



Development of a computerized adaptive screening tool for overall psychopathology (“p”)*

Tyler M. Moore^{a,*}, Monica E. Calkins^a, Theodore D. Satterthwaite^a, David R. Roalf^a,
Adon F.G. Rosen^a, Ruben C. Gur^{a,b}, Raquel E. Gur^a

^a Department of Psychiatry, Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

^b VISN4 Mental Illness Research, Education, and Clinical Center at the Philadelphia VA Medical Center, Philadelphia, PA, 19104, USA

ARTICLE INFO

Keywords:

Psychopathology
Computer adaptive testing
Item response theory
Psychometrics
Validation
Clinical screen

ABSTRACT

A substantial body of work supports the existence of a general psychopathology factor (“p”). Psychometrically, this is important because it implies that there is a psychological phenomenon (overall psychopathology) that can be measured and potentially used in clinical research or treatment. The present study aimed to construct, calibrate, and begin to validate a computerized adaptive (CAT) screener for “p”. In a large community sample (N = 4544; age 11–21), we modeled 114 clinical items using a bifactor multidimensional item response theory (MIRT) model and constructed a fully functional (and public) CAT for assessing “p” called the Overall mental illness (OMI) screener. In a random, non-overlapping sample (N = 1019) with extended phenotyping (neuroimaging) from the same community cohort, adaptive versions of the OMI screener (10-, 20-, and 40-item) were simulated and compared to the full 114-item test in their ability to predict demographic characteristics, common mental disorders, and brain parameters. The OMI screener performed almost as well as the full test, despite being only a small fraction of the length. For prediction of 13 mental disorders, the mid-length (20-item) adaptive version showed mean area under the receiver operating characteristic curve of 0.76, compared to 0.79 for the full version. For prediction of brain parameters, mean absolute standardized relationship was 0.06 for the 20-item adaptive version, compared to 0.07 for the full form. This brief, public tool may facilitate the rapid and accurate measurement of overall psychopathology in large-scale studies and in clinical practice.

While psychiatric disorders are generally conceptualized as distinct but overlapping clusters of symptoms, such as embodied in the Diagnostic and Statistical Manual of the American Psychiatric Association (American Psychiatric Association, 2011), there is an accumulating body of work supporting the existence of a general psychopathology factor (“p”) (Caspi et al., 2014). The discovery of the p factor was presaged by the overwhelming comorbidity consistently found among mental disorders (Kessler et al., 2005), and indeed, most factor analyses of psychopathology data find that the higher-order classifications of psychopathologies (e.g. internalizing, externalizing, etc.) are themselves correlated. For example, Krueger (1999), Krueger et al. (1998), and others have shown that mental disorders can usually be grouped into three *correlated* dimensions: externalizing, anxious-misery, and fear. Our clinical interview included a fourth disorder type, Psychosis, for a total of 4 factors.

An important takeaway from the above is that endorsement of any

psychopathological symptom increases the probability of endorsement of any other symptom, regardless of whether the symptoms are part of the same disorder (or even higher-order domain). Psychometrically this is important because it implies that there is a psychological phenomenon (overall psychopathology) that can be measured and potentially used in clinical research or treatment. Despite the overwhelming evidence for the existence of “p”, at present it is usually measured by aggregating responses across extensive, time-consuming clinical interviews. In other domains including cognitive assessment and “high-stakes” testing (e.g. Graduate Record Examinations), modern measurement tools such as item response theory (IRT) and computerized adaptive testing (CAT) provide nearly-equivalent accuracy in a fraction of the time. Despite this, the considerable advantages of such procedures have not been applied to the measurement of “p”.

IRT (Embretson and Reise, 2009; Lord, 1980) is a psychometric method that focuses on various characteristics of individual test or scale

* This work was supported by the Lifespan Brain Institute (LiBI); NIMH grants MH089983, MH019112, MH096891, MH107235, and MH107703; and the Dowshen Program for Neuroscience.

* Corresponding author. Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104, USA.

E-mail address: tymoore@pennmedicine.upenn.edu (T.M. Moore).

items (rather than a test/scale as a whole). One of the most common IRT models is the 2-parameter model described by the following equation:

$$p_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \quad (1)$$

where $p_i(\theta)$ is the probability of endorsement (or a correct response, in the case of cognitive testing), a_i is the item discrimination, b_i is the item difficulty, and θ is the trait level of the person (in our case, overall mental illness, OMI). The discrimination parameter, a_i , determines how precisely the item can place an individual on a trait spectrum; higher discrimination is always better. The difficulty parameter, b_i , determines how high on the latent trait continuum one has to be in order to have a 50% chance of endorsing the item. For example, a depression item that asked about current plans to commit suicide would presumably be a “difficult” item because one would need to be fairly high on the Depression continuum to endorse that item. The more likely an item is to be endorsed at any given trait level, the lower its difficulty (i.e. the “easier” it is).

Because we anticipated the item response data analyzed here to be multidimensional, we used multidimensional IRT (MIRT) (Edwards et al., 2014; Reckase, 2009), an extension that allows modeling of multiple dimensions simultaneously. The unidimensional model shown in Eq. (1) can be expanded to account for multiple dimensions as follows:

$$p_i(x = 1 | \theta_1, a_{i1}, b_{i1}, \theta_2, a_{i2}, b_{i2}) = \frac{e^{a_{i1}(\theta_1 - b_{i1}) + a_{i2}(\theta_2 - b_{i2})}}{1 + e^{a_{i1}(\theta_1 - b_{i1}) + a_{i2}(\theta_2 - b_{i2})}} \quad (2)$$

where $p_i(x = 1)$ is the probability of a correct response, a_{i1} is the item discrimination for dimension 1, a_{i2} is the item discrimination for dimension 2, b_{i1} is the item difficulty for dimension 1, b_{i2} is the item difficulty for dimension 2, θ_1 is the ability level of the person on dimension 1, and θ_2 is the ability level of the person on dimension 2. This model can be expanded to account for as many dimensions as necessary and in almost any structural configuration necessary. Here, because we wanted to model not only the multiple dimensions but also a single overall (“general”) factor, we used a specific type of MIRT model configuration called a bifactor model (Reise, Steven P., 2012; Reise, Steven P. et al., 2010). In a bifactor model, items load on up to two factors simultaneously—their own “specific” factor (e.g. “internalizing symptoms”) and a general factor (OMI). The bifactor model allows accurate estimation of item parameters on the overall factor in the presence of multidimensionality, which can cause biased discrimination parameter estimates if ignored (Reise, S. P. et al., 2015; Reise, Steven P. et al., 2011).

An advantage of IRT (in this case, MIRT) is that its emphasis on individual items allows using the parameter estimates produced by the model to assess the overall quality of the items. Here, “quality” is determined by the amount of information provided by the item at any given trait level, where information is represented by the following equation:

$$I(\theta) = a_i^2 p_i(\theta) q_i(\theta) \quad (3)$$

where $I(\theta)$ is the information produced by the item, a_i is the item discrimination, $p_i(\theta)$ is the probability of endorsement, and $q_i(\theta)$ is the probability of non-endorsement. Three key features are notable about this equation. First, as a_i goes up, information always goes up, thus higher discrimination parameters are always good. Second, the amount of information provided by the item differs by person. For example, keeping a_i constant at 1.0, a person low on the trait might have $p_i(\theta) = 0.1$ and $q_i(\theta) = 0.9$, for total information of $0.1 \times 0.9 = 0.09$, while a person slightly higher on the trait might have $p_i(\theta) = 0.2$ and $q_i(\theta) = 0.8$, for total information of $0.2 \times 0.8 = 0.16$. Finally, all items produce maximum information when administered to a person with a 50% probability of endorsement—i.e. where both $p_i(\theta)$ and $q_i(\theta)$ are 0.5. An implication of this last point is that items provide maximum

information when administered to someone whose trait level (θ) is the same as the item difficulty (b_i).

