11.7	11.0
DSM-5 Panic disorder		5.9	2.3	5.0	1.7
DSM-5 Social anxiety disorder		17.4	6.5	14.3	6.0
DSM-5 OCD		14.1	8.3	11.9	7.6

Questionnaire – panic scale (Spitzer, Kroenke, & Williams, 1999) in detecting panic disorder in a large community sample (Batterham, MacKinnon, & Christensen, 2016a).

The Social Phobia Screener (SOPHS) contains five items to assess the degree of fear, embarrassment, avoidance, and interference caused by social or performance situations in the past 30 days. The SOPHS has strong psychometric properties and comparable accuracy to the 17-item Social Phobia Inventory and the three-item Mini-SPIN (Batterham et al., 2016a).

The Short Obsessions and Compulsions Screener (SOCS) is a seven-item scale. The first five questions enquire about checking, touching, cleanliness/washing, repeating and exactness and the final two items are designed to gauge the associated impairment and difficulty to resist obsessions and compulsions. The SOCS has comparable or better discriminant characteristics in comparison to other self-report tools for OCD and high sensitivity and negative predictive value in a sample of young adults (Uher, Heyman, Mortimore, Frampton, & Goodman, 2007). The histograms and descriptive statistics for the total raw sum scores of the three legacy scales are provided in the Supporting Information alongside the item banks. Again, the full range of severity was covered by the sample but with relatively higher rates observed at the lower severity range.

2.3 | Statistical analyses

2.3.1 | Development of static screeners

Brief static screeners were developed with the explicit goal of replicating the functionality across the spectrum of severity that is measured by the full item banks. To achieve this goal, items were selected based on IRT discrimination and difficulty parameters as well as item information curves when measuring a single unidimensional construct representing either panic disorder, SAD, or OCD severity (Embretson & Reise, 2000). Items with particularly high discrimination were preferred (given that these items maximize information and increase the level of precision associated with the scale), in comparison to the remaining items, and item characteristic curves were inspected to ensure each response option provided sufficient information relative to the other response options. However, some items could be included with lower discrimination parameters in comparison to unselected items if they captured more information at lower or higher levels of severity. This was to ensure that the items contained in the brief screeners captured sufficient information along the same range of the continua that was captured by the full item bank.

The test information curves of brief static screeners, full item banks, and legacy scales were compared visually by placing all items (including those from the legacy scales) on the same metric as the full item banks. This was achieved by estimating a single factor confirmatory factor analysis (CFA) with the parameters for the full item banks fixed at their previously estimated values whilst the legacy items were freely estimated. Scores from the static screeners were compared to scores generated from the full item banks using Pearson's correlation, bias (signed mean difference), and root-mean squared deviation (RMSD). The scores were estimated using the expected a posteriori (EAP) method (Embretson & Reise, 2000). The IRT analysis and CFA were conducted using Mplus version 7 (Muthen & Muthen, 2015). EAP scores were estimated and compared using the mirtCAT package for R, version 0.5 (Chalmers, 2016).

2.3.2 | Development of adaptive screeners

Computerized adaptive tests (CATs) were developed using the IRT parameters estimated in the full development sample. CATs seek to maximize concordance with the full item banks whilst minimizing the number of items administered. The level of precision can be maintained for the adaptive test by setting termination criteria based on the minimum standard error associated with the theta scores that is pre-determined as acceptable. The pointwise Kullback–Liebler information divergence method was used to select the initial item (at a mean trait level of zero) and to adaptively select subsequent items from the full bank. This method has been shown to reduce bias and mean square error in the early stages of the test when few items have been administered (Chang & Ying, 1996). Preliminary and final theta (latent trait) scores were estimated using the EAP method (Embretson & Reise, 2000). Several termination rules were examined using multiple simulations to determine an optimal balance between efficiency and precision. The first termination rule required the CAT to terminate if the standard error associated with the theta score decreased below 0.3. This rule was then relaxed to examine the impact of increasing the standard error (i.e. less precision) associated with the theta score to <0.35 and <0.4, respectively.

