Psychological Assessment
2017, Vol. 29, No. 7, 868–880

1040-3590/17/$12.00 http://dx.doi.org/10.1037/pas0000384

Efficiently Measuring Dimensions of the Externalizing Spectrum Model: Development of the Externalizing Spectrum Inventory-Computerized Adaptive Test (ESI-CAT)

Matthew Sunderland and Tim Slade
University of New South Wales Australia

Robert F. Krueger
University of Minnesota

Kristian E. Markon
University of Iowa

Christopher J. Patrick
Florida State University

Mark D. Kramer
Minneapolis Veterans Affairs Health Care System and Center for Chronic Disease Outcomes Research, Minneapolis, Minnesota

The development of the Externalizing Spectrum Inventory (ESI) was motivated by the need to comprehensively assess the interrelated nature of externalizing psychopathology and personality using an empirically driven framework. The ESI measures 23 theoretically distinct yet related unidimensional facets of externalizing, which are structured under 3 superordinate factors representing general externalizing, callous aggression, and substance abuse. One limitation of the ESI is its length at 415 items. To facilitate the use of the ESI in busy clinical and research settings, the current study sought to examine the efficiency and accuracy of a computerized adaptive version of the ESI. Data were collected over 3 waves and totaled 1,787 participants recruited from undergraduate psychology courses as well as male and female state prisons. A series of 6 algorithms with different termination rules were simulated to determine the efficiency and accuracy of each test under 3 different assumed distributions. Scores generated using an optimal adaptive algorithm evidenced high correlations ($r > .9$) with scores generated using the full ESI, brief ESI item-based factor scales, and the 23 facet scales. The adaptive algorithms for each facet administered a combined average of 115 items, a 72% decrease in comparison to the full ESI. Similarly, scores on the item-based factor scales of the ESI-brief form (57 items) were generated using on average of 17 items, a 70% decrease. The current study successfully demonstrates that an adaptive algorithm can generate similar scores for the ESI and the 3 item-based factor scales using a fraction of the total item pool.

Keywords: externalizing, disinhibition, computerized adaptive test, item response theory, assessment

Supplemental materials: http://dx.doi.org/10.1037/pas0000384.supp

Disorders of impulse control, such as conduct disorder, antisocial personality disorder, substance use disorders, and their constituent symptoms, have a tendency to covary more often than chance. Analyses of *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) symptom- and syndrome-level data in adult populations have consistently identified a single coherent liability dimension that links these putatively distinct disorders together as a means of explaining their high comorbidity and distinguishes them from other forms of psychopathology (Krueger, 1999; Lahey et al., 2008; Slade & Watson, 2006; Vollebergh et al., 2001; Wolf et al., 1988; Wright et al., 2013). This psychopathological construct, labeled externalizing, has been shown to have a strong genetic basis and has been linked with a range of disinhibitory personality traits such as impulsivity and aggression (Hicks, Krueger, Iacono, McGue, & Patrick, 2004; Krueger & South, 2009). Indeed, similar constructs have been observed in work on the structure of psychopathology in children dating back more than 50 years (Achenbach & Edelbrock, 1984). As such, these findings have led researchers to postulate that significant advances in the literature might be achieved if these various externalizing problems and disinhibitory personality traits were brought together under a single integrative and hierarchical framework (Krueger, Markon, Patrick, & Iacono, 2005). Previous studies have investi-

This article was published Online First November 14, 2016.

Matthew Sunderland and Tim Slade, NHMRC Centre for Research Excellence in Mental Health and Substance Use, National Drug and Alcohol Research Centre, University of New South Wales; Robert F. Krueger, Department of Psychology, University of Minnesota; Kristian E. Markon, Department of Psychology, University of Iowa; Christopher J. Patrick, Department of Psychology, Florida State University; Mark D. Kramer, Minneapolis Veterans Affairs Health Care System and Center for Chronic Disease Outcomes Research, Minneapolis, Minnesota.

Correspondence concerning this article should be addressed to Matthew Sunderland, National Drug and Alcohol Research Centre, Building R1, Randwick Campus, University of New South Wales, Sydney, NSW, Australia. E-mail: matthews@unsw.edu.au

gated dimensional models of disinhibitory personality and its disorders and developed scales based on the Minnesota Multiphasic Personality Inventory (MMPI)-2, most notably the Aggressiveness and Constraint factors of the Personality Psychopathology Five (Harkness, McNulty, & Ben-Porath, 1995). The ESI was more recently developed by Krueger, Markon, Patrick, Benning, and Kramer (2007) to expand on the measurement of various externalizing problems and disinhibitory personality traits and serve as a common empirically driven framework, thus representing a comprehensive model of the externalizing spectrum.

The ESI and Brief Forms

The ESI was developed empirically using a bottom-up process, meaning that items were constructed to target specific elements of externalizing identified from the existing literature. The resultant inventory contained 415 items that targeted 23 unidimensional facets representing theoretically distinct yet related externalizing constructs. Jointly, these 23 facet scales were then shown to exhibit a bifactor or hierarchical structure whereby the variance across all facets can be explained using three orthogonal superordinate factors: a single externalizing (disinhibition) factor and two additional factors representing the remaining variance that is shared among certain facets. After externalizing, the second factor represents specific variance associated with callous aggression facets such as aggression, (lack of) empathy, blame externalizing, fraud, (dis)honesty, dependability, impatient urgency, rebelliousness, boredom proneness, and excitement seeking. The third factor represents specific variance associated with substance abuse facets such as alcohol use and problems, marijuana use and problems, and drug use and problems. In short, a significant benefit of developing the ESI in this manner was to integrate a coherent and empirically based conceptual model of externalizing with a corresponding self-report assessment (Krueger et al., 2007).

Since publication of the initial ESI development study, validation of the instrument has relied heavily on use of various brief forms of the 415-item version targeting the three superordinate factors. Hall et al. (2007) demonstrated that overall scores on a 100-item version of the ESI (measuring externalizing in general) were related to the higher order factors of negative emotionality and constraint measured using the Multidimensional Personality Questionnaire (MPQ). Similarly, higher scores on externalizing were related to a higher incidence of rule-breaking behaviors, alcohol dependence, and drug abuse as well as reduced amplitude of the error-related negativity, a reliable neurophysiological index of the ability to self-monitor ongoing behavior for errors or inappropriate actions. Using the same 100-item version, Blonigen et al. (2011) reported that externalizing scores were significantly and negatively correlated with scores on a measure of integrity. Finally, Venables and Patrick (2012) administered an extended 159-item version as a means of measuring the three superordinate factors directly and to examine their relationship with a range of *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) symptoms of externalizing disorders, personality, and psychopathy in a sample of incarcerated adults. They found evidence of convergent and discriminant validity of the superordinate ESI factors. Specifically, the general factor was related to child and adult symptoms of antisocial personality disorder and substance dependence, antisocial deviance features of psychopathy, and scores on the constraint and negative emotionality dimensions of personality. Additionally, the substance abuse factor predicted unique variance in the symptoms of substance dependence over and above the externalizing factor whereas the callous aggression factor was specifically related to aggression symptoms of antisocial personality disorder and the affective-interpersonal components of psychopathy. Taking all of these results together provides compelling evidence for validity of the superordinate factors measured by the ESI in relation to clinically relevant diagnostic criteria and personality traits.

One particular disadvantage that has the potential to limit the utility of the ESI is the sheer length of administration at 415 items. In response to this need, Patrick, Kramer, Krueger, & Markon, (2013) developed a brief form of the inventory (the ESI-BF) that provides coverage of each of the 23 facets and effectively indexes the three superordinate factors. The ESI-BF consists of items selected from the full length instrument to form short facet scales that: (a) faithfully reflected the content of each full length scale, (b) demonstrated effective measurement of the construct associated with each scale, and (c) functioned in a manner similar to the full scale within the overall externalizing measurement model. The resultant scale included 160 items, with each of the facets measured using three to 11 items, which maintained high internal consistency, replicated the structure of the full ESI, and demonstrated similar validity in relation to the MPQ.