These parameters pave the way for arguably one of the most groundbreaking applications of IRT, computerized adaptive testing (CAT) (Weiss and Kingsbury, 1984; Wainer and Dorans, 2000), which we have previously used to develop short and adaptive forms of cognitive and clinical tests (Moore et al., 2015, 2018; Roalf et al., 2016a,b). In CAT, after the first item-administration (and response), a scoring algorithm estimates the examinee's trait level, and based on this rough estimate, chooses the most appropriate next item to administer, where “most appropriate” is determined by how much information it will provide. After this next item administration (and response), the algorithm then uses both item responses to estimate the examinee's ability. Then the next most appropriate item is selected, and so on. The test stops when some stopping criterion is met—e.g. when the examinee's standard error of measurement reaches some lower bound. A step-by-step example of the CAT item-selection process is provided in the Supplement. Note that because IRT scoring is based on the item parameters (a_i and b_i), endorsement of an “easy” item will affect a person's score differently than endorsement of a “difficult” item. Because all items were calibrated in the same model (the bifactor MIRT model described above), examinees can be scored on the same scale even if the sets of items administered to each examinee are completely non-overlapping. This feature is especially useful for longitudinal studies, where item repetition can be problematic.

The purpose of the present study was to develop and validate a CAT assessment of overall mental illness (OMI). To do so, we calibrated a set of 114 clinical items for use as an item bank for CAT assessment. Then, we examined the external validity of scores from this adaptive OMI assessment, compared to validity of scores on the full 114-item scale. Multiple criteria were used for predictive validity assessment, including demographic information, environmental experiential information (e.g. neighborhood crime rates), and multimodal brain parameters (volume, cerebral blood flow, and fractional anisotropy). The brain areas selected for use as validity criteria were based on established research suggesting relationships among multiple psychopathologies and the amygdala (Bogerts et al., 1993; Cullen et al., 2014; Feinstein et al., 2013; Oler et al., 2016; Rajarethinam et al., 2001; Siegle et al., 2007; Sterzer et al., 2007), anterior insula (Alvarez et al., 2015; Makris et al., 2006; Moran et al., 2013; Paulus and Stein, 2006; Penner et al., 2016; Senatorov et al., 2014; Uddin and Menon, 2009), and hippocampus (Bogerts et al., 1993; Bois et al., 2015; Bremner et al., 2000; Erk et al., 2014; Kwon, 2003; Plessen et al., 2006; Videbech and Ravnkilde, 2004). Similarly, fractional anisotropy of the cingulate gyrus (Han et al., 2013; Nakamae et al., 2008; Peterson et al., 2011; Wang et al., 2004), superior longitudinal fascicles (Bennett et al., 2013; Harris et al., 2008; Svátkova et al., 2016; Zou et al., 2008), and the uncinate fasciculus (Kitis et al., 2012; Konrad et al., 2010; Kubicki et al., 2002; Phan et al., 2009; Sarkar et al., 2012; Von Der Heide, R J et al., 2013; Zarei et al., 2011) have been implicated in diverse psychopathologies. Consequently, we used volume and cerebral blood flow (CBF) of the amygdala, anterior insula, and hippocampus, and we used fractional anisotropy (FA) of the cingulate gyrus, superior longitudinal fascicles, and uncinate fasciculus. We also selected a fMRI measure of activation of the anterior insula during a working memory task. Task-Activation in various brain regions during functional MRI has been implicated in numerous psychiatric disorders, including depression (Smith et al., 2011), bipolar depression (Marchand et al., 2007), psychosis (Fusar-Poli et al., 2011), anorexia (Castro-Fornieles et al., 2010), and others (Dibbets et al., 2010).

Specific predictions related to the above validity analyses were as follows: 1) OMI will be negatively associated with mother's education, positively associated with trauma exposure, and unrelated to age; 2) OMI will be negatively related to neighborhood SES and positively related to neighborhood crime rates; 3) OMI will be positively associated with resting CBF and negatively associated with regional brain volumes

Table 1
Clinical diagnostic rates and demographic characteristics of samples.

Characteristic	Full Sample	Calibration Sample	Validation Sample
N	5563	4544	1019
Mean Age, years (SD)	15.7 (2.6)	15.7 (2.7)	15.7 (2.5)
Proportion Female	0.54	0.54	0.52
Proportion European-American	0.57	0.6	0.45
Proportion African-American	0.32	0.3	0.45
Proportion Other Race	0.11	0.1	0.1
Mean Maternal Edu, years (SD)	14.4 (2.4)	14.5 (2.4)	14.2 (2.4)
Depression, prop.	0.15	0.14	0.17
GAD, prop.	0.03	0.03	0.02
Psychosis Risk, prop.†	0.25	0.24	0.33
Mania, prop.	0.01	0.01	0.01
Separation Anxiety, prop.	0.04	0.04	0.04
Specific Phobia, prop.	0.30	0.30	0.31
Social Phobia/Anxiety, prop.	0.23	0.22	0.27
Panic Disorder, prop.	0.02	0.02	0.01
Agoraphobia, prop.	0.06	0.06	0.08
OCD, prop.	0.03	0.03	0.03
ADD, prop.	0.16	0.15	0.17
ODD, prop.	0.33	0.33	0.37
Conduct Disorder, prop.	0.07	0.07	0.11

Note. SD = standard deviation; prop. = proportion; Edu = education; GAD = generalized anxiety disorder; OCD = obsessive-compulsive disorder; ADD = attention deficit disorder; ODD = oppositional defiant disorder; †Psychosis Risk is not a diagnosis as such, but rather, comparable to Clinical High Risk reported in other studies.

and FA; 4) OMI will be negatively associated with activation of the anterior insula during a working memory task.

1. Method

1.1. Participants

Participants in this study were from the Philadelphia Neurodevelopmental Cohort (PNC), which has been described in detail elsewhere (Calkins et al., 2015). Briefly, the PNC is a community sample of $N = 9498$ individuals age 8–21 years (mean = 14.2), 51.7% female, 55.8% European American, 32.9% African-American, from the greater Philadelphia area. In the present study, age range was limited to 11–21 (participants age 8–10 were not directly interviewed), and the sample was randomly split into calibration and validation samples. Table 1 shows the demographic characteristics and clinical diagnosis rates for these samples, as well as the combined (full) sample for the study. As expected in a community sample, rates of clinical diagnoses in this age range mostly reflect national rates. For example, depression, obsessive-compulsive disorder, panic disorder, and ADHD have lifetime prevalence rates (Table 1) of 15%, 3%, 2%, and 16%, respectively. These (and other diagnostic proportions in Table 1) approximate those found by others in the US (Merikangas et al., 2010).

1.2. Measures

Clinical. The clinical assessment used in the PNC has been described in detail elsewhere (Calkins et al., 2014, 2015). Briefly, psychopathology was assessor administered using a computerized, structured interview (GOASSESS), which was abbreviated and modified from the NIMH Genetic Epidemiology Research Branch Kiddie-SADS (Merikangas et al., 2009). The interview included a timeline of life events, demographics and medical history, psychopathology screen, Children's Global Assessment Scale (C-GAS) (Shaffer, 1983), and interviewer observations. Items for the current investigation were 114

dichotomous items drawn from the psychopathology screen. These items assessed psychiatric and psychological treatment history, and lifetime occurrence of major domains of psychopathology including psychosis, mood (major depressive episode, manic episode), anxiety (generalized anxiety disorder, separation anxiety disorder, specific phobia, social phobia, panic disorder, agoraphobia, obsessive-compulsive disorder, post-traumatic stress disorder), attention deficit/hyperactivity disorder (ADHD), behavioral (oppositional defiant disorder, conduct disorders), eating disorders (anorexia, bulimia), and suicidal thinking and behavior. Computerized algorithms assigned rankings reflecting DSM-IV categories of each psychopathology domain based on endorsement of contributing items by either the proband or collateral (behavioral domains), or by the proband (all other domains). Psychopathology categories were considered significant if sufficient symptoms were endorsed with frequency and duration approximating DSM-IV disorder or episode criteria, accompanied by significant distress or impairment rated ≥ 5 on an 11 point.