There could be some instances where an adaptive test based on standard error termination rules may continue to needlessly administer items until the maximum number of items from the bank is reached. This is due to the fact that the total item bank may not contain a sufficient number of items to meet the required standard error cut-point for some individuals with theta scores that are not indexed well by the bank (e.g. extremely low severity scores). As such, an additional termination rule was applied that would terminate the CAT if the change in theta scores from one item to the next dropped below 0.05. This ensures that the CAT would terminate once there were no remaining items in the bank that would alter an individual's theta score to a large extent. Each of the three standard error termination rules (< 0.3, < 0.35, and <0.4) were then re-analysed in tandem with the additional rule where the CAT would terminate if the difference between theta estimates from one item to the next dropped below 0.05. A total of six simulations were conducted using the mirtCAT package for R, version 0.5 (Chalmers, 2016).

Scores from the CAT simulations were compared to scores generated from the full item banks using Pearson's correlation, bias (signed mean difference), and RMSD. Efficiency of each CAT was determined using the mean number of items administered. Additional Pearson's correlations were estimated between scores from an optimal CAT, brief static screener, full item bank, and legacy scale for each disorder.

2.3.3 | Validation of brief screeners

Scores were generated from responses obtained in the validation sample using the static screeners and the optimal CATs identified in the development sample. The performance of each screener was examined

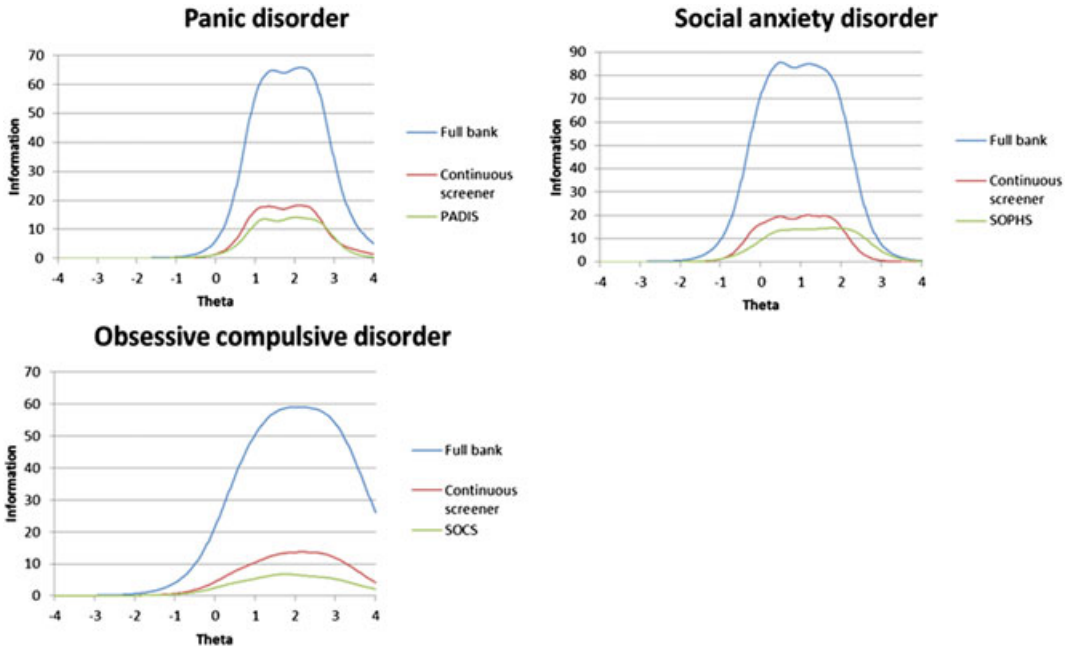

FIGURE 1 Information curves for the full item banks, static continuous screeners, and legacy scales in development sample (n = 3175)

TABLE 2 Comparisons between computerized adaptive test (CAT) simulations and static screeners with the full item banks in the development sample (n = 3175)