In addition, Patrick, Kramer, et al. (2013) developed three item-based factor scales of modest length embedded in the brief form that indexed the three superordinate factors directly, rather than indirectly, through the 23 facet scale scores. These scales were developed by first selecting items from the ESI-BF facet scales that showed robust loadings on target superordinate factors of the full-form ESI model. Candidate items from each facet scale were then selected that exhibited robust and selective associations with scores on the target factor and effective item response theory (IRT) parameters. This process resulted in three unidimensional scales that effectively target general externalizing or disinhibition (20 items), callous aggression (19 items), and substance abuse (18 items). Patrick, Kramer, and colleagues (2013) concluded that the Disinhibition scale could be considered a measure of general externalizing liability free from criterion contamination related to substance abuse or aggressive behavior given that it contains no items that measure alcohol or drug use or aggression. They also concluded that the Callous Aggression scale appears to reflect an aggressive-dominant interpersonal style given strong links to trait aggression and robust relationships with social potency, harm avoidance, and traditionalism measured by the MPQ. Finally, they demonstrated that the Substance Abuse scale reflects the tendency toward alcohol and substance abuse for reasons of experience-seeking rather than due to a lack of behavioral or emotional control (i.e., reasons distinct from general disinhibition) given associations with traits reflecting stimulation-seeking and nonconformity measured by the MPQ. Thus, in combination, these brief scores represent an efficient manner for studies to index general externalizing liability as well as unique callous aggression and substance abuse expressions of externalizing.

Computerized Adaptive Tests (CATs) in Psychopathology and Personality

While these brief scales offer a highly useful, readily amenable approach to administration of the ESI in time-poor settings, they possess disadvantages that are shared with all brief forms that are operationalized in a static fashion, meaning that the same fixed number of items is administered to all individuals regardless of their true underlying score. An alternative to brief form inventories is provided by CATs, which make use of IRT and computerized algorithms to tailor the administration of items from a large pool to each individual person (Embretson & Reise, 2000; Meijer & Nering, 1999). The tailoring seeks to triangulate individual scores associated with a precalibrated measurement model until a desired level of precision is met (Bjorner, Chang, Thissen, & Reeve, 2007). In other words, the CAT algorithm only selects items that maximize information about the person's likely score after each response and terminates once meaningful gains in information can no longer be attained by administering more items. As such, this can result in individual scores across the continuum of severity that are comparable in terms of precision to the full item bank and generally utilize fewer items than a static brief form (Choi, Reise, Pilkonis, Hays, & Cella, 2010). Indeed, previous studies have demonstrated that CATs can feasibly measure a range of psychopathological constructs, such as depression, anxiety, addiction, emotional distress, anger, and suicidal behavior with an acceptable rate of precision using only a fraction of the total item pool (De Beurs, de Vries, de Groot, de Keijser, & Kerkhof, 2014; Fliege et al., 2005; Kirisci et al., 2012; Pilkonis et al., 2011; Walter et al., 2007). A study by Choi and colleagues (2010) also demonstrated that a CAT version for depression was able to provide better precision across the full continuum of severity using a similar number of items in comparison to a static brief form. The primary explanation for their finding was that static forms contain fewer items that optimize information at each specific point on the continuum given they are required to optimize information across a larger area of severity at once.

In the field of personality testing, Simms and Clark (2005) developed and validated a CAT version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). They found the CAT version yielded significant item savings (60% increase in efficiency) with similar descriptive statistics, test-retest stability, internal factor structure, and convergent and discriminant validity pattern in comparison with the full-scale SNAP. The SNAP-CAT also demonstrated greater item savings than those found for non-IRT CAT applications in personality testing, such as the countdown method for the MMPI-2 (Roper, Ben-Porath, & Butcher, 1995). Similarly, Reise and Henson (2000) demonstrated that an IRT-based adaptive version of the NEO Personality Inventory—Revised (NEO PI-R) was able to replicate the full-scale scores using 50% of the items for each scale (four items instead of eight) with correlations >0.91. However, they also found that there was little variability in the four items that were administered by the CAT to each of the respondents for 23 out of 30 facets. They concluded that the CAT algorithm, despite providing greater efficiency, was not required to administer the NEO PI-R, given that similar results in efficiency and precision could have been obtained by administering a static version that comprised the four best items in terms of psychometric information. Indeed, studies have shown that the relative efficiency, precision, and variation in items administered (exposure rate) can differ across item banks from different scales depending on the quality of the item parameters and the number of items providing information at targeted severity levels (Sahin & Weiss, 2015). Generally, scales with larger item banks provide greater flexibility for the algorithms to select a variety of items that target a broad range of severity.

The Current Study

It might be possible that additional gains in efficiency are achievable without a substantial loss of precision in relation to the full ESI, the item-based factor scales of the ESI-BF, and the individual facet scales of the ESI-BF by applying an IRT-based computerized adaptive algorithm. The ESI facet scales were developed under the primary assumptions of IRT, therefore making the application of IRT-based CAT methods well suited for the current purpose. However, some of the ESI facet scales are limited in terms of the number of items and therefore an adaptive algorithm might be restricted in terms of the variability in items presented and the overall efficiency obtained. Moreover, when developing a new CAT for an existing instrument, the choice of various control parameters (e.g., termination criteria) can influence the balance between efficiency and precision of the final scores (Thompson & Weiss, 2011). As such, simulations are required to provide some indication of whether the added complexity associated with CAT administration might be justifiable with regard to gains in efficiency in comparison to simpler static brief forms. Currently, there are no simulation studies that have examined the performance of various adaptive algorithms in relation to the measurement of the externalizing spectra.

The current study examined the gains in efficiency and loss of precision associated with a computerized adaptive version of the three item-based factor scales of the ESI-BF as well as the 23 unidimensional facet scales of the full ESI in a series of Monte Carlo and real data simulations. The three item-based factor scales were selected as the primary focus of this paper, given that the broad factors of the externalizing spectrum model have received increasing attention and interest in the literature. Multiple studies have examined the links between these broad psychopathological constructs and various neurobiological, cognitive, physiological, and genetic correlates (Dick, 2007; Hall, Bernat, & Patrick, 2007; Nelson, Strickland, Krueger, Arbisi, & Patrick, 2016; Patrick & Drislane, 2015; Patrick, Durbin, & Moser, 2012; Patrick, Venables, et al., 2013; Young et al., 2009). This is due, in part, to calls from the National Institute of Mental Health Research Domain Criteria initiative to focus research efforts on dimensional and biologically meaningful constructs of clinically relevant phenomena, such as externalizing (or cognitive control as it is referred to in the research domain criteria framework; Cuthbert & Kozak, 2013; Insel et al., 2010). That being said, greater information regarding the individual profiles of externalizing, callous aggression, and substance abuse might be gleaned from a thorough and rapid investigation of all 23 facets that comprise the superordinate dimensions. As such, simulations of an adaptive algorithm for each of the 23 facets sought to investigate the costs in precision versus the benefits in efficiency.

Method

Sample

The current study utilized data collected by Krueger et al. (2007) as part of the original development study of the ESI. Participants were recruited from undergraduate psychology courses as well as male and female minimum/medium security state prisons. Written informed consent was obtained from all participants prior to administration of the questionnaires and verbal confirmation was given regarding confidentiality by research staff. Student participants were compensated with extra credit toward their psychology course grade or a payment of $10, whereas all prison participants were paid $10. Given the length and iterative development of the original ESI, data were collected in three separate waves of nonoverlapping participants. The first wave of data collection comprised 289 students and 286 prisoners, the second wave comprised 299 students and 314 prisoners, and the third wave comprised 283 students and 316 prisoners. Approximately 22 participants were removed prior to the analysis due to an invalid response pattern. There were 1,787 participants (49% male) in the combined sample with a mean age of 26.8 years (SD = 9.4, range = 18–63). For additional details regarding the sample, see Krueger et al. (2007).

ESI Item Pool

The original development sample of the ESI was used to obtain data from the 415 items that formed the item pools for each of the 23 facet scales (ranging between 9 and 31 items) and the 57 items that formed the item pool for the brief item-based factor scales. Each wave of data collection from the original development study included additional items so that participants in the third wave were administered all 415 items of the ESI. In contrast, the first and second waves contained a number of missing responses to the final item set. Responses for individuals from prior waves to new items included in the second and third waves were treated as missing in order to retain the maximum amount of information relevant to the total number of items. The missing data were treated in all subsequent analyses using full information missing data analytic techniques, which have been recommended for use when data are missing by design (Graham, Hoffer, & MacKinnon, 1996). The percentage of missing data for each item is provided in supplementary Table S1. As detailed further in Krueger et al. (2007), the items of the ESI were developed to target several thematically distinct but related constructs that represent the full range of externalizing behaviors identified in a detailed review of the literature. The items were rated on a 4-point response scale (i.e., 0 = *false*, 1 = *somewhat false*, 2 = *somewhat true*, 3 = *true*).