In addition, two screening tools to assess subthreshold psychosis spectrum symptoms were embedded within the psychopathology screen (Calkins et al., 2014). First, positive subschizophrenic symptoms were assessed using the PRIME Screen-Revised [PS-R (Kobayashi et al., 2008; Miller et al., 2004)] which contains 12 items rated on a 7-point Likert scale (Extremely Disagree to Extremely Agree) and which were split dichotomously for the current analyses. Second, negative/disorganized symptoms were assessed using six assessor-rated subscales of the Scale of Prodromal Symptoms (SOPS) from the Structured Interview for Prodromal Syndromes (SIPS v. 4.0) (Miller et al., 2003): Avolition; Expression of Emotion; Experience of Emotions and Self; Occupational Functioning; Trouble with Focus and Attention; Disorganized Communication. Ratings on the SOPS were based on selected closed (yes/no) questions from the SIPS, which were used in the present investigation.

Neuroimaging. Neuroimaging acquisition, processing, and quality assurance has been described in detail elsewhere (Satterthwaite et al., 2016a,b; Satterthwaite et al., 2014a,b). Briefly, 1601 youth who completed clinical and cognitive assessment as part of the PNC also completed multi-modal neuroimaging (Satterthwaite et al., 2014a,b; Satterthwaite et al., 2016a,b). All participants were imaged at a single site using the same scanner (Siemens 3T Tim Trio; 32 channel head coil). Imaging sequences provided complementary information regarding brain structure, organization, metabolism, and function. A T1-weighted image provided information regarding brain structure (Satterthwaite et al., 2016a,b; Rosen et al., 2018); an arterial-spin labeled (ASL) image provided an estimate of cerebral perfusion (Kaczurkin et al., 2018; Satterthwaite, T. D. et al., 2014a,b); a diffusion-weighted image provided information regarding white matter microstructure (Roalf et al., 2016a,b; Vandekar et al., 2015). Finally, a fractal version of the n-back working memory task provided information regarding functional responses to working memory load, with the contrast of interest being 2-back versus 0-back (Satterthwaite, T. D. et al., 2013; Shanmugan et al., 2016). All image processing and quality assurance is as previously described. In order to derive consistent regional summary measures across sequences, the structural image was segmented using an advanced multi-atlas labelling procedure with joint label fusion (Wang, H. et al., 2013; Wang, Hongzhi and Yushkevich, 2013). Within these regions, volume (from the T1 image), cerebral blood flow (from the ASL image), fractional anisotropy (from the diffusion image), and activation to working memory load (from the n-back task) was measured. To reduce multiple comparisons, specific regions with previously-described associations with overall psychopathology were selected for further analysis.

Geographic/Environmental. Two measures used here—“Neighborhood SES” and “Neighborhood Crime”—are based on census and crime data linked to participants' block groups. The methods and analyses used to derive these scores are detailed in Moore et al. (2016). Briefly, Neighborhood SES comprises census variables reflective of socio-economic status in that area—e.g. median family income,

percent of residents who are married, percent of residents who are in poverty, percent of real estate lots that are vacant, etc. Higher values of Neighborhood SES indicate higher SES. Neighborhood Crime comprises rates (per 100 people) of 49 crimes in each block group—e.g. robbery, aggravated assault, arson, etc. Higher values of Neighborhood Crime indicate more crime in that participant's living environment.

1.3. Analysis

First, a validation sample ($N = 1019$) was selected (based on whether they had neuroimaging) and set aside for validation purposes (see below); all non-validation analyses described below were therefore run on the remaining ($N = 4544$) participants meeting age criteria after removing the validation sample. In preparation for the bifactor IRT model, we conducted an exploratory factor analysis (EFA) on the 114 psychopathology items based on the matrix of tetrachoric inter-item correlations (within the *psych* package in R). The number of factors to extract (4) was determined by theory, as described above. Shanmugan et al. (2016) performed a nearly identical analysis on a super-set of these data using 112 items instead of 114, and they found that the four-factor solution cleanly replicates previous literature, with the 4 factors being anxious-misery, fear, psychosis, and externalizing.

The EFA results were used to assign items to factors for the bifactor IRT model—specifically, a bifactor 2-parameter logistic (2PL) model (Gibbons and Hedeker, 1992; Gibbons et al., 2007) implemented in the *mirt* package (Chalmers, 2012) in R using Bock-Aitkin expectation-maximization (BAEM) estimation (Bock and Aitkin, 1981). Parameter estimates from the bifactor general factor (OMI) were used as the final CAT parameters for the 114-item bank. That is, with the bifactor model estimated, discrimination parameter estimates from the general factor only were used in a unidimensional 2-parameter model along with the intercepts (transformed to traditional difficulty parameters) from the bifactor model (Weiss and Gibbons, 2007).

Parameter estimates for the item bank were entered into Firestar (Choi, 2009), a CAT-simulation program. CAT administrations were then simulated in the validation sample ($N = 1019$) to determine what would have happened if these participants had been administered an adaptive version of the full clinical interview. CAT simulations were set to stop after 10, 20, and 40 items (regardless of standard error of measurement), and participants thus received three scores corresponding to a 10-item, 20-item, and 40-item CAT. Full 114-item scores were also examined for comparison. Note that most CATs use the standard error of measurement to determine when to stop—i.e. an examinee is administered items until his/her standard error of measurement has reached an acceptably low level, and different examinees can thus receive different numbers of items. However, a problem with the SE-based approach is that it is difficult to predict how long the test will take—e.g. some examinees might reach the stopping rule threshold after only 10 items, while others must answer 20 items before reaching the stopping threshold. The difficulty in predicting administration time is not a major problem in clinical practice, but in large-scale genomic studies requiring precise scheduling of study participants' time, the variability in administration time using the SE-based stopping rule might be unacceptable. We therefore use the fixed-length stopping rule here, although the stopping rule can easily be changed in the CAT program (see below and Supplement). To be clear, the 10-, 20-, and 40-item “short forms” are administered adaptively—i.e. individuals receive different sets of items depending on their response patterns—but the length of the test is fixed.

Finally, scores from the CAT simulations described above were related to clinical categories, demographic variables, and brain parameters. For clinical categories, receiver operating characteristic (ROC) curve analysis was performed to compare the sensitivity and specificity of the four scores (10-item CAT, 20-item CAT, 40-item CAT, and full test). For demographic and brain parameters, Pearson correlations were used to compare test scores. Significance of differences among these

correlations was determined by bootstrapped confidence intervals of the difference (Efron and Tibshirani, 1986). Note that for the clinical categories, many items used in the OMI adaptive screener are identical to the items used in the diagnosis process itself, and the ROC results are therefore inflated (sometimes highly). However, the goal of this analysis is not to test whether the adaptive screener is sufficient for diagnosis, but rather to compare the (albeit inflated) ROC results for the adaptive versions with the ROC results from the full 114-item version. All results from this analysis will be inflated (including for the full version), making them useful only for comparison across forms. A related goal is to examine how the relationships of OMI scores to diagnoses change as items are added—e.g. as items are added, does the OMI score correlate more and more highly with a particular disorder?

1.4. Building the CAT-OMI for public release

The adaptive OMI for public use was built using the free, open-source software Concerto (Scalise and Allen, 2015). Information about how to install and use Concerto is provided in the Supplementary Materials, and the test itself is provided in the Supplementary.json file.