	$M_{(items)}$	$M_{(theta)}$	$M_{(SE)}$	Correlation	Bias	RMSD
Panic disorder						
Full bank	19	0.00	0.43	1.00	–	–
CAT 1: SE < 0.30	14	0.00	0.47	0.99	<0.01	0.14
CAT 2: SE < 0.35	13	0.00	0.48	0.98	<0.01	0.16
CAT 3: SE < 0.40	12	0.00	0.49	0.98	<0.01	0.18
CAT 4: SE < 0.30, Δthetas < 0.05	4	0.00	0.54	0.93	<0.01	0.32
CAT 5: SE < 0.35, Δthetas < 0.05	4	0.00	0.55	0.93	<0.01	0.33
CAT 6: SE < 0.40, Δthetas < 0.05	4	0.00	0.56	0.92	<0.01	0.33
Static brief screener	4	0.00	0.59	0.88	0.17	0.41
Social anxiety disorder (SAD)						
Full bank	26	0.00	0.19	1.00	–	–
CAT 1: SE < 0.30	8	0.00	0.31	0.97	<0.01	0.22
CAT 2: SE < 0.35	7	0.00	0.33	0.97	<0.01	0.25
CAT 3: SE < 0.40	6	0.00	0.35	0.96	<0.01	0.27
CAT 4: SE < 0.30, Δthetas < 0.05	4	0.00	0.33	0.96	<0.01	0.27
CAT 5: SE < 0.35, Δthetas < 0.05	4	0.00	0.35	0.96	<0.01	0.28
CAT 6: SE < 0.40, Δthetas < 0.05	3	0.00	0.36	0.95	<0.01	0.30
Static brief screener	4	0.00	0.36	0.94	<0.01	0.34
Obsessive compulsive disorder (OCD)						
Full bank	44	0.00	0.27	1.00	–	–
CAT 1: SE < 0.30	22	0.00	0.34	0.98	<0.01	0.21
CAT 2: SE < 0.35	18	0.00	0.37	0.96	<0.01	0.25
CAT 3: SE < 0.40	15	−0.02	0.41	0.95	0.02	0.28
CAT 4: SE < 0.30, Δthetas < 0.05	8	−0.01	0.41	0.93	0.01	0.34
CAT 5: SE < 0.35, Δthetas < 0.05	7	−0.01	0.42	0.93	0.01	0.35
CAT 6: SE < 0.40, Δthetas < 0.05	6	−0.01	0.44	0.92	0.01	0.37
Static brief screener	7	0.01	0.49	0.90	<0.01	0.41

Note: SE, standard error; Δthetas, difference in provisional theta scores from one item to the next; RMSD, root mean square deviation.

by generating the mean number of items administered, mean theta score, mean standard error, and Pearson's correlation, bias, and RMSD in comparison to scores generated using the full item bank.

3 | RESULTS

Static screeners consisted of four items for panic disorder (79% reduction), four items for SAD (85% reduction), and seven items for OCD (84% reduction). The specific items selected from the item banks are detailed further in the Supporting Information along raw score to IRT and *T*-score conversion tables to facilitate scoring on a metric weighted to the Australian population. The test information curves for the item banks, static continuous screeners, and legacy scales based on the development sample are provided in Figure 1. The item banks provided information along the mild to severe range of the continuum. The static screeners demonstrated a similar coverage of the latent continuum in comparison to the full item banks despite the reduced magnitude of information. The reduced magnitude of information among the screeners in comparison to the full bank was expected given the substantially reduced number of items contained in the screeners. The panic disorder and OCD static screeners provided more information in comparison to the comparable length PADIS and SOCS legacy screeners, suggesting that the static screeners generate a more precise indication of latent severity in comparison to the legacy scales. Similarly, the SAD static screener provided more information between -1 and 2 standard deviations from the mean on the latent trait in comparison to the SOPHS. However, the SOPHS provided more information in the upper severity range (> 2 standard deviations) in comparison to the static screener.