The unidimensional model was applied to the data for each facet, which assumes that variance between the items within each facet can be adequately explained by a single common latent variable. This variable may be interpreted as an indicator of underlying severity with higher scores representing greater severity. Fit statistics pertaining to the unidimensional model for each of the item-based factor scales and the 23 facets scales were generated using confirmatory factor analysis with weighted least squares mean and variance adjusted estimation. Model fit was determined using the comparative fit index (CFI; Bentler, 1990), Tucker-Lewis fit index (TLI; Tucker & Lewis, 1973), and root-mean-square error of approximation (RMSEA; Browne & Cudeck, 1993). Values on the CFI and TLI $\geq$0.90 indicate adequate fit and scores $\geq$0.95 represent good fit, whereas values on the RMSEA $\leq$0.08 represent adequate fit and values $>$0.10 indicate poor fit (Browne & Cudeck, 1993; Byrne, 2012; Hu & Bentler, 1999; MacCallum, Browne, & Sugawara, 1996). Scores for each of the item-based factor scales and the 23 facet scales were calibrated (i.e., estimation of IRT parameters) according to Samejima's graded response model using full information marginal maximum likelihood via Bock-Aitkin expectation maximization algorithm (Bock & Aitkin, 1981). Item parameters were estimated using FlexMIRT version 2.0 (Cai, 2013).

CAT Simulations

The adaptive algorithm was examined using a series of simulations. The IRT parameters estimated in the calibration phase were utilized to generate item responses for each of the ESI-BF item-based factor scales and the 23 facet scales for 1,000 data points under each assumed population model. The efficiency and precision of potential CAT algorithms for each factor and facet scale were examined under three assumed population distributions: (a) the normal distribution, (b) the uniform distribution, and (c) the empirical distribution of each factor estimated using the full development sample. The empirical distributions were generated using empirical histograms or the normalized accumulated posterior densities for all response patterns at each quadrature node (Woods, 2007). For each population distribution, the CAT algorithm would commence by selecting the single best item that maximized Fisher information at a theta estimate of zero for all respondents (Meijer & Nering, 1999). Using the simulated response to this particular item, the algorithm estimates a preliminary theta score and selects the next best item that maximizes the Fisher information evaluated at the updated theta estimates (Meijer & Nering, 1999). The minimum number of items to be administered by each of the CAT algorithms was specified at two items.

Several CAT algorithms were simulated to test different combinations of termination rules based on the standard error of measure for each theta estimate and the relative change in theta score estimates from one item to the next. The first termination rule specified that the CAT would terminate if the standard error dropped below 0.3 or if all items in the scale were administered. The second rule increased the standard error threshold to 0.4, reflecting less precision. Previous research has demonstrated that using standard errors as the sole termination rule can result in a loss of efficiency depending on the nature of the item pool. Specifically, individuals with true theta scores on the severity continuum that are not well indexed by the item pool (i.e., extremely high or extremely low scores) may never generate theta scores with enough precision to meet the required termination rules and therefore the algorithm will needlessly present these individuals with every item. To overcome this effect, additional termination rules were specified so that the algorithm would terminate if the change between preliminary theta scores estimated from one item to the next dropped below 0.01 and 0.05, respectively. The termination rules were designed to work in combination; for example, one possible combination specified that the CAT would terminate if the *SE* dropped below 0.3 *or* the change in theta

estimates from one item to the next dropped below 0.01. This resulted in a total of four additional algorithms with different combinations of termination rules. A description of the terminations rules is provided in Table 1. For comparative purposes a CAT algorithm was run with no termination rules in order to generate scores based on administration of the full item pool.

Final theta estimates and standard errors were produced for each CAT simulation using maximum a posteriori with standard normal priors. Pearson's correlations were generated to compare theta scores estimated using the CAT algorithms with true theta scores (used to generate the data) and theta scores generated by administration of the full item pool. In addition, the root-mean-square deviation (RMSD) was calculated conditional on the underlying true theta scores (grouped in deciles). All CAT simulations were run using the mirtCAT package for R (Chalmers, 2016).

After the simulations, one CAT algorithm with an optimal balance between efficiency and precision was selected and examined using a real world dataset, that is, the responses from participants with complete data in the third wave of data collection from the ESI development sample. To test the validity of the factor and facet scores generated by the CAT among this population, Pearson's correlations were generated between scores estimated by the CAT with the superordinate factor scores estimated using scores for the 23 facet scales from the full ESI and scores for the ESI-BF item-based factor scales. Correlations were also generated between the factor and facet scores estimated by the CAT with sum scores for the ESI-100 item brief version to measure externalizing and scores for each of the 23 lower order facet scales of the full ESI.

Results

Item Pool Calibration

The unidimensional model was applied separately to items from the three item-based factor scales as well as the 23 facet scales of the ESI. Model fit statistics for the three item-based factor scales were within the acceptable range for Disinhibition (CFI = 0.97, TLI = 0.97, RMSEA = 0.07) and Callous Aggression (CFI = 0.91, TLI = 0.90, RMSEA = 0.08). For Substance Abuse, the CFI and TLI values were 0.97 and 0.96 indicating good model fit but the RMSEA value was poor at 0.11. The mean factor loading for the Disinhibition scale items was 0.75 (range = 0.50–0.90), with corresponding values of 0.72 (range = 0.61–0.88) and (range = 0.70–0.97) for items comprising the Callous Aggression and Substance Abuse scales. For the individual facet scales, the unidimensional model provided good to excellent fit according to the CFI and TLI with values exceeding 0.93 for all facet scales (mean CFI = 0.97, mean TLI = 0.97). The RMSEA values ranged from good to mediocre (M = 0.07, range = 0.04–0.10) with Drug Use evidencing the best fit and Impatient Urgency evidencing the worst fit. All items had loadings above 0.5 on each of the respective facet scales. The item response parameters for the three item-based factor scales and the 23 facet scales of the ESI estimated by the FlexMIRT program are provided in supplementary Table S1.

Table 1
Termination Rules Associated With Each CAT Algorithm

Standard error termination rules	Difference in theta estimates termination rules		
	Disabled	$\Delta\theta < .01$	$\Delta\theta < .05$
$SE(\theta_{ext}) < .3$	Algorithm 1	Algorithm 3	Algorithm 5
$SE(\theta_{ext}) < .4$	Algorithm 2	Algorithm 4	Algorithm 6

CAT Simulations

The results of the simulations for the item-based factor scales are provided in Table 2. For the simulations assuming a normal distribution (see Table 2), the CAT algorithms resulted in an average number of items administered for each scale between four items (for the least conservative termination rules) and 11 items (for the most conservative termination rules). Focusing on Algorithm 1, the Substance Abuse scale evidenced the most efficient CAT with a reduction from 18 items to six items while maintaining an average standard error of 0.3 followed by the Disinhibition and Callous Aggression scales with approximately a 50% reduction in items. There were further gains in efficiency when the additional termination rules that place a minimum difference in theta estimates generated from one item to the next were incorporated. The precision began to drop, particularly for the Callous Aggression and Disinhibition factors when the minimum difference in theta estimates between one item and the next was raised to 0.05, however nearly all CAT simulations resulted in an average standard error of 0.4 or less. Correlations associated with the CAT scores from all three factors and true and full pool theta estimates were very high (rs > 0.9). Figure 1 provides the RMSD values for each CAT algorithm conditional on true theta; for all three scales the error remains relatively consistent across the dimension despite some indication that error increases at the extreme scores (top and bottom deciles). The RMSD is broken down into bias and variance components for the normal, uniform, and empirical distributions and provided in supplementary Tables S2 and S3. The CAT algorithm that provided a good balance between efficiency and precision included the termination rule of $SE < 0.3$ in the estimated scores or a minimum difference in theta estimates between one item and the next at <0.05 (CAT Algorithm 5).