2. Results

Supplementary Table S1 shows the 4-factor EFA solution based on tetrachoric correlations, along with item text. As anticipated, the item groupings make theoretical and clinical sense; specifically, F1 = Externalizing, F2 = Anxious-Misery, F3 = Psychosis, and F4 = Fear.

Supplementary Table S2 shows the IRT parameter estimates (normal metric) for the bifactor model estimated using item assignment based on the EFA described above. Supplementary Table S3 shows the same results where item discrimination parameters have been converted to factor analysis loadings (standardized item-trait correlations). Discrimination parameter estimates were highest for items MAN002 (“Has there ever been a time when you felt so full of energy that you couldn't stop doing things and didn't get tired?”) and MAN004 (“Have there been times when you ... couldn't stop talking ... or had so many ideas in your head that you could hardly keep track of them?”), whereas they were the lowest for items PHB005 (“Have you ever been very nervous or afraid of doctors, needles, or blood?”) and PHB002 (“Have you ever been very nervous or afraid of being in really high places, like a roof or tall building?”). Difficulty parameter estimates were highest for items CDD011 (“Have you ever tricked or threatened someone into having sex with you ... ?”) and CDD010 (“Have you ever attacked someone to steal from them?”), and the lowest for items PHB001 and GAD001 (“Have you ever been a worrier?”). Average discrimination across the whole item bank was 0.84 (range = 0.26–1.94), and average difficulty was 1.15 (range = −0.04 – 2.79).

All results described below are from analyses performed using the validation sample ($N = 1019$). Supplementary Table S4 shows the results of the ROC analysis examining the ability of the adaptive scores and the full version to predict diagnosis on 13 psychological disorders, as well as the Global Assessment Scale (GAS), and Fig. 1 shows these results graphically (ordered by AUC of the full form). For the shortest (10-item) adaptive version, AUC is highest for mania and panic disorder, and lowest for specific phobia, ADD, and separation anxiety. For the 20-item adaptive version, AUC is highest for mania and OCD, and lowest for specific phobia, ADD, and separation anxiety. For the 40-item version, the AUC is highest for OCD and panic disorder, and lowest for specific phobia and separation anxiety. Finally, for the full form, AUC is highest for OCD and depression, and lowest for specific phobia and ADD. Overall, as items are added, AUC increases for all disorders except mania and GAD, both of which remain about the same regardless of the number of items. When the adaptive and full forms are compared, the full form AUC is significantly higher than all adaptive forms for separation anxiety, specific phobia, social anxiety, conduct disorder, agoraphobia, and psychosis. For the remaining seven disorders (ADD,

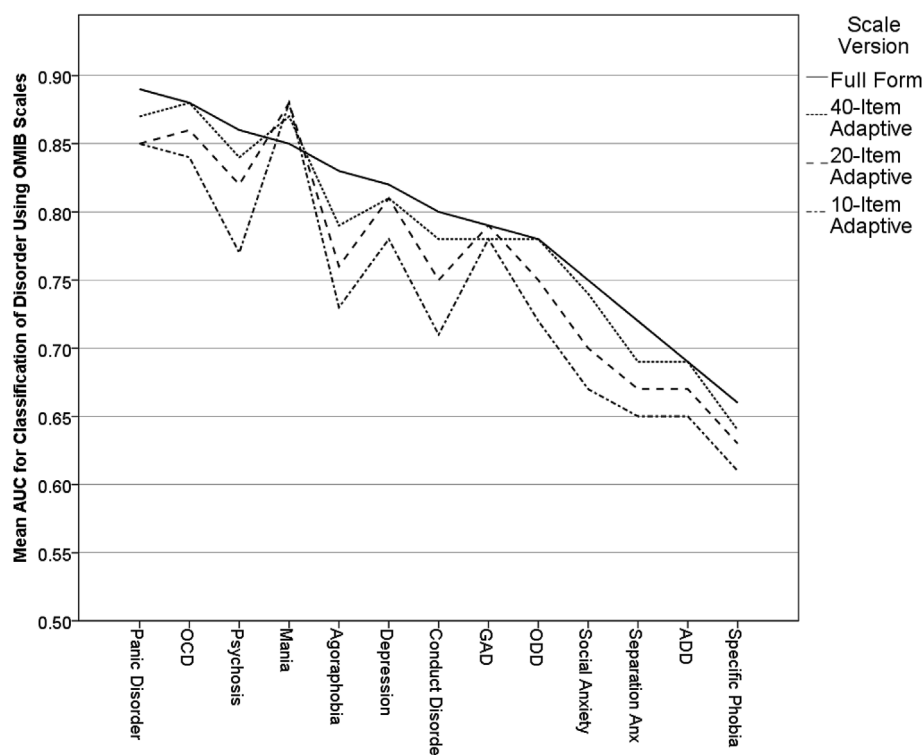


Fig. 1. Areas under the receiver operating characteristic curves for prediction of 13 common mental disorders using the full and adaptive versions of the OMI screener. Note. Ocd = obsessive-compulsive disorder; GAD = generalized anxiety disorder; ODD = oppositional defiant disorder; Anx = anxiety; add = attention deficit disorder.

ODD, GAD, depression, mania, OCD, and panic disorder), at least one of the adaptive forms shows no significant difference from the full form. For GAD and mania specifically, all adaptive forms are statistically indistinguishable from the full form.

Table 2 shows the correlations of the adaptive and full OMI scores with demographic-experiential and brain parameters. Mother's education, neighborhood SES, and trauma exposure all show significant relationships with all OMI forms (including full). Significant differences between the correlations of the full form and adaptive forms were found

Table 2
Correlations of Demographic, Environmental, and Brain Variables with Three Adaptive Versions of the Clinical Screener for Overall mental illness.

Variable	Adaptive Versions			Full (114-Item)
	10-Item	20-Item	40-Item	
Age	−0.011	−0.028	−0.020	−0.025
Mother's Edu	−0.097*	−0.127*	−0.128*	−0.148
Neighborhood SES	−0.141*	−0.185*	−0.205*	−0.227
Neighborhood Crime	0.016	0.042	0.044	0.054
Trauma Exposure	0.307*	0.340*	0.354*	0.378
Brain Measures				
Amygdala (volume)	−0.034*	−0.061	−0.060*	−0.079
Amygdala (CBF)	0.049	0.050	0.060	0.060
A Insula (volume)	−0.044*	−0.073*	−0.076*	−0.101
A Insula (CBF)	0.043	0.029	0.032	0.026
A Insula (task activation)	−0.078	−0.085	−0.092	−0.101
Hippocampus (volume)	−0.044*	−0.088	−0.082*	−0.099
Hippocampus (CBF)	0.061	0.055	0.057	0.055
Cingulate Gyrus (FA)	−0.038	−0.033	−0.032	−0.027
Sup Longitudinal Fasciculus (FA)	−0.062	−0.065	−0.071	−0.080
Uncinate Fasciculus (FA)	−0.093	−0.097	−0.101	−0.102

Note. Significant effects bolded; *indicates that the correlation is significantly different from the correlation of the full form (rightmost column); †indicates that the correlation is significantly different from the correlation with the full form, but neither correlation is significantly different from 0; SES = socioeconomic status; CBF = cerebral blood flow; A = anterior; FA = fractional anisotropy; Sup = superior.

for three of the five demographic-experiential variables. Specifically, the full form correlated more strongly than all three adaptive forms with mother's education, neighborhood SES, and trauma exposure. For age and neighborhood crime, there were no significant differences between the full and adaptive forms.