The results of the CAT simulations using the development sample are provided in Table 2. In terms of striking a balance between efficiency and precision, the CAT simulations with the termination rules of standard error < 0.3 and difference in theta scores from one item to the next <0.05 performed the best across all three disorders. On average, the number of items presented by the CAT was four (79% reduction), four (85% reduction), and eight (82% reduction) for panic disorder, SAD and OCD, respectively. The mean standard errors for

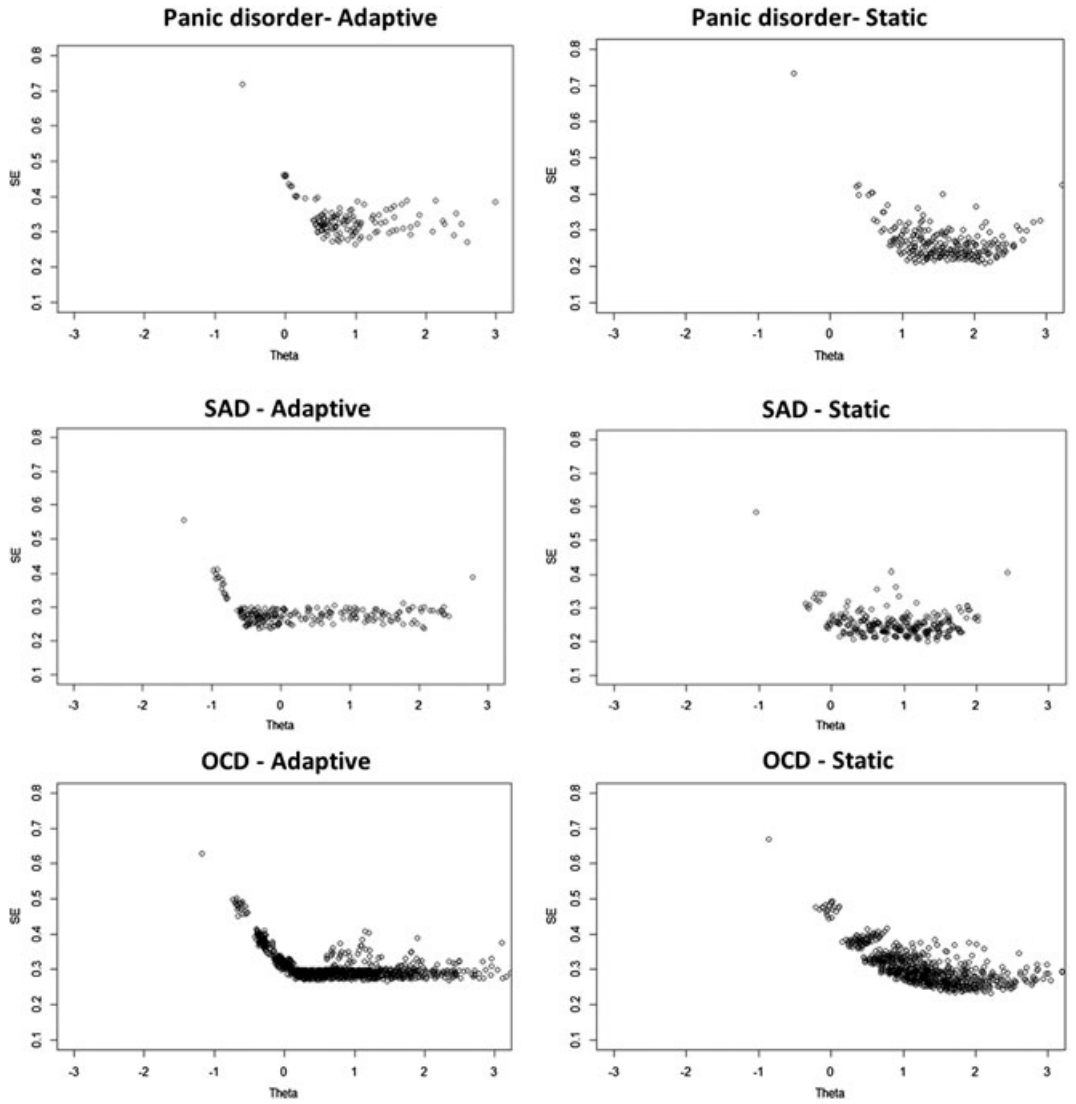

FIGURE 2 Scatterplots for standard errors by estimated theta scores in adaptive and static continuous screeners for panic disorder, social anxiety disorder (SAD), and obsessive compulsive disorder (OCD) in development sample (n = 3175)