For the simulations assuming a uniform distribution (see Table 2), the CAT algorithms resulted in a higher average number of items administered and higher standard errors than those observed in the simulations using the normal distribution. Despite this, the overall trend in terms of efficiency and precision across the algorithms remained similar to the simulations using the normal distribution. The correlations between all CAT algorithms and true and full pool theta estimates were very high (rs ≥ 0.95). Again, the algorithm that terminated the assessment after reaching an $SE <$ 0.3 or a difference between theta estimates of <0.05 provided a good balance between efficiency and precision. Finally, for the simulations assuming an empirical distribution (see Table 2), the CAT algorithms provided very similar results to the simulations assuming a normal distribution. There was some indication that the CAT simulation for the Callous Aggression scale resulted in relatively weak correlations with the true theta estimates but they still remained high with $r > .81$. Once again, the CAT algorithm that included a termination rule of $SE < 0.3$ or minimum difference in theta estimates between one item and the next at <0.05 resulted in a good balance between efficiency and precision. Theta

Table 2
CAT Simulations, Average Number of Items Administered, Mean Theta Scores, Mean Standard Errors, and Correlations With True and Full Pool Theta Estimates in the Normal, Uniform, and Empirical Distributions

Algorithm	Avg items			Mean theta and *SE*						True theta			Full pool theta		
	Dis	Agg	Sub	$\theta_{(dis)}$	$SE_{(\theta dis)}$	$\theta_{(agg)}$	$SE_{(\theta agg)}$	$\theta_{(sub)}$	$SE_{(\theta sub)}$	$r^2_{(\theta dis)}$	$r^2_{(\theta agg)}$	$r^2_{(\theta sub)}$	$r^2_{(\theta dis)}$	$r^2_{(\theta agg)}$	$r^2_{(\theta sub)}$
Normal distribution															
Full pool	20	19	18	.00	.25	.03	.28	.01	.23	.97	.96	.96	1.00	1.00	1.00
1	10	11	6	.00	.30	.03	.32	.00	.29	.95	.94	.94	.98	.98	.98
2	5	7	5	.00	.38	.04	.39	.00	.30	.91	.91	.93	.95	.95	.97
3	9	11	5	.00	.30	.03	.32	.00	.29	.95	.94	.94	.98	.98	.97
4	5	6	4	.00	.38	.04	.39	.00	.30	.91	.91	.93	.95	.95	.97
5	6	7	4	.01	.34	.02	.37	.00	.30	.93	.92	.93	.96	.96	.96
6	4	5	4	.01	.39	.03	.41	.00	.31	.91	.90	.93	.94	.93	.96
Uniform distribution															
Full pool	20	19	18	.09	.31	.21	.31	.09	.34	.98	.97	.97	1.00	1.00	1.00
1	13	12	12	.08	.34	.18	.35	.07	.37	.98	.97	.97	.99	.99	.99
2	8	9	10	.00	.40	.16	.41	.07	.37	.95	.96	.97	.97	.98	.99
3	12	10	9	.09	.34	.18	.36	.07	.37	.98	.97	.97	.99	.99	.99
4	6	8	8	.01	.40	.16	.41	.07	.37	.95	.96	.96	.97	.98	.99
5	7	7	5	.05	.37	.23	.40	.05	.38	.96	.96	.96	.98	.98	.99
6	5	5	5	.02	.41	.20	.44	.05	.38	.95	.95	.96	.97	.98	.99
Empirical distribution															
Full pool	20	19	18	.01	.25	.04	.26	.14	.20	.96	.88	.95	1.00	1.00	1.00
1	10	11	4	.02	.30	.04	.31	.11	.29	.94	.87	.92	.98	.98	.95
2	5	6	4	.02	.38	.05	.39	.11	.29	.90	.84	.92	.94	.94	.95
3	10	10	4	.01	.30	.04	.31	.11	.29	.94	.87	.92	.98	.97	.95
4	5	6	3	.02	.38	.05	.39	.11	.29	.90	.84	.91	.94	.94	.95
5	7	7	3	.02	.34	.03	.35	.10	.29	.92	.84	.91	.97	.95	.95
6	4	5	3	.02	.39	.04	.41	.11	.29	.90	.82	.91	.94	.92	.95

Note. CAT = computerized adaptive test; Dis = Disinhibition; Agg = Callous Aggression; Sub = Substance Abuse.

estimates generated by these algorithms also demonstrated strong associations with full pool theta estimates ($rs \geq 0.92$).

Correlations between theta scores generated by CAT Algorithm 5 and the theta scores generated using the full ESI facet scales, ESI-BF item-based factor scales, ESI-100 sum scores, and the 23 lower-order facet scales of the full ESI among a subsample of the development sample with complete data are provided in Table 3. Overall, the theta scores for the Disinhibition factor generated by the CAT algorithm reflected externalizing scores generated using the full ESI facet scales and ESI-100 sum scores. The relationship between scores on the respective CAT item-based factor scales and the ESI-BF item based factor scales were very strong with $r > .97$, indicating that the CAT algorithm replicated scores for the ESI-BF item based factor scales using on average one third of the items (18 vs. 57 items). The correlations between the CAT item-based factor scales and the 23 lower-order facet scales provided evidence that the CAT Callous Aggression factor indexed a lack of empathy and all forms of aggression whereas the CAT Substance Abuse factor indexed marijuana, drug, and alcohol use and problems.

Estimated scores using all items for each of the 23 facet scales in the three assumed distributions are provided in Table 4. The estimated scores generally reflected the true theta scores with high correlations observed ($r > .90$) for 22 of the 23 facet scales. The Destructive Aggression scale demonstrated a relatively lower correlation of 0.83 when estimated in the empirical distribution. The scores were estimated with a high degree of precision with standard errors averaging between 0.20 and 0.35 for the facet scales in the normal and empirical distributions. The precision decreased in the uniform distribution for all facet scales with standard errors averaging between 0.28 and 0.37. The combination of termination rules used in CAT Algorithm 5 was used to examine the efficiency and precision associated with a CAT version of the 23 facet scales of the ESI.

As can be seen in Table 5, the CAT algorithm resulted in a substantial decrease in the average number of items administered to estimate scores for each of the 23 facet scales. Focusing on the normal distribution, the largest decrease in items was associated with Drug Problems (88%), Alcohol Problems (87%), and Alcohol Use (83%) scales. The smallest decrease in items was associated with Alienation (33%), followed by Planful Control (45%). The total number of items administered on average across the 23 facet scales was 115 compared with 415 of the full ESI (a 72% reduction in items). The average standard error ranged between 0.29 and 0.39 for all facet scales with Marijuana Use resulting in the best precision and Marijuana Problems the worst. The correlations between scores estimated by the adaptive algorithm versus the full item pools were very high ($r > .96$). The adaptive algorithm was also examined for each of the 23 facet scales in participants with complete data from the development sample. Again, there was a substantial reduction in average number of items administered (ranging between 4 and 8) while maintaining an acceptable level of precision (standard errors between 0.31 and 0.43). Correlations with the full item pools were uniformly high with values between 0.93 and 0.98.

Figure 2 outlines the item exposure rate (i.e., the percent of respondents exposed to each item during the CAT simulations) for