For correlations with brain parameters, regional volume showed significant correlations with OMI scores for the amygdala (full form only), anterior insula (20-item, 40-item, and full form), and hippocampus (20-item, 40-item, and full form). Regional CBF did not show any significant relationships with OMI scores. Regional FA showed significant correlations with OMI scores for the superior longitudinal fasciculus (40-item and full form) and uncinate fasciculus (all forms). Finally, task activation of the anterior insula showed significant relationships with the OMI for all forms. For the brain variables listed in Table 2, no significant differences were found between correlations with the full form versus any of the three adaptive forms. Overall, correlations of OMI with brain parameters are in the expected directions, with all showing negative correlations except for CBF.

3. Discussion

The purpose of the present study was to construct, calibrate, and validate a CAT screener for overall mental illness. The results were generally encouraging. All relationships of CAT-OMI scores with demographic characteristics, common mental disorders, and brain measures were in the expected direction. Specifically, higher OMI was associated with, 1) lower socio-economic status and mother's education, 2) higher neighborhood crime (not significant) and trauma exposure, 3) higher probability of all mental disorders measured here, 4) lower volume in relevant brain regions, 5) lower FA in relevant brain regions, and 6) lower activation of the anterior insula during task-fMRI. Importantly, most of these relationships were almost as strong for the adaptive versions as for the full versions of the scale. Note that the work done here builds on previous work aimed at rapidly (and sometimes adaptively) assessing clinical outcomes, often at intake. Perhaps the largest example is the Patient-Reported Outcomes Measurement Information System (PROMIS) (Fries et al., 2005). Another

representative example of smaller scales designed for quick screening is the Brief Risk-Resilience Index for SCreening (BRISC) (Williams et al., 2012).

The “target user” of the CAT-OMI would be a researcher who wishes to rapidly assess a participant's overall level of psychopathology. This test would clearly be inappropriate if the topic of interest were a specific disorder (e.g. depression); however, many studies even peripherally related to psychopathology would benefit from a measure of psychopathology as a covariate. In such cases, where psychopathology is not a focus but should not be ignored entirely, conventional scales (e.g. depression scales) might be too narrow, while a combination of conventional scales will almost certainly be too time-consuming. The adaptive OMI assessment meets potential needs of non-clinical researchers insofar as it is broad enough to cover most mental disorders, but efficient enough to be of minimal time burden (3–10 min, depending on version).

Another potential use of the OMI assessment is as a screener in recruitment of participants. A benefit of CAT is that, after each response, it estimates a standard error of measurement (SEM), which can itself be used to estimate confidence intervals around the test-taker's interim trait estimate. Thus, one can set the CAT to stop when the 95% confidence interval around a trait estimate no longer contains a particular target trait level, such as the cutoff level that would classify someone as “ill” or not; see “Clinical-Decision Adaptive Testing” in Waller and Reise (1989, p. 1054). This could allow recruiters to decide, with very few items, whether a participant meets psychopathology-based exclusion criteria.

It is unclear whether the adaptive OMI assessment would be useful to clinicians. On one hand, clinicians usually prefer to obtain a diagnosis before deciding on treatment (if any), and knowing a patient's OMI would likely be insufficiently informative. On the other hand, there are scenarios in which the efficiency of the OMI assessment might be optimal. For example, in disaster zones or very low-income areas with scarce resources (where the professional to patient ratio is very small), the only option might be to assess people very roughly but rapidly to help prioritize professional attention.

Overall, results support the feasibility and utility of the CAT-OMI for research purposes and, possibly, clinical purposes; however, the present study has several important limitations. First, there are several forms of replication required to bolster the current findings. The 114 items used in the current analyses were extracted from a larger K-SADS based interview and therefore it is possible that participants may respond to the questions in that context differently than they do within the truly adaptive framework. Thus, the adaptive form still needs to be validated using concurrent assessment with the full GOASSESS. Second, the age range is relatively young (11–21 years), and given differences in manifestations of psychopathology in children, adults, and the elderly, replication in samples with wider and/or different age ranges will be important. Relatedly, although the CAT-OMI includes symptoms from a wide range of common mental illnesses, some important ones are not covered—e.g. autism spectrum disorders, substance abuse, eating disorders, dissociative disorders, and some others. Fourth, for the neuroimaging validation, while we undertook great effort to specify *a priori* brain regions based on existing literature, there are undoubtedly many more brain-psychopathology relationships that we simply failed to find. Thus, validity results related to brain regions must be interpreted with caution, and their specificity of effect cannot be strongly claimed. Fifth, the association between OMI and SES (including with maternal education) does not add much to the validity evidence because it is so non-specific (SES correlates with many things). The finding that the full form correlated more highly with SES than did any of the adaptive forms could therefore be interpreted in a number of ways and requires further study. Sixth, related to the problem of age range mentioned above, there are known sex differences in brain development that have been unaccounted for here. This, combined with the development-rich age range, increases the likelihood of both Type 1 and Type 2 error.

Finally, measurement invariance (differential item functioning) has not been explored in the present study. Future research should explore the possibility that the measurement model is different across groups (e.g. by race, sex, age group, disease status, socioeconomic class, etc.).

Despite the above limitations, the CAT-OMI screener shows promise and may facilitate the rapid and accurate measurement of overall psychopathology in large-scale studies and in clinical practice. The CAT-OMI is brief—i.e. as few as 10 items, depending on the desired level of measurement precision—and most importantly, it is publicly available and customizable, free of charge. It is thus poised to become part of the repertoire of common tests available for large-scale studies and with minimal resource cost.

Conflict of interest (COI)

No conflict of interest.

Acknowledgements

This work was supported by the Lifespan Brain Institute (LiBI); National Institute of Mental Health (NIMH) grants MH089983, MH019112, MH096891, MH107235, and MH107703; and the Dowshen Program for Neuroscience.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jpsychires.2019.05.028>.

References

- Alvarez, R.P., Kirlic, N., Misaki, M., Bodurka, J., Rhudy, J.L., Paulus, M.P., Drevets, W.C., 2015. Increased anterior insula activity in anxious individuals is linked to diminished perceived control. *Transl. Psychiatry* 5, e591.
- American Psychiatric Association, 2011. In: *Diagnostic and Statistical Manual of Mental Disorders*, fourth ed. Text Revision (DSM-IV-TR).
- Bennett, M.R., Farnell, L., Gibson, W.G., 2013. Fiber pathway pathology, synapse loss and decline of cortical function in schizophrenia. *PLoS One* 8, e60518.
- Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459.
- Bogerts, B., Lieberman, J.A., Ashtari, M., Bilder, R.M., Degreaf, G., Lerner, G., Johns, C., Masiar, S., 1993. Hippocampus-amygdala volumes and psychopathology in chronic schizophrenia. *Biol. Psychiatry* 33, 236–246.
- Bois, C., Levita, L., Ripp, I., Owens, D.C.G., Johnstone, E.C., Whalley, H.C., Lawrie, S.M., 2015. Hippocampal, amygdala and nucleus accumbens volume in first-episode schizophrenia patients and individuals at high familial risk: a cross-sectional comparison. *Schizophr. Res.* 165, 45–51.
- Bremner, J.D., Narayan, M., Anderson, E.R., Staib, L.H., Miller, H.L., Charney, D.S., 2000. Hippocampal volume reduction in major depression. *Am. J. Psychiatry* 157, 115–118.
- Calkins, M.E., Merikangas, K.R., Moore, T.M., Burstein, M., Behr, M.A., Satterthwaite, T.D., Ruparel, K., Wolf, D.H., Roalf, D.R., Mentch, F.D., Qiu, H., Chiavacci, R., Connolly, J.J., Sleiman, P.M.A., Gur, R.C., Hakonarson, H., Gur, R.E., 2015. The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. *JCPP (J. Child Psychol. Psychiatry)* 56, 1356–1369.
- Calkins, M.E., Moore, T.M., Merikangas, K.R., Burstein, M., Satterthwaite, T.D., Bilker, W.B., Ruparel, K., Chiavacci, R., Wolf, D.H., Mentch, F., Qiu, H., Connolly, J.J., Sleiman, P.A., Hakonarson, H., Gur, R.C., Gur, R.E., 2014. The psychosis spectrum in a young U.S. community sample: findings from the Philadelphia Neurodevelopmental Cohort. *World Psychiatr.* 13, 296–305.
- Caspi, A., Houts, R.M., Belsky, D.W., Goldman-Mellor, S., Harrington, H., Israel, S., Meier, M.H., Ramrakha, S., Shalev, I., Poulton, R., Moffitt, T.E., 2014. The p Factor. *Clinical Psychological Science* 2, 119–137.
- Castro-Fornieles, J., Caldú, X., Andró-Pépi, S., Lázaro, L., Bargalló, N., Falcón, C., Plana, M.T., Junqué, C., 2010. A cross-sectional and follow-up functional MRI study with a working memory task in adolescent anorexia nervosa. *Neuropsychologia* 48, 4111–4116.
- Chalmers, R.P., 2012. Mirt: a multidimensional item response theory package for the R Environment. *J. Stat. Softw.* 48.
- Choi, S.W., 2009. Firestar: computerized adaptive testing simulation program for polytomous item response theory models. *Appl. Psychol. Meas.* 33, 644–645.
- Cullen, K.R., Westlund, M.K., Klimes-Dougan, B., Mueller, B.A., Houry, A., Eberly, L.E., Lim, K.O., 2014. Abnormal amygdala resting-state functional connectivity in adolescent depression. *JAMA Psychiatry* 71, 1138.
- Dibbets, P., Evers, E.A., Hurks, P.P., Bakker, K., Jolles, J., 2010. Differential brain activation patterns in adult attention-deficit hyperactivity disorder (ADHD) associated with task switching. *Neuropsychology* 24, 413.