all three CATs were 0.54 (panic disorder), 0.33 (SAD) and 0.41 (OCD) and all three were highly correlated with the respective item bank (*r* values >0.93). In comparison, the mean standard errors for the static continuous screeners were 0.59, 0.36, and 0.49 for panic disorder, SAD, and OCD, respectively. Correlations between the static and full banks were slightly lower than correlations between the adaptive tests and full banks. Similarly, bias, and RMSD values were all slightly higher between the static versions and the full banks in comparison to the adaptive tests and the full banks. Figure 2 provides the standard errors plotted against the estimated theta scores in the development sample for both static and adaptive screeners. Inspection of the plots in Figure 2 suggests that the adaptive tests resulted in a more consistent level of precision, as indicated by a larger number of individuals at or below the 0.3 cutoff across the moderate to severe range of the severity than the static screeners, particularly for the SAD and OCD screeners.

Pearson's correlations comparing the full bank, static screener, adaptive screener, and legacy scales in the development sample are provided in Table 3. The item banks and screeners demonstrated convergent and divergent validity with correlations substantially higher with legacy scales purportedly measuring the same disorder than correlations between legacy scales measuring different disorders. Correlations between the item banks and the static and adaptive screeners were very high (*r* values >0.88) indicating a high degree of similarity in the theta scores.

In the separate validation sample, scores generated by the static and adaptive screeners were compared to scores generated by the item banks. As can be seen in Table 4, the correlations remained very high (*r* values >0.89) and the adaptive tests administered on average a similar or lower number of items in the validation sample than the development sample. Likewise, the mean standard error for the adaptive tests in the validation sample was slightly lower than those demonstrated in the development sample. There was some indication that the scores generated by the static screeners in the validation sample were slightly biased towards a higher severity in comparison to the item bank. However, this bias was minimal with the largest mean difference of -0.17 standard deviations from the mean for the panic disorder static screener.

4 | DISCUSSION

The current study sought to develop a series of adaptive and static screeners from recently constructed item banks for panic disorder,

TABLE 3 Pearson's correlations between screeners in development sample (*n* = 3175)

	1	2	3	4	5	6	7	8	9	10	11	12
1. PADIS	**1.00**											
2. Panic full item bank	**0.70**	**1.00**										
3. Panic static screener	**0.73**	**0.88**	**1.00**									
4. Panic adaptive (CAT 4) screener	**0.70**	**0.93**	**0.88**	**1.00**								
5. SOPHS	0.49	0.56	0.50	0.54	**1.00**							
6. SAD full item bank	0.44	0.59	0.49	0.57	**0.78**	**1.00**						
7. SAD static screener	0.45	0.58	0.49	0.56	**0.77**	**0.94**	**1.00**					
8. SAD adaptive (CAT 4) screener	0.43	0.56	0.47	0.54	**0.76**	**0.96**	**0.95**	**1.00**				
9. SOCS	0.35	0.40	0.37	0.38	0.41	0.44	0.41	0.41	**1.00**			
10. OCD full item bank	0.38	0.52	0.45	0.48	0.47	0.58	0.52	0.53	**0.70**	**1.00**		
11. OCD static screener	0.36	0.49	0.43	0.45	0.46	0.55	0.50	0.50	**0.70**	**0.90**	**1.00**	
12. OCD adaptive (CAT 4) screener	0.35	0.47	0.41	0.44	0.43	0.53	0.48	0.49	**0.68**	**0.93**	**0.91**	**1.00**

Note: bold typeface indicates correlations among screeners measuring the same disorder.