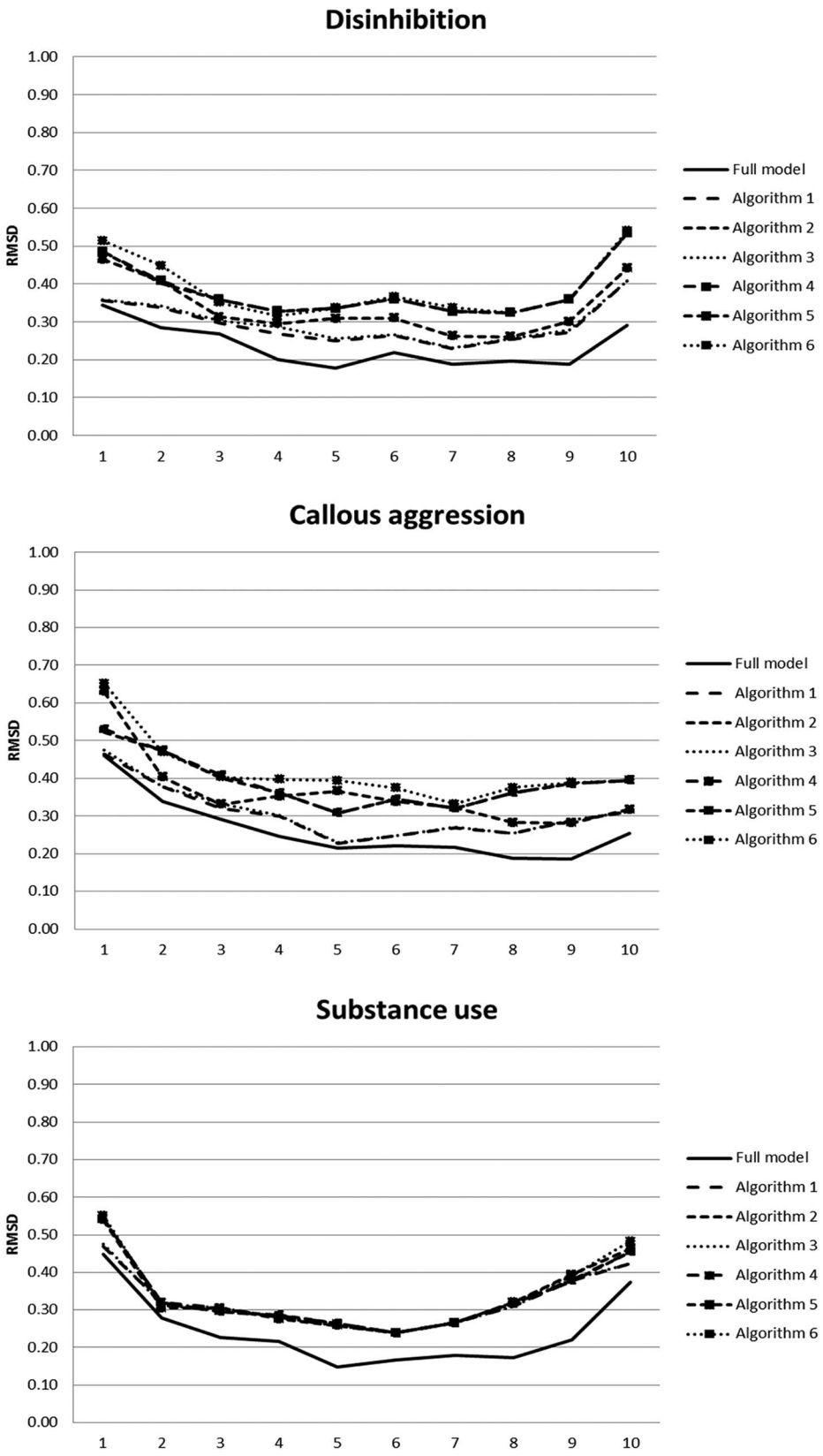

Figure 1. Root-mean-square deviation (RMSD) conditional on true theta (deciles) for the three item-based factor scales under each computerized adaptive test (CAT) algorithm assuming a normal distribution.

Table 3
Correlations Between Factor Scores Generated Using CAT Algorithm 5 and the Full Form ESI-Factor Scores, Brief Item-Based Factor Scales, 100-Item Version of the ESI, and Full Form ESI-Facet Scales Amongst Respondents With Complete Data From the Development Sample

Factor/Facet	CAT Dis	CAT Agg	CAT Sub
Superordinate factors			
ESI-Ext	.94	.48	.69
ESI-Cal	.12	.73	.02
ESI-Sub	.20	.03	.73
ESI-BF-Dis	.97	.49	.65
ESI-BF-Cal	.53	.97	.32
ESI-BF-Sub	.70	.31	.97
ESI-100-Ext	.91	.49	.80
Lower-order facets			
Alcohol problems	.66	.33	.67
Alcohol use	.43	.27	.66
Alienation	.52	.24	.37
Blame externalization	.51	.41	.29
Boredom proneness	.58	.42	.40
(lack of) Dependability	.70	.49	.46
Destructive aggression	.69	.66	.48
Drug problems	.80	.31	.76
Drug use	.75	.31	.88
(lack of) Empathy	.49	.88	.31
Excitement seeking	.59	.52	.51
Fraud	.83	.56	.57
(lack of) Honesty	.56	.55	.32
Impulsive urgency	.75	.43	.51
Irresponsibility	.89	.41	.66
Marijuana problems	.73	.37	.82
Marijuana use	.73	.34	.91
Physical aggression	.75	.63	.55
(lack of) Planful control	.70	.42	.51
Problematic impulsivity	.90	.42	.63
Rebelliousness	.77	.57	.61
Relational aggression	.64	.75	.38
Theft	.88	.44	.67

Note. CAT = computerized adaptive test; ESI = Externalizing Spectrum Inventory; Dis = Disinhibition; Agg = Callous Aggression; Sub = Substance Abuse.

the three CAT item-based factor scales under the normal distribution. Exposure rates are typically examined in educational and aptitude testing where overexposure is particularly problematic for maintaining confidentiality of the item bank. However, the exposure rates were provided here to examine the variability in item presentation and determine whether there is sufficient benefit in an adaptive version in comparison to a brief static version for each scale. All items from the Disinhibition scale and Callous Aggression scale were presented at least once during the simulation whereas 16 items (89%) from the Substance Abuse scale were presented at least once. Excluding the first item automatically administered to all participants, several items from each scale were utilized more or less frequently as evidenced by a downward trend in the exposure rates. The Substance Abuse scale evidenced a relatively flatter exposure rate across the full item bank. The exposure rate for each of the facet scales using CAT Algorithm 5 indicated that all items were presented at least once in 16 out of 23 facets. The Alcohol Problems, Alcohol Use, Drug Problems, Drug Use, Empathy, Marijuana Problems, and Marijuana Use facets contained items that were never administered ranging from 1 item (Drug Use) through to 12 items (Drug Problems). In terms of overexposure, the exposure rate in the total sample for all items was below 80% for 14 of the 23 facets, apart from the initial starting item administered to all respondents. The remaining scales contained items that were administered to more than 80% of the sample ranging from one item (Blame Externalizing, Boredom Proneness, Dependability, Excitement Seeking, Impatient Urgency, and Relational Aggression) to three items (Alienation).

Discussion

Results of the current study reveal that it is feasible to accurately and efficiently administer the ESI-BF item-based factor scales and individual facet scales of the ESI using a computerized adaptive algorithm (i.e., the ESI-CAT). A series of simulations revealed that, using a specific set of termination rules (CAT Algorithm 5), similar scores estimated using the ESI-BF item-based factor scales can be generated using on average 17 items assuming a normal and empirical distribution. The average number of items increased to 19 when simulating the algorithm using the uniform distribution. Similarly, the standard errors associated with scores on the uniform distribution were slightly higher for all three factors compared to the normal and empirical distributions. This is most likely due to the larger proportion of individuals in the upper and lower extremes of the theta continuum (± 2 standard deviations from the mean) for each of the three factors in comparison with the normal and empirical distributions. More items were required to attain the precision termination rules for those simulated respondents with extreme scores suggesting the that item pool utilized in the current study may not target individuals with extreme scores on the three factors as precisely as individuals with more moderate scores. That being said, the parameter estimates utilized by the CAT algorithm were generated using both student and incarcerated adult samples in order to provide information across a wider continuum of the traits in comparison with using a general population sample. The uniform distribution was utilized to test the performance of the CAT under a more extreme condition that would normally not be seen in practice. Thus, the performance a CAT when applied to the majority of populations that wish to efficiently measure the superordinate factors of the ESI should align with the results demonstrated by the normal or empirical simulations.