- Edwards, M.C., Wirth, R.J., Houts, C.R., Bodine, A.J., 2014. Three (or four) factors, four (or three) models. In: Reise, S.P., Revicki, D. (Eds.), *Handbook of Item Response Theory Modeling*. Routledge, New York.
- Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1, 54–75.
- Embretson, S.E., Reise, S.P., 2009. In: *Item Response Theory for Psychologists*, Repr. Erlbaum, Mahwah, NJ [u.a.].
- Erk, S., Meyer-Lindenberg, A., Schmierer, P., Mohnke, S., Grimm, O., Garbusow, M., Haddad, L., Poehland, L., Mühleisen, T.W., Witt, S.H., Tost, H., Kirsch, P., Romanczuk-Seiferth, N., Schott, B.H., Cichon, S., Nöthen, M.M., Rietschel, M., Heinz, A., Walter, H., 2014. Hippocampal and frontolimbic function as intermediate phenotype for psychosis: evidence from healthy relatives and a common risk variant in CACNA1C. *Biol. Psychiatry* 76, 466–475.
- Feinstein, J.S., Buzza, C., Hurlmann, R., Follmer, R.L., Dahdaleh, N.S., Coryell, W.H., Welsh, M.J., Tranel, D., Wemmie, J.A., 2013. Fear and panic in humans with bilateral amygdala damage. *Nat. Neurosci.* 16, 270–272.
- Fries, J.F., Bruce, B., Cella, D., 2005. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clin. Exp. Rheumatol.* 23, S53.
- Fusar-Poli, P., Howes, O.D., Allen, P., Broome, M., Valli, I., Asselin, M.C., Montgomery, A.J., Grasby, P.M., McGuire, P., 2011. Abnormal prefrontal activation directly related to pre-synaptic striatal dopamine dysfunction in people at clinical high risk for psychosis. *Mol. Psychiatry* 16, 67.
- Gibbons, R.D., Bock, R.D., Hedeker, D., Weiss, D.J., Segawa, E., Bhaumik, D.K., Kupfer, D.J., Frank, E., Grochocinski, V.J., Stover, A., 2007. Full-information item bifactor Analysis of graded response data. *Appl. Psychol. Meas.* 31, 4–19.
- Gibbons, R.D., Hedeker, D.R., 1992. Full-information item bi-factor analysis. *Psychometrika* 57, 423–436.
- Han, K., Choi, S., Jung, J., Na, K., Yoon, H., Lee, M., Ham, B., 2013. Cortical thickness, cortical and subcortical volume, and white matter integrity in patients with their first episode of major depression. *J. Affect. Disord.* 155, 42–48.
- Harris, G.J., Jaffin, S.K., Hodge, S.M., Kennedy, D., Caviness, V.S., Marinkovic, K., Papadimitriou, G.M., Makris, N., Oscar-Berman, M., 2008. Frontal white matter and cingulum diffusion tensor imaging deficits in alcoholism. *Alcohol Clin. Exp. Res.* 32, 1001–1013.
- Kaczurkin, A.N., Moore, T.M., Calkins, M.E., Ciric, R., Detre, J.A., Elliott, M.A., Foa, E.B., Garcia de la Garza, A., Roalf, D.R., Rosen, A., Ruparel, K., Shinohara, R.T., Xia, C.H., Wolf, D.H., Gur, R.E., Gur, R.C., Satterthwaite, T.D., 2018. Common and dissociable regional cerebral blood flow differences associate with dimensions of psychopathology across categorical diagnoses. *Mol. Psychiatry*.
- Kessler, R.C., Chiu, W.T., Demler, O., Walters, E.E., 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. *Arch. Gen. Psychiatr.* 62, 617.
- Kitis, O., Ozalay, O., Zengin, E.B., Haznedaroglu, D., Eker, M.C., Yalvac, D., Oguz, K., Coburn, K., Gonul, A.S., 2012. Reduced left uncinate fasciculus fractional anisotropy in deficit schizophrenia but not in non-deficit schizophrenia. *Psychiatr. Clin. Neurosci.* 66, 34–43.
- Kobayashi, H., Nemoto, T., Koshikawa, H., Osono, Y., Yamazawa, R., Murakami, M., Kashima, H., Mizuno, M., 2008. A self-reported instrument for prodromal symptoms of psychosis: testing the clinical validity of the PRIME Screen—revised (PS-R) in a Japanese population. *Schizophr. Res.* 106, 356–362.
- Konrad, A., Dielentheis, T.F., El Masri, D., Bayerl, M., Fehr, C., Gesierich, T., Vucurevic, G., Stoeter, P., Winterer, G., 2010. Disturbed structural connectivity is related to inattention and impulsivity in adult attention deficit hyperactivity disorder. *Eur. J. Neurosci.* 31, 912–919.
- Krueger, R.F., 1999. The structure of common mental disorders. *Arch. Gen. Psychiatr.* 56, 921.
- Krueger, R.F., Caspi, A., Moffitt, T.E., Silva, P.A., 1998. The structure and stability of common mental disorders (DSM-III-R): a longitudinal-epidemiological study. *J. Abnorm. Psychol.* 107, 216–227.
- Kubicki, M., Westin, C., Maier, S.E., Frumin, M., Nestor, P.G., Salisbury, D.F., Kikinis, R., Jolesz, F.A., McCarley, R.W., Shenton, M.E., 2002. Uncinate fasciculus findings in schizophrenia: a magnetic resonance diffusion tensor imaging study. *Am. J. Psychiatry* 159, 813–820.
- Kwon, J.S., 2003. Similarity and disparity of obsessive-compulsive disorder and schizophrenia in MR volumetric abnormalities of the hippocampus-amygdala complex. *J. Neuro. Neurosurg. Psychiatry* 74, 962–964.
- Lord, F.M., 1980. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ.
- Makris, N., Goldstein, J.M., Kennedy, D., Hodge, S.M., Caviness, V.S., Faraone, S.V., Tsuang, M.T., Seidman, L.J., 2006. Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophr. Res.* 83, 155–171.
- Marchand, W.R., Lee, J.N., Thatcher, G.W., Jensen, C., Stewart, D., Dilda, V., Thatcher, J., Creem-Regehr, S.H., 2007. A functional MRI study of a paced motor activation task to evaluate frontal-subcortical circuit function in bipolar depression. *Psychiatr. Res. Neuroimaging* 155, 221–230.
- Merikangas, K.R., Avenevoli, S., Costello, E.J., Koretz, D., Kessler, R.C., 2009. National comorbidity survey replication adolescent supplement (NCS-a): I. Background and measures. *J. Am. Acad. Child Adolesc. Psychiatry* 48, 367–379.
- Merikangas, K.R., He, J., Burstein, M., Swanson, S.A., Avenevoli, S., Cui, L., Benjet, C., Georgiades, K., Swendsen, J., 2010. Lifetime prevalence of mental disorders in US adolescents: results from the national comorbidity survey replication—adolescent supplement (NCS-a). *J. Am. Acad. Child Adolesc. Psychiatry* 49, 980–989.
- Miller, T.J., Cicchetti, D., Markovich, P.J., McGlashan, T.H., Woods, S.W., 2004. The SIPS screen: a brief self-report screen to detect the schizophrenia prodrome. *Schizophr. Res.* 70, 78.
- Miller, T.J., McGlashan, T.H., Rosen, J.L., Cadenhead, K., Ventura, J., McFarlane, W., Perkins, D.O., Pearlson, G.D., Woods, S.W., 2003. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. *Schizophr. Bull.* 29, 703–715.
