Examining the Factor Structure and Incremental Validity of the Barkley Deficits in Executive Functioning Scale – Short Form in a Community Sample

Kate Clauss, Tracy K. Witte and Joseph R. Bardeen

To cite this article: Kate Clauss, Tracy K. Witte and Joseph R. Bardeen (2021): Examining the Factor Structure and Incremental Validity of the Barkley Deficits in Executive Functioning Scale – Short Form in a Community Sample, Journal of Personality Assessment, DOI: 10.1080/00223891.2021.1887879

To link to this article: https://doi.org/10.1080/00223891.2021.1887879

Published online: 09 Mar 2021.
Examining the Factor Structure and Incremental Validity of the Barkley Deficits in Executive Functioning Scale – Short Form in a Community Sample

Kate Clauss, Tracy K. Witte and, and Joseph R. Bardeen

Department of Psychological Sciences, Auburn University, Auburn, AL

ABSTRACT

The Barkley Deficits in Executive Functioning Scale – Short Form (BDEFS-SF; Barkley, 2011) was developed to assess deficits in five facets of executive functioning. Theoretical assumptions surrounding the BDEFS-SF presume that executive dysfunction is an overarching construct that consists of five domain-specific factors (i.e., a hierarchical model; Barkley, 2011). Prior research has supported a correlated five-factor model, but the tenability of hierarchical or bifactor models of the BDEFS-SF have not yet been tested. In the present study (N = 1,120 community adults), confirmatory factor analysis was used to compare four theoretically relevant models of the BDEFS-SF (i.e., one-factor, correlated five-factor, hierarchical, and bifactor models). The bifactor model provided the best fit to the data. However, the general factor accounted for the overwhelming majority of variance in BDEFS-SF scores and none of the domain-specific factors exhibited adequate construct replicability or factor determinacy. Further, the general factor accounted for the overwhelming majority of variance in criterion variables (i.e., executive attention and health anxiety); the Organization and Emotion factors accounted for a small amount of unique variance in executive attention and the Emotion factor accounted for a small amount of unique variance in health anxiety. Taken together, study findings suggest that the BDEFS-SF has a strong general factor and independent use of the domain-specific factors is contraindicated.
accumulated regarding aspects of EF that were underrepresented in the Prototype-BDEFS (P-BDEFS), additional items were added and the subdomains were renamed according to new factor analytic data (to be discussed later in the introduction).

In the initial investigation of the P-BDEFS, principal components analysis (PCA) was conducted in a sample of 352 adults. The sample was made up of adults (a) with ADHD (41.4%), (b) seeking treatment for problems other than ADHD (i.e., clinical control [27.6%]), and (c) from the community (i.e., control participants [31.0%]). PCA resulted in ten factors with eigenvalues greater than one. Five of those factors consisted of at least 10 items each with each factor explaining at least 2% of the variance in the model, and thus, were retained. Three items were removed because they did not have a primary loading of .40 or greater on any factor. The 88-item solution accounted for approximately 63% of the variance in the model, with the factor relating to time management explaining 51% of the total variance. At this point, researchers decided to employ a varimax rotation to account for possible correlations between the subscales, which resulted in each factor accounting for 8.6 – 15.7% of the variance in the model. Using this model, individuals with ADHD endorsed significantly higher EF deficits than control participants.

After this initial investigation, the authors determined that one limitation of the P-BDEFS was a lack of items assessing emotion regulation. While the self-activation/concentration scale included some emotion regulation items, the construct was weakly represented (Barkley, 2011). Thus, 10 items, based on Gross’s (1998) process model of emotion regulation, were added to better assess self-regulation of emotion, and two items were added to better assess self-motivation. The resultant 100-item scale was administered to a nationally representative sample of adults (N = 1,249). PCA resulted in five factors that had eigenvalues greater than two, and at least 12 items per factor with primary loadings greater than .4. Eighty-nine items were retained for the BDEFS-Long Form (LF). The self-activation/concentration scale from P-BDEFS was not included in the BDEFS-LF because only four of its original items exhibited factor loadings > .40 on a sixth factor. This factor was replaced with a factor consisting of the newly added emotion regulation items. Additional items that had previously loaded onto the self-activation/concentration scale either no longer exhibited meaningful primary loadings (< .4) or were accounted for by one of the five remaining factors. The five factors were renamed: self-organization/problem solving, self-management to time, self-restraint, self-regulation of emotion, and self-motivation. Using varimax rotation, each of the five factors accounted for 8-13.6% of the variance in the model.

The original validation study provided psychometric support for the BDEFS-LF and -SF, including evidence of internal consistency, retest reliability over a 2- to 3-week period in a subsample of a community adult sample (n = 62), and construct validity (Barkley, 2011). Of note, the intercorrelations of the BDEFS factor scores were all large in size (r’s from .55 to .77; Barkley & Murphy, 2011; Cohen, 1992). As described by Barkley (2011), “this implies that there is likely to be an underlying unitary construct of EF (perhaps comparable to g in research on intelligence tests), yet with unique variance accounted for by each subscale apart from the others,” (p. 68). As such, in addition to calculating domain-specific subscales scores, the 20 items of the BDEFS-SF are summed and used as a general measure of deficits in EF in both clinical and research settings.

Consistent with the proposition that the executive functioning deficits are a transdiagnostic risk factor for wide variety of maladaptive outcomes, the BDEFS total and subscale scores have exhibited medium to large magnitude correlations with measures of posttraumatic stress symptoms (e.g., r = .56; community sample; Bardeen & Fergus, 2018), negative metacognitive beliefs (e.g., r = .61; community sample; Bardeen & Fergus, 2018), depressive symptoms (e.g., r = .51; student sample; Feldman et al., 2013), and both state and trait anxiety (e.g., rs from .29 to .62; college student sample; Franklin et al., 2018). Additionally, evidence suggests that while the BDEFS-LF exhibits large magnitude correlations with measures of ADHD symptoms, it is statistically and conceptually distinct from ADHD (Jarett, 2016). A substantial number of individuals in both community and ADHD-specific samples score in the significantly impaired range on the P-BDEFS, suggesting that the measure is capable of identifying behavioral deficits related to EF that are not simply an artifact of overlap with ADHD symptoms or only evident in clinical samples (Barkley & Murphy, 2011). In fact, the initial validation of the BDEFS was in a community sample. The brief nature of the BDEFS-SF and its ability to be used in easier-to-obtain, unslected samples (i.e., samples not chosen based on clinical symptoms or other specific characteristics), make it particularly appealing for use in this type of research. Similarly, the BDEFS may be an optimal screener for EF deficits in clinical settings because it can be used with both ADHD and non-ADHD samples and creates minimal patient burden.

Since its creation more than eight years ago, there has not been a subsequent evaluation of the psychometric properties of the BDEFS-SF. This is problematic because the BDEFS-SF is currently used in research and clinical settings (i.e., as a screener for EF deficits, or to track EF deficits over the course of treatment). As discussed, there are several ways of using the BDEFS-SF. Use of the total score assumes that they have incremental utility beyond the general factor. Theoretical assumptions surrounding the BDEFS-SF presume that executive dysfunction is an overarching construct with sub-domains subsumed below it (i.e., a hierarchical model; Barkley, 2011). To date, research has only examined a correlated five-factor model, which is unable to test the existence of an overarching EF construct. Moreover, the correlated five-factor model does not test the...
assumption that the domain-specific factors provide unique information beyond the total score (Reise, 2012).

To adequately test these assumptions, examination of a hierarchical and bifactor