The correlations for scores for the item-based factor scales and the 23 facet scales generated by the ESI-CAT with scores estimated using the ESI-BF item-based factor scales and the full ESI facet scales were consistently high. This provides some evidence that the ESI-CAT item-based factor and facet scales index constructs similar to those defined by the full ESI model despite the reduced number of items administered. As such, evidence of validity demonstrated by the previous brief form scales of the ESI in relation to similar constructs of externalizing can be expected to carry over to the CAT version of the ESI. However, it should be noted that item content was a major consideration in selection of items for inclusion in the ESI-BF facet and factor scales, and the CAT algorithm operates on measurement properties of items without regard to content. Considering this, more research is needed to evaluate the impact of omitting scale items reflecting particular content from computation of facet and factor scores (for perspec-

Table 4
Number of Items, Mean Score, Standard Error, and Correlation With True Theta for Each ESI Facet Estimated in the Normal, Uniform, and Empirical Distributions

Facet	Items	Normal			Uniform			Empirical		
		M	*SE*	$r^2_{(true)}$	*M*	*SE*	$r^2_{(true)}$	*M*	*SE*	$r^2_{(true)}$
Alcohol problems	30	.08	.21	.96	.24	.28	.97	.06	.21	.97
Alcohol use	23	.02	.21	.97	.03	.30	.98	.07	.20	.95
Alienation	9	.03	.32	.94	.07	.37	.97	.02	.32	.94
Blame externalizing	14	.07	.24	.97	.15	.30	.98	.09	.23	.93
Boredom proneness	12	.04	.24	.97	.06	.31	.98	.05	.24	.96
Dependability	23	.07	.22	.97	.17	.37	.98	.07	.22	.96
Destructive aggression	15	.13	.35	.91	.39	.38	.94	.10	.34	.83
Drug problems	25	.14	.24	.94	.35	.33	.95	.16	.24	.94
Drug use	13	.05	.26	.95	.12	.36	.96	.05	.25	.94
Empathy	31	.08	.20	.97	.21	.25	.98	.07	.20	.96
Excitement seeking	18	.06	.24	.97	.09	.29	.98	.00	.23	.95
Fraud	14	.06	.30	.94	.21	.37	.97	.06	.30	.94
Honesty	15	.06	.25	.97	.16	.29	.98	.04	.24	.96
Impatient urgency	12	.04	.28	.96	.02	.32	.98	.04	.28	.95
Irresponsibility	25	.06	.24	.96	.10	.30	.98	.09	.24	.97
Marijuana problems	18	.16	.29	.92	.39	.36	.94	.16	.28	.90
Marijuana use	17	.05	.24	.95	.12	.35	.96	.01	.25	.97
Physical aggression	21	.06	.25	.96	.19	.32	.97	.06	.25	.95
Planful control	11	.04	.28	.96	.12	.31	.98	.03	.28	.96
Problematic impulsivity	20	.04	.21	.97	.09	.29	.98	.05	.21	.98
Rebelliousness	15	.06	.22	.97	.10	.29	.98	.07	.22	.97
Relational aggression	19	.07	.25	.97	.16	.30	.98	.06	.24	.94
Theft	15	.09	.27	.94	.25	.36	.95	.10	.28	.94

Note. ESI = Externalizing Spectrum Inventory.

tive on the issue of item content, see, e.g., Patrick, Curtin, & Tellegen, 2002; Tellegen & Waller, 2008), in terms of relationships with criterion measures in assessment domains including but not limited to self-report (Patrick et al., 2012, Patrick, Venables, et al., 2013).

The ESI-CAT item-based factor scales, on average, demonstrate an 83% reduction in items from the 100-item brief form and a 70% reduction in items from the 57-item brief form (Hall et al., 2007; Patrick, Kramer, et al., 2013). This added efficiency comes at a cost of losing the ability to administer the ESI in a static fashion. This potential disadvantage may not be too detrimental given the high saturation and relatively low cost of personal computers, tablet computers, smart phones, and Internet use in modern clinical and research settings, which are all capable of administering the CAT version. Moreover, the exposure rates calculated by the optimal CAT algorithm across the various scales indicate that there is added value in utilizing the more complex adaptive algorithm over brief static scales. There was no indication that the same smaller subset of items was administered to every participant across the simulations. The majority of the CAT simulations utilized all items within the item banks when determining scores and there were very few items within each of the scales that were administered to a significant number (>80%) of the total sample. At this point in time however, additional work is required to implement the ESI-CAT in an accessible and open-source online software platform that can facilitate the administration and scoring of CATs in research and clinical settings.

The high correlations (between scores generated using the adaptive algorithms and scores generated using the full item pool) and the average standard errors observed in the current study are in line with previous simulation studies that have utilized CAT algorithms to measure latent levels of psychopathology. Fliege et al. (2005) developed a CAT for depression and found that, on average, six items out of a total of 64 were required to estimate scores with a predefined standard error of 0.32. Similarly, Gibbons et al. (2012) demonstrated that a bifactor CAT could measure general levels of depression using a mean of 12 items per person and a correlation of 0.95 with scores generated using an item pool of 389 items. Walter et al. (2007) demonstrated that between six and eight items from a larger pool of 50 were sufficient to measure latent levels of anxiety with high precision ($SE < 0.32$) and high correlation with the full item pool ($r = .97$). Likewise, Gibbons et al. (2014) found that a bifactor CAT algorithm was able to estimate general anxiety scores with a correlation of 0.94 using, on average, 12 items from a total item pool of 431 items. Kirisci et al. (2012) applied a CAT algorithm to determine the predictive validity of the transmissible liability index for addiction compared with a pen and paper version and found only minor reductions in accuracy (4%) were observed for a large reduction in items administered (79%).Finally, Weiss and Gibbons (2007) utilized a bifactor CAT algorithm to reproduce general psychopathology scores from the Mood-Anxiety Spectrum Scales with a correlation above 0.90 using 25 to 30 items from a total of 615 items.

The favorable results of the current study beg the question of which version of the ESI should be utilized and under what conditions? We provide some tentative suggestions, but interested readers should first identify the core aspects of their own research and clinical work in order to determine how best to proceed. The full 415-item version is one of the most comprehensive assessment tools to measure the externalizing spectrum to date (Krueger et al.,

Table 5
CAT Simulations Using Algorithm 5 for Each of the 23 Facets, Average Number of Items, Means and Standard Errors, Correlations With True Theta, and Full Item Pool Theta for Normal, Uniform, and Empirical Distributions

	Normal					Uniform					Empirical				
Facet	Avg items	*M*	*SE*	$r^2_{(true)}$	$r^2_{(full)}$	Avg items	*M*	*SE*	$r^2_{(true)}$	$r^2_{(full)}$	Avg items	*M*	*SE*	$r^2_{(true)}$	$r^2_{(full)}$
Alcohol problems	4	.08	.33	.93	.96	6	.25	.36	.95	.98	4	.07	.33	.93	.96
Alcohol use	4	.01	.31	.94	.96	5	.05	.37	.97	.98	4	.05	.31	.91	.95
Alienation	6	.03	.36	.93	.99	7	.08	.39	.96	.99	7	.03	.37	.93	.98
Blame externalizing	4	.07	.31	.94	.98	6	.18	.36	.97	.99	4	.08	.31	.90	.97
Boredom proneness	4	.03	.30	.95	.98	6	.05	.34	.97	.99	4	.04	.30	.94	.98
Dependability	7	.07	.32	.94	.97	8	.17	.35	.97	.99	7	.09	.32	.93	.97
Destructive aggression	5	.15	.43	.87	.96	5	.41	.46	.92	.98	5	.10	.43	.78	.94
Drug problems	3	.15	.36	.90	.96	4	.36	.41	.93	.98	3	.17	.36	.90	.95
Drug use	4	.04	.32	.92	.97	5	.15	.39	.95	.99	4	.05	.32	.91	.97
Empathy	6	.09	.32	.94	.97	7	.21	.35	.97	.99	6	.08	.32	.91	.96
Excitement seeking	5	.05	.33	.94	.97	7	.08	.36	.97	.99	5	.01	.33	.92	.97
Fraud	6	.08	.38	.91	.97	6	.21	.43	.96	.99	5	.06	.38	.91	.97
Honesty	6	.05	.32	.95	.98	6	.15	.34	.97	.99	6	.03	.32	.93	.97
Impatient urgency	6	.04	.35	.93	.97	7	.01	.36	.97	.99	6	.03	.35	.93	.97
Irresponsibility	7	.06	.34	.92	.96	7	.07	.38	.97	.99	7	.09	.34	.93	.96
Marijuana problems	4	.18	.39	.89	.96	5	.40	.43	.92	.99	4	.17	.38	.87	.96
Marijuana use	4	.06	.29	.92	.97	5	.16	.37	.95	.99	4	.04	.31	.96	.98
Physical aggression	5	.06	.35	.93	.96	6	.19	.39	.96	.99	6	.05	.35	.91	.96
Planful control	6	.04	.33	.94	.98	7	.11	.35	.98	.99	6	.05	.33	.93	.98
Probl impulsivity	5	.05	.31	.95	.97	7	.09	.36	.97	.99	5	.05	.30	.95	.97
Rebelliousness	4	.06	.31	.94	.97	6	.10	.34	.97	.99	4	.07	.31	.94	.97
Relational aggression	6	.07	.34	.93	.96	7	.16	.38	.97	.99	6	.06	.34	.90	.96
Theft	4	.08	.34	.92	.97	5	.23	.40	.95	.99	4	.10	.34	.92	.98

Note. CAT = computerized adaptive test.