TABLE 4 Comparisons between static and adaptive screeners and the full item banks in a separate validation sample (*n* = 3577)

	$M_{(items)}$	$M_{(theta)}$	$M_{(SE)}$	Correlation	Bias	RMSD
Panic disorder						
Panic full item bank	19	0.13	0.38	—	—	—
Panic static screener	4	0.30	0.50	0.92	−0.17	0.37
Panic adaptive screener	4	0.19	0.49	0.92	−0.06	0.34
Social anxiety disorder (SAD)						
SAD full item bank	26	0.16	0.16	—	—	—
SAD static screener	4	0.23	0.31	0.93	−0.07	0.33
SAD adaptive screener	3	0.21	0.30	0.92	−0.06	0.33
Obsessive compulsive disorder (OCD)						
OCD full item bank	44	0.14	0.25	—	—	—
OCD static screener	7	0.21	0.45	0.89	−0.07	0.43
OCD adaptive screener	7	0.10	0.40	0.92	0.04	0.39

SAD, and OCD using data-driven techniques. This involved comparing the efficiency and precision of static versus adaptive approaches to screening for mental disorders in a community sample of Australian adults. In general, the data-driven techniques were able to select a subset of items for each disorder that provided similar scores in comparison to the full item banks. These methods led to considerable reduction in administered items ranging from 79% to 89%. The screeners demonstrated high rates of precision across the moderate to severe range of severity and sufficient convergent and discriminant validity with respect to legacy scales. Importantly, the results indicated that all of the screeners performed similar if not better in an independent validation sample of Australian adults, compared to the original development sample.

The results of the current study demonstrated similar gains in efficiency and minimal declines in precision using an adaptive algorithm in comparison to other studies that have utilized CAT for mental health constructs, such as depression, anxiety, and addiction (Becker et al., 2008; Fliege et al., 2005; Gibbons et al., 2008, 2012, 2014; Kirisci et al., 2012; Smits, Cuijpers, & van Straten, 2011; Walter et al., 2007). Yet the results of the static screeners demonstrated a similar level of precision, albeit with slightly higher mean standard errors, using equivalent number of items as the adaptive algorithms. Based on these results, the question can be asked as to whether the added complexity imposed by the CAT is justified for mental health assessment and monitoring?

To answer this question, a closer inspection of the results across the continuum of latent severity is required. As can be seen in Figure 2, the adaptive algorithms were able to maintain a more consistent level of precision across a wider range of the continuum than the static screeners. This is despite the fact that items were purposely selected to maintain information across the spectrum of the trait continuum covered by the full bank during the static form development. Higher rates of precision are critical when mental health outcomes are continually monitored over multiple occasions and when priority is given to detecting minor or slight changes in an individual's health status rather than noise in the assessment procedure. Adaptive algorithms might consequently be most suitable for use in tailored treatments or measurement-based (precision) medicine that seek to alter or modify an individual's mental health treatment regime based on observed changes in mental health status (Fortney et al., 2017; Harding, Rush, Arbuckle, Trivedi, & Pincus, 2011; Zimmerman, McGlinchey, & Chelminski, 2009). However, if basic screening in the community is desired, particularly for initial assessment and mental health triage, then the simplicity and efficiency of the static forms may be sufficient.

The current study had several limitations that require further discussion. First, the use of Facebook as an advertising medium could introduce some degree of self-selection bias. However, Facebook users represent a substantial majority of the Australian public and previous studies have indicated that the use of Facebook advertising generates samples that are similarly representative to other traditional mediums (Thornton et al., 2016). We also utilized a weighting scheme based on key variables of the Australian population to reduce the impact of self-selection bias on the final results. Regardless, the screeners developed in the current study require additional testing in other samples to ensure replicability of these findings, particularly in clinical samples where higher rates of symptomatology are typically observed. Second, the current study utilized methods that contain multiple parameters that can influence efficiency and precision. While the current study employed a comprehensive range of methods, it was unable to examine all possible combinations of parameters.

The screeners developed in the current study represent some of the newest data-driven techniques to determine severity of panic disorder, SAD, and OCD in the general population. These screeners demonstrated significant reductions in items administered in comparison to the item banks without substantial losses in precision. Moreover, performance of these screeners was comparable to existing legacy screeners and the results replicated in a separate sample of the Australian community. In conclusion, the screeners developed for panic disorder, SAD, and OCD could be used to measure and monitor the severity of psychopathology in tailored treatment protocols.