- Moore, T.M., Martin, I.K., Gur, O.M., Jackson, C.T., Scott, J.C., Calkins, M.E., Ruparel, K., Port, A.M., Nivar, I., Krinsky, H.D., Gur, R.E., Gur, R.C., 2016. Characterizing social environment's association with neurocognition using census and crime data linked to the Philadelphia Neurodevelopmental Cohort. *Psychol. Med.* 46, 599–610.
- Moore, T.M., Calkins, M.E., Reise, S.P., Gur, R.C., Gur, R.E., 2018. Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. *Psychiatr. Res.* 263, 250–256.
- Moore, T.M., Scott, J.C., Reise, S.P., Port, A.M., Jackson, C.T., Ruparel, K., Savitt, A.P., Gur, R.E., Gur, R.C., 2015. Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. *Psychol. Assess.* 27, 955–964.
- Moran, L.V., Tagamets, M.A., Sampath, H., O'Donnell, A., Stein, E.A., Kochunov, P., Hong, L.E., 2013. Disruption of anterior insula modulation of large-scale brain networks in schizophrenia. *Biol. Psychiatry* 74, 467–474.
- Nakamae, T., Narumoto, J., Shibata, K., Matsumoto, R., Kitabayashi, Y., Yoshida, T., Yamada, K., Nishimura, T., Fukui, K., 2008. Alteration of fractional anisotropy and apparent diffusion coefficient in obsessive-compulsive disorder: a diffusion tensor imaging study. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* 32, 1221–1226.
- Oler, J.A., Fox, A.S., Shackman, A.J., Kalin, N.H., 2016. The central nucleus of the amygdala is a critical substrate for individual differences in anxiety. In: Amaral, D.G., Adolphs, R. (Eds.), *Living without an Amygdala*. Guilford, New York.
- Paulus, M.P., Stein, M.B., 2006. An insular view of anxiety. *Biol. Psychiatry* 60, 383–387.
- Penner, J., Ford, K.A., Taylor, R., Schaefer, B., Théberge, J., Neufeld, R.W.J., Osuch, E.A., Menon, R.S., Rajakumar, N., Allman, J.M., Williamson, P.C., 2016. Medial prefrontal and anterior insular connectivity in early schizophrenia and major depressive disorder: a resting functional MRI evaluation of large-scale brain network models. *Front. Hum. Neurosci.* 10.
- Peterson, D.J., Ryan, M., Rimrod, S.L., Cutting, L.E., Denckla, M.B., Kaufmann, W.E., Mahone, E.M., 2011. Increased regional fractional anisotropy in highly screened attention-deficit hyperactivity disorder (ADHD). *J. Child Neurol.* 26, 1296–1302.
- Phan, K.L., Orlichenko, A., Boyd, E., Angstadt, M., Coccaro, E.F., Liberzon, I., Arfanakis, K., 2009. Preliminary evidence of white matter abnormality in the uncinate fasciculus in generalized social anxiety disorder. *Biol. Psychiatry* 66, 691–694.
- Plessen, K.J., Bansal, R., Zhu, H., Whiteman, R., Amat, J., Quackenbush, G.A., Martin, L., Durkin, K., Blair, C., Royal, J., Hugdahl, K., Peterson, B.S., 2006. Hippocampus and amygdala morphology in attention-deficit/hyperactivity disorder. *Arch. Gen. Psychiatr.* 63, 795.
- Rajarethinam, R., DeQuardo, J.R., Miedler, J., Arndt, S., Kirbat, R., Brunberg, J., A. Tandon, R., 2001. Hippocampus and amygdala in schizophrenia: assessment of the relationship of neuroanatomy to psychopathology. *Psychiatr. Res. Neuroimaging* 108, 79–87.
- Reckase, M.D., 2009. Multidimensional item response theory models. *Multidimensional Item Response Theory* 79–112.
- Reise, S.P., Cook, K.F., Moore, T.M., 2015. Evaluating the impact of multidimensionality on unidimensional Item Response Theory model parameters. In: Reise, S.P., Revicki, D. (Eds.), *Handbook of Item Response Theory Modeling*. Routledge, New York, pp. 13–40.
- Reise, S.P., 2012. The rediscovery of bifactor measurement models. *Multivariate Behav. Res.* 47, 667–696.
- Reise, S.P., Moore, T.M., Maydeu-Olivares, A., 2011. Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educ. Psychol. Meas.* 71, 684–711.
- Reise, S.P., Moore, T.M., Haviland, M.G., 2010. Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J. Personal. Assess.* 92, 544–559.
- Roalf, D.R., Moore, T.M., Wolk, D.A., Arnold, S.E., Mechanic-Hamilton, D., Rick, J., Kabadi, S., Ruparel, K., Chen-Plotkin, A., Chahine, L.M., Dahodwala, N.A., Duda, J.E., Weintraub, D.A., Moberg, P.J., 2016a. Defining and validating a short form Montreal Cognitive Assessment (s-MoCA) for use in neurodegenerative disease. *J. Neurol. Neurosurg. Psychiatry* 87, 1303–1310.
- Roalf, D.R., Quarmley, M., Elliott, M.A., Satterthwaite, T.D., Vandekar, S.N., Ruparel, K., Gennatas, E.D., Calkins, M.E., Moore, T.M., Hopson, R., Prabhakaran, K., Jackson, C.T., Verma, R., Hakonarson, H., Gur, R.C., Gur, R.E., 2016b. The impact of quality assurance assessment on diffusion tensor imaging outcomes in a large-scale population-based cohort. *Neuroimage* 125, 903–919.
- Rosen, A.F.G., Roalf, D.R., Ruparel, K., Blake, J., Seelaus, K., Villa, L.P., Ciric, R., Cook, P.A., Davatzikos, C., Elliott, M.A., Garcia de la Garza, L.G., Gennatas, E.D., Quarmley, M., Schmitt, J.E., Shinohara, R.T., Tisdall, M.D., Craddock, R.C., Gur, R.E., Gur, R.C., Satterthwaite, T.D., 2018. Quantitative assessment of structural image quality. *Neuroimage* 169, 407–418.
- Sarkar, S., Craig, M.C., Catani, M., Dell'Acqua, F., Fahy, T., Deeley, Q., Murphy, D.G.M., 2012. Frontotemporal white-matter microstructural abnormalities in adolescents with conduct disorder: a diffusion tensor imaging study. *Psychol. Med.* 43, 401–411.
- Satterthwaite, T.D., Shinohara, R.T., Wolf, D.H., Hopson, R.D., Elliott, M.A., Vandekar, S.N., Ruparel, K., Calkins, M.E., Roalf, D.R., Gennatas, E.D., Jackson, C., Erus, G., Prabhakaran, K., Davatzikos, C., Detre, J.A., Hakonarson, H., Gur, R.C., Gur, R.E., 2014a. Impact of puberty on the evolution of cerebral perfusion during adolescence. *Proc. Natl. Acad. Sci. Unit. States Am.* 111, 8643–8648.
- Satterthwaite, T.D., Wolf, D.H., Erus, G., Ruparel, K., Elliott, M.A., Gennatas, E.D., Hopson, R., Jackson, C., Prabhakaran, K., Bilker, W.B., Calkins, M.E., Loughhead, J., Smith, A., Roalf, D.R., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R.C., Gur, R.E., 2013. Functional maturation of the executive system during adolescence. *J.*