2007). This version should be considered when the goal is to study the externalizing spectrum in fine detail with a high level of precision. Importantly, the detail provided by the full version is particularly suitable for those who wish to focus on the individual symptoms that comprise both the lower- and higher-order factors. The static brief forms or adaptive algorithms should be considered when the costs to administer the full version become too prohibitive and researchers are willing to accept some loss of precision and detail. The adaptive algorithms offer the most efficient method to estimate scores for the brief item-based factor scales as well as the 23 individual facets. If the use of computers in a particular setting is not feasible or pen-and-paper administration is more desirable, than the alternative option is the static brief version to measure the 23 facets and the Disinhibition, Callous Aggression, and Substance Abuse factors directly through the item-based factors scales embedded within the 160 items (Patrick, Kramer, et al., 2013).

There are some limitations of the current study that require further discussion. The ESI-CAT was developed under the assumption that the high correlations with the full model would imply that similar constructs were being measured. This would also imply that the good validity demonstrated by previous studies on the ESI would translate to the ESI-CAT. However, further validation studies are required using the ESI-CAT as a means of confirming the validity specific to the adaptive administration of the ESI as well as to examine any loss of information in comparison to the ESI-BF due to the adaptive algorithms. Related to this limitation, the current study developed and simulated the adaptive algorithm using item data obtained under a single static administration of the ESI. Thus the correlations between the adaptive algorithm and the full pool may be optimized due to this single administration and the fact that the items selected by the adaptive algorithm form a subset of the full measure. Additionally, the order of the item administration can change dramatically across individuals using the adaptive algorithm and the effects of this remain untested in the current study. Further live testing of the ESI-CAT as well as separate administration of the brief and full forms are required to examine if the strong correlations remain, as well as to examine if the order of item presentation influences final scores in any way. It is also possible that different subgroups of the population with the same latent scores might display a different probability of endorsing various items on the ESI. Future studies are required to investigate levels of differential item functioning in the full item pools of the ESI to determine the effect, if any, on various subgroup comparisons. Finally, the current study estimated the IRT parameters among a combined sample of student and incarcerated adult populations. As outlined in Krueger et al. (2007), student and incarcerated adult populations were combined in order to have a sufficient range of responses to various items when estimating the item parameters. These samples were also selected and combined to provide a wider distribution of the externalizing spectrum so that the resulting model could be applicable to a wide range of people. Indeed, O'Connor (2002) previously demonstrated that psychometric models tend to generalize well across a diverse range of samples. However, these samples differ in a number of characteristics from the general population as well as treatment-seeking clinical populations and it remains to be seen whether these differences are large enough to influence the generaliz-

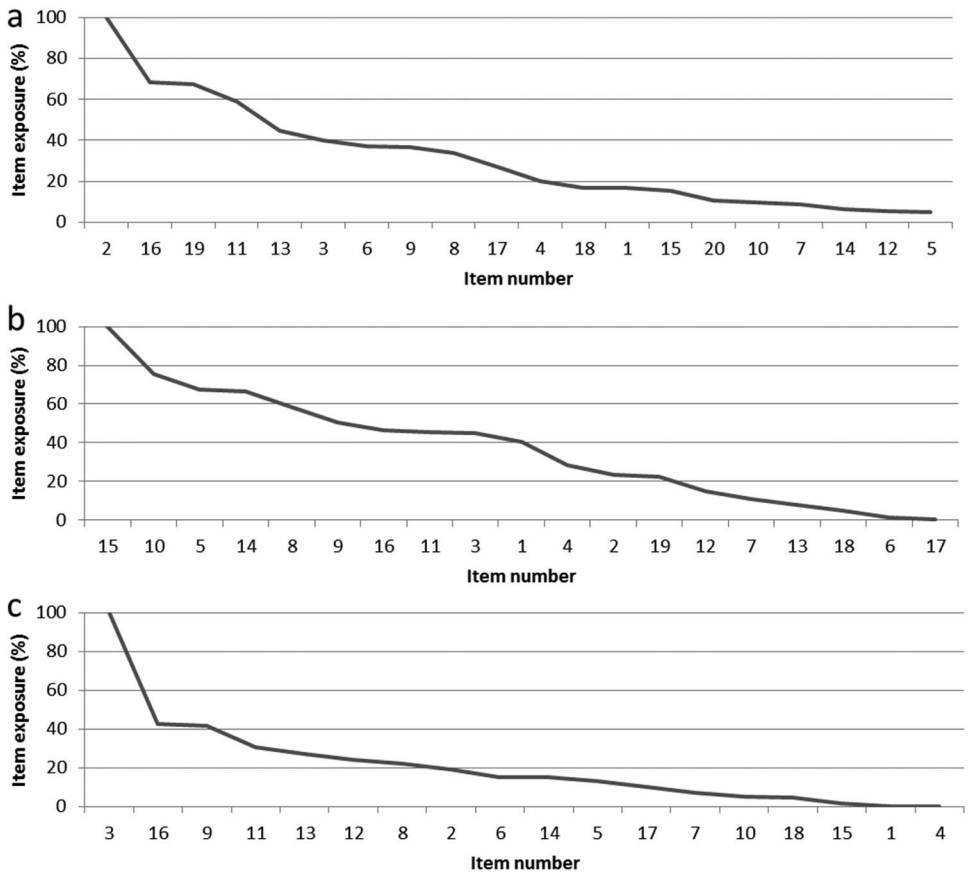

Figure 2. Item exposure rates for each specific item for the computerized adaptive test (CAT) version of the item-based factor scales in a normal distribution: (a) Disinhibition, (b) Callous Aggression, and (c) Substance Abuse.

ability of these findings. Therefore, it would be valuable for additional work to collect information regarding the performance of the ESI-CAT from a variety of different populations.

Recent calls by the National Institute of Mental Health Research Domain Criteria initiative to focus research efforts on dimensional and biologically meaningful constructs of clinically relevant phenomena have resulted in an increasing number of studies investigating the biological basis of externalizing problems using the externalizing spectrum model (Dick, 2007; Hall, Bernat, & Patrick, 2007; Nelson et al., 2016; Patrick & Drislane, 2015; Patrick, Durbin, & Moser, 2012; Young et al., 2009). For the foreseeable future, the use of psychometric instruments and self-report symptom data provide an essential link between the identification of novel biomarkers and the varying manifestations of externalizing problems. The current study successfully demonstrates that the use of an adaptive algorithm can generate similar scores on the item-based factor scales and the individual facets of the ESI using a fraction of the total item pool. Indeed, the current study provides researchers and clinicians with an additional psychometrically valid tool to measure the ESI factors and facets in a highly efficient and precise manner.

References

Achenbach, T. M., & Edelbrock, C. S. (1984). Psychopathology of childhood. *Annual Review of Psychology, 35,* 227–256. http://dx.doi.org/10.1146/annurev.ps.35.020184.001303

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16,* 95–108. http://dx.doi.org/10.1007/s11136-007-9168-6

Blonigen, D. M., Patrick, C. J., Gasperi, M., Steffen, B., Ones, D. S., Arvey, R. D., . . . do Nascimento, E. (2011). Delineating the construct network of the Personnel Reaction Blank: Associations with externalizing tendencies and normal personality. *Psychological Assessment, 23,* 18–30. http://dx.doi.org/10.1037/a0021048

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459. http://dx.doi.org/10.1007/BF02293801

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model

fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–172). Newbury Park, CA: Sage.

Byrne, B. M. (2012). *Structural equation modelling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge/Taylor and Francis.

Cai, L. (2013). flexMIRT Version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

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.