ACKNOWLEGEMENTS

MS, PB, and AC are supported by National Health and Medical Research Council (NHMRC) Fellowships 1052327, 1083311, and 1013199, respectively. The current study was funded by NHMRC project grant 1043952.

DECLARATION OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

REFERENCES

Batterham, P. J., Brewer, J. L., Tjhin, A., Sunderland, M., Carragher, N., & Calear, A. L. A. L. (2015). Systematic item selection process applied to developing item pools for assessing multiple mental health problems. *Journal of Clinical Epidemiology*, *68*(8), 913–919. doi:10.1016/j.jclinepi.2015.03.022

Batterham, P. J., Mackinnon, A. J., & Christensen, H. (2016a). Community-based validation of the social phobia screener (SOPHS). *Assessment*. doi:10.1177/1073191116636448

Batterham, P. J., Sunderland, M., Carragher, N., & Calear, A. L. (2016b). Development and community-based validation of eight item banks to assess mental health. *Psychiatry Research*, *243*, 452–463. doi:10.1016/j.psychres.2016.07.011

Batterham, P. J., Sunderland, M., Carragher, N., Calear, A. L., Mackinnon, A. J., & Slade, T. (2016b). The distress questionnaire-5: Population screener for psychological distress was more accurate than the K6/K10. *Journal of Clinical Epidemiology*, *71*, 35–42. doi:10.1016/j.jclinepi.2015.10.005

Becker, J., Fliege, H., Kocalevent, R. D., Bjorner, J. B., Rose, M., Walter, O. B., & Klapp, B. F. (2008). Functioning and validity of a computerized adaptive test to measure anxiety (A CAT). *Depression and Anxiety*, *25*(12), 182–194. doi:10.1002/da.20482

Burgess, P. M., Pirkis, J. E., Slade, T. N., Johnston, A. K., Meadows, G. N., & Gunn, J. M. (2009). Service use for mental health problems: Findings from the 2007 National Survey of mental health and wellbeing. *The Australian and New Zealand Journal of Psychiatry*, *43*(7), 615–623. doi:10.1080/00048670902970858

Cella, D., Gershon, R., Lai, J.-S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*(S1), 133–141. doi:10.1007/s11136-007-9204-6

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–38.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229. doi:10.1177/014662169602000303

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*(1), 125–136. doi:10.1007/s11136-009-9560-5

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression. *Journal of Medical Internet Research*, *11*(2), e13. doi:10.2196/jmir.1194

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, *14*(10), 2277–2291. doi:10.1007/s11136-005-6651-9

Fortney, J. C., Unützer, J., Wrenn, G., Pyne, J. M., Smith, G. R., Schoenbaum, M., & Harbin, H. T. (2017). A tipping point for measurement-based care. *Psychiatric Services*, *68*(2), 179–188. doi:10.1176/appi.ps.201500439

Gibbons, R., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., … Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, *59*(4), 361–368. doi:10.1176/ps.2008.59.4.361

Gibbons, R., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). The CAT-DI: Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*(11), 1104–1112. doi:10.1001/archgenpsychiatry.2012.14

Gibbons, R., Weiss, D., Pilkonis, P., Frank, E., Moore, T., Kim, J., & Kupfer, D. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *American Journal of Psychaitry*, *171*(2), 187–194. doi:10.1038/nature13314.A

Gibbons, R., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, *12*(1), 83–104. doi:10.1146/annurev-clinpsy-021815-093634

Harding, K. J. K., Rush, A. J., Arbuckle, M., Trivedi, M. H., & Pincus, H. A. (2011). Measurement-based care in psychiatric practice. *The Journal of Clinical Psychiatry*, *72*(8), 1136–1143. doi:10.4088/JCP.10r06282whi

Hilvert-Bruce, Z., Rossouw, P. J. P. J., Wong, N., Sunderland, M., & Andrews, G. (2012). Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders. *Behaviour Research and Therapy*, *50*(7–8), 463–468. doi:10.1016/j.brat.2012.04.001

Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & Ustün, T. B. (2007). Age of onset of mental disorders: A review of recent literature. *Current Opinion in Psychiatry*, *20*(4), 359–364. doi:10.1097/YCO.0b013e32816ebc8c

Kessler, R. C., Ruscio, A. M., Shear, K., & Wittchen, H.-U. (2010). Epidemiology of anxiety disorders. In M. B. Stein, & T. Steckler (Eds.), *Behavioral neurobiology of anxiety and its treatment*. (pp. 21–37). New York: Springer.

Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M., & Wittchen, H. U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research*, *21*(3), 169–184. doi:10.1002/mpr.1359

Kirisci, L., Tarter, R., Reynolds, M., Ridenour, T., Stone, C., & Vanyukov, M. (2012). Computer adaptive testing of liability to addiction: Identifying individuals at risk. *Drug and Alcohol Dependence*, *123*(Suppl. 1), S79–S86. doi:10.1016/j.drugalcdep.2012.01.016

Muthen, L. K., & Muthen, B. O. (2015). *Mplus users' guide*. Los Angeles, CA: Muthen & Muthen.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the patient reported outcomes measurement information system (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*(3), 263–283. doi:10.1177/1073191111411667

Rose, M., Bjorner, J. B., Fischer, F., Anatchkova, M., Gandek, B., Klapp, B. F., & Ware, J. E. (2012). Computerized adaptive testing - ready for ambulatory monitoring? *Psychosomatic Medicine*, *74*(4), 338–348. doi:10.1097/PSY.0b013e3182547392

Siu, A. L., Bibbins-Domingo, K., Grossman, D. C., Baumann, L. C., Davidson, K. W., Ebell, M., … Pignone, M. P. (2016). Screening for depression in adults. *JAMA*, *315*(4), 380–387. doi:10.1001/jama.2015.18392

Slade, T., Johnston, A., Oakley Browne, M. A., Andrews, G., & Whiteford, H. (2009). 2007 National Survey of mental health and wellbeing: Methods and key findings. *Australian and New Zealand Journal of Psychiatry*, *43*(7), 594–605.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, *188*(1), 147–155. doi:10.1016/j.psychres.2010.12.001

Spitzer, R. L., Kroenke, K., & Williams, J. B. W. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *Journal of the American Medical Association*, *282*(18), 1737–1744.

Mendlowicz, M. V., & Stein, M. B. (2000). Quality of life in individuals with anxiety disorders. *American Journal of Psychiatry*, *157*(5), 669–682. doi:10.1176/appi.ajp.157.5.669

Thompson, A., Issakidis, C., & Hunt, C. (2008). Delay to seek treatment for anxiety and mood disorders in an Australian clinical sample. *Behaviour Change*, *25*(2), 71–84. doi:10.1375/bech.25.2.71

Thornton, L., Batterham, P. J., Fassnacht, D. B., Kay-Lambkin, F., Calear, A. L., Hunt, S., … Ho, R. C. M. (2016). Recruiting for health, medical or psychosocial research using Facebook: Systematic review. *Internet Interventions*, *4*(1), 72–81. doi:10.1016/j.invent.2016.02.001

Uher, R., Heyman, I., Mortimore, C., Frampton, I., & Goodman, R. (2007). Screening young people for obsessive-compulsive disorder. *British Journal of Psychiatry*, *191*, 353–354. doi:10.1192/bjp.bp.106.034967

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for "Anxiety" (Anxiety-CAT). In *Quality of Life Research* (Vol. 16). (pp. 143–155). Dordrecht: Springer DOI: 10.1007/s11136-007-9191-7.

Zimmerman, M., McGlinchey, J. B., & Chelminski, I. (2009). Measurement-based care and outcome measures: Implications for practice. In T. Schwartz, & T. Petersen (Eds.), *Depression: Treatment strategies and management* (2nd ed.). (pp. 127–138). New York: Informa Healthcare.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Sunderland M, Batterham PJ, Calear AL, Carragher N. The development and validation of static and adaptive screeners to measure the severity of panic disorder, social anxiety disorder, and obsessive compulsive disorder. *Int J Methods Psychiatr Res*. 2017;e1561. https://doi.org/10.1002/mpr.1561