- Neurosci. 33, 16249–16261.
- Satterthwaite, T.D., Connolly, J.J., Ruparel, K., Calkins, M.E., Jackson, C., Elliott, M.A., Roalf, D.R., Hopson, R., Prabhakaran, K., Behr, M., Qiu, H., Mentch, F.D., Chiavacci, R., Sleiman, P.M.A., Gur, R.E., Gur, R.C., Hakonarson, H., 2016a. The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage* 124, 1115–1119.
- Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Loughhead, J., Prabhakaran, K., Calkins, M.E., Hopson, R., Jackson, C., Keefe, J., Riley, M., Mentch, F.D., Sleiman, P., Verma, R., Davatzikos, C., Hakonarson, H., Gur, R.E., Gur, R.C., 2016b. Neuroimaging of the Philadelphia neurodevelopmental cohort. *Neuroimage* 86, 544–553.
- Satterthwaite, T.D., Wolf, D.H., Calkins, M.E., Vandekar, S.N., Erus, G., Ruparel, K., Roalf, D.R., Linn, K.A., Elliott, M.A., Moore, T.M., Hakonarson, H., Shinohara, R.T., Davatzikos, C., Gur, R.C., Gur, R.E., 2016b. Structural brain abnormalities in youth with psychosis spectrum symptoms. *JAMA Psychiatry* 73, 515–524.
- Scalise, K., Allen, D.D., 2015. Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *Br. J. Math. Stat. Psychol.* 68, 478–496.
- Senatorov, V.V., Damadzic, R., Mann, C.L., Schwandt, M.L., George, D.T., Hommer, D.W., Heilig, M., Momenan, R., 2014. Reduced anterior insula, enlarged amygdala in alcoholism and associated depleted von Economo neurons. *Brain* 138, 69–79.
- Shaffer, D., 1983. A children's global assessment scale (CGAS). *Arch. Gen. Psychiatr.* 40, 1228.
- Shanmugan, S., Wolf, D.H., Calkins, M.E., Moore, T.M., Ruparel, K., Hopson, R.D., Vandekar, S.N., Roalf, D.R., Elliott, M.A., Jackson, C., Gennatas, E.D., Leibenluft, E., Pine, D.S., Shinohara, R.T., Hakonarson, H., Gur, R.C., Gur, R.E., Satterthwaite, T.D., 2016. Common and dissociable mechanisms of executive System dysfunction across psychiatric disorders in youth. *Am. J. Psychiatry* 173, 517–526.
- Siegle, G.J., Thompson, W., Carter, C.S., Steinhauer, S.R., Thase, M.E., 2007. Increased amygdala and decreased dorsolateral prefrontal BOLD responses in unipolar depression: related and independent features. *Biol. Psychiatry* 61, 198–209.
- Smith, R., Fadok, R.A., Purcell, M., Liu, S., Stonnington, C., Spetzler, R.F., Baxter, L.C., 2011. Localizing sadness activation within the subgenual cingulate in individuals: a novel functional MRI paradigm for detecting individual differences in the neural circuitry underlying depression. *Brain imaging and behavior* 5, 229–239.
- Sterzer, P., Stadler, C., Poustka, F., Kleinschmidt, A., 2007. A structural neural deficit in adolescents with conduct disorder and its association with lack of empathy. *Neuroimage* 37, 335–342.
- Svatkova, A., Nestrasil, I., Rudser, K., Goldenring Fine, J., Bledsoe, J., Semrud-Clikeman, M., 2016. Unique white matter microstructural patterns in ADHD presentations-a diffusion tensor imaging study. *Hum. Brain Mapp.* 37, 3323–3336.
- Uddin, L.Q., Menon, V., 2009. The anterior insula in autism: under-connected and under-examined. *Neurosci. Biobehav. Rev.* 33, 1198–1203.
- Vandekar, S.N., Shinohara, R.T., Raznahan, A., Roalf, D.R., Ross, M., DeLeo, N., Ruparel, K., Verma, R., Wolf, D.H., Gur, R.C., Gur, R.E., Satterthwaite, T.D., 2015. Topologically dissociable patterns of development of the human cerebral cortex. *J. Neurosci.* 35, 599–609.
- Videbech, P., Ravnkilde, B., 2004. Hippocampal volume and depression: a meta-analysis of MRI studies. *Am. J. Psychiatry* 161, 1957–1966.
- Von Der Heide, R.J., Skipper, L.M., Klobusicky, E., Olson, I.R., 2013. Dissecting the uncinate fasciculus: disorders, controversies and a hypothesis. *Brain : J. Neurol.* 136, 1692–1707.
- Wainer, H., Dorans, N.J., 2000. In: *Computerized Adaptive Testing*, second ed. Routledge Ltd, Mahwah.
- Waller, N.G., Reise, S.P., 1989. Computerized adaptive personality assessment: an illustration with the Absorption scale. *J. Personal. Soc. Psychol.* 57, 1051–1058.
- Wang, F., Sun, Z., Cui, L., Du, X., Wang, X., Zhang, H., Cong, Z., Hong, N., Zhang, D., 2004. Anterior cingulum abnormalities in male patients with schizophrenia determined through diffusion tensor imaging. *Am. J. Psychiatry* 161, 573–575.
- Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. *TPAMI* 35, 611–623.
- Wang, H., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation. *Front. Neuroinf.* 7, 27.
- Weiss, D.J., Gibbons, R.D., 2007. Computerized adaptive testing with the bifactor model. In: Weiss, D.J. (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. IACAT.
- Weiss, D.J., Kingsbury, G.G., 1984. Application of computerized adaptive testing to educational problems. *J. Educ. Meas.* 21, 361–375.
- Williams, L.M., Cooper, N.J., Wisniewski, S.R., Gatt, J.M., Koslow, S.H., Kulkarni, J., DeVarney, S., Gordon, E., Rush, A.J., 2012. Sensitivity, specificity, and predictive power of the “Brief Risk-resilience Index for SC reining,” a brief pan-diagnostic web screen for emotional health. *Brain and behavior* 2, 576–589.
- Zarei, M., Mataix-Cols, D., Heyman, I., Hough, M., Doherty, J., Burge, L., Winmill, L., Nijhawan, S., Matthews, P.M., James, A., 2011. Changes in gray matter volume and white matter microstructure in adolescents with obsessive-compulsive disorder. *Biol. Psychiatry* 70, 1083–1090.
- Zou, K., Huang, X., Li, T., Gong, Q., Li, Z., Ou-yang, L., Deng, W., Chen, Q., Li, C., Ding, Y., Sun, X., 2008. Alterations of white matter integrity in adults with major depressive disorder: a magnetic resonance imaging study. *J. Psychiatry Neurosci.* : JPN 33, 525.