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: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 19,* 125–136. http://dx.doi.org/10.1007/s11136-009-9560-5

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology, 122,* 928–937. http://dx.doi.org/10.1037/a0034028

De Beurs, D. P., de Vries, A. L., de Groot, M. H., de Keijser, J., & Kerkhof, A. J. (2014). Applying computer adaptive testing to optimize online assessment of suicidal behavior: A simulation study. *Journal of Medical Internet Research, 16,* e207. http://dx.doi.org/10.2196/jmir.3511

Dick, D. M. (2007). Identification of genes influencing a spectrum of externalizing psychopathology. *Current Directions in Psychological Science, 16,* 331–335. http://dx.doi.org/10.1111/j.1467-8721.2007.00530.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 14,* 2277–2291. http://dx.doi.org/10.1007/s11136-005-6651-9

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *JAMA Psychiatry, 69,* 1104–1112. http://dx.doi.org/10.1001/archgenpsychiatry.2012.14

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *The American Journal of Psychiatry, 171,* 187–194. http://dx.doi.org/10.1176/appi.ajp.2013.13020178

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31,* 197–218. http://dx.doi.org/10.1207/s15327906mbr3102_3

Hall, J. R., Bernat, E. M., & Patrick, C. J. (2007). Externalizing psychopathology and the error-related negativity. *Psychological Science, 18,* 326–333. http://dx.doi.org/10.1111/j.1467-9280.2007.01899.x

Harkness, A. R., McNulty, J. L., & Ben-Porath, Y. S. (1995). The Personality Psychopathology Five (PSY-5): Constructs and MMPI-2 scales. *Psychological Assessment, 7,* 104–114. http://dx.doi.org/10.1037/1040-3590.7.1.104

Hicks, B. M., Krueger, R. F., Iacono, W. G., McGue, M., & Patrick, C. J. (2004). Family transmission and heritability of externalizing disorders: A twin-family study. *Archives of General Psychiatry, 61,* 922–928. http://dx.doi.org/10.1001/archpsyc.61.9.922

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry, 167,* 748–751. http://dx.doi.org/10.1176/appi.ajp.2010.09091379

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,* S79–S86. http://dx.doi.org/10.1016/j.drugalcdep.2012.01.016

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56,* 921–926. http://dx.doi.org/10.1001/archpsyc.56.10.921

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: An integrative quantitative model of the adult externalizing spectrum. *Journal of Abnormal Psychology, 116,* 645–666. http://dx.doi.org/10.1037/0021-843X.116.4.645

Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for *DSM-V*. *Journal of Abnormal Psychology, 114,* 537–550. http://dx.doi.org/10.1037/0021-843X.114.4.537

Krueger, R. F., & South, S. C. (2009). Externalizing disorders: Cluster 5 of the proposed meta-structure for *DSM-V* and ICD-11. *Psychological Medicine, 39,* 2061–2070. http://dx.doi.org/10.1017/S0033291709990328

Lahey, B. B., Rathouz, P. J., Van Hulle, C., Urbano, R. C., Krueger, R. F., Applegate, B., . . . Waldman, I. D. (2008). Testing structural models of *DSM–IV* symptoms of common forms of child and adolescent psychopathology. *Journal of Abnormal Child Psychology, 36,* 187–206. http://dx.doi.org/10.1007/s10802-007-9169-5

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modelling. *Psychological Methods, 1,* 130–149. http://dx.doi.org/10.1037/1082-989X.1.2.130

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23,* 187–194. http://dx.doi.org/10.1177/01466219922031310

Nelson, L. D., Strickland, C., Krueger, R. F., Arbisi, P. A., & Patrick, C. J. (2016). Neurobehavioral traits as transdiagnostic predictors of clinical problems. *Assessment, 23,* 75–85.

O'Connor, B. P. (2002). The search for dimensional structure differences between normality and abnormality: A statistical review of published data on personality and psychopathology. *Journal of Personality and Social Psychology, 83,* 962–982.

Patrick, C. J., Curtin, J. J., & Tellegen, A. (2002). Development and validation of a brief form of the Multidimensional Personality Questionnaire. *Psychological Assessment, 14,* 150–163. http://dx.doi.org/10.1037/1040-3590.14.2.150

Patrick, C. J., & Drislane, L. E. (2015). Triarchic model of psychopathy: Origins, operationalizations, and observed linkages with personality and general psychopathology. *Journal of Personality, 83,* 627–643.

Patrick, C. J., Durbin, C. E., & Moser, J. S. (2012). Reconceptualizing antisocial deviance in neurobehavioral terms. *Development and Psychopathology, 24,* 1047–1071. http://dx.doi.org/10.1017/S0954579412000533

Patrick, C. J., Kramer, M. D., Krueger, R. F., & Markon, K. E. (2013). Optimizing efficiency of psychopathology assessment through quantitative modeling: Development of a brief form of the Externalizing Spectrum Inventory. *Psychological Assessment, 25,* 1332–1348. http://dx.doi.org/10.1037/a0034864

Patrick, C. J., Venables, N. C., Yancey, J. R., Hicks, B. M., Nelson, L. D., & Kramer, M. D. (2013). A construct-network approach to bridging diagnostic and physiological domains: Application to assessment of

externalizing psychopathology. *Journal of Abnormal Psychology, 122,* 902–916. http://dx.doi.org/10.1037/a0032807

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D., & the PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment, 18,* 263–283. http://dx.doi.org/10.1177/1073191111411667

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347–364. http://dx.doi.org/10.1177/107319110000700404

Roper, B. L., Ben-Porath, Y. S., & Butcher, J. N. (1995). Comparability and validity of computerized adaptive testing with the MMPI-2. *Journal of Personality Assessment, 65,* 358–371. http://dx.doi.org/10.1207/s15327752jpa6502_10

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory and Practice, 15,* 1585–1595. http://dx.doi.org/10.12738/estp.2015.6.0102

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17,* 28–43. http://dx.doi.org/10.1037/1040-3590.17.1.28

Slade, T., & Watson, D. (2006). The structure of common *DSM–IV* and ICD-10 mental disorders in the Australian general population. *Psychological Medicine, 36,* 1593–1600. http://dx.doi.org/10.1017/S0033291706008452

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *Handbook of personality theory and testing: Personality measurement and assessment* (Vol. II, pp. 261–292). London, UK: Sage. http://dx.doi.org/10.4135/9781849200479.n13

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16,* 1–9.

Tucker, L., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10. http://dx.doi.org/10.1007/BF02291170

Venables, N. C., & Patrick, C. J. (2012). Validity of the Externalizing Spectrum Inventory in a criminal offender sample: Relations with disinhibitory psychopathology, personality, and psychopathic features. *Psychological Assessment, 24,* 88–100. http://dx.doi.org/10.1037/a0024703

Vollebergh, W. A., Iedema, J., Bijl, R. V., de Graaf, R., Smit, F., & Ormel, J. (2001). The structure and stability of common mental disorders: The NEMESIS study. *Archives of General Psychiatry, 58,* 597–603. http://dx.doi.org/10.1001/archpsyc.58.6.597

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). *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16,* 143–155. http://dx.doi.org/10.1007/s11136-007-9191-7

Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.), *Proceedings of the GMAC Conference on Computerized Adaptive Testing*. Retrieved from http://publicdocs.iacat.org/cat2010/cat07weiss%26gibbons.pdf

Wolf, A. W., Schubert, D. S., Patterson, M. B., Grande, T. P., Brocco, K. J., & Pendleton, L. (1988). Associations among major psychiatric diagnoses. *Journal of Consulting and Clinical Psychology, 56,* 292–294. http://dx.doi.org/10.1037/0022-006X.56.2.292

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement, 67,* 73–87. http://dx.doi.org/10.1177/0013164406288163

Wright, A. G., Krueger, R. F., Hobbs, M. J., Markon, K. E., Eaton, N. R., & Slade, T. (2013). The structure of psychopathology: Toward an expanded quantitative empirical model. *Journal of Abnormal Psychology, 122,* 281–294. http://dx.doi.org/10.1037/a0030133

Young, S. E., Friedman, N. P., Miyake, A., Willcutt, E. G., Corley, R. P., Haberstick, B. C., & Hewitt, J. K. (2009). Behavioral disinhibition: Liability for externalizing spectrum disorders and its genetic and environmental relation to response inhibition across adolescence. *Journal of Abnormal Psychology, 118,* 117–130. http://dx.doi.org/10.1037/a0014657

Received August 28, 2015
Revision received June 17, 2016
Accepted July 22, 2016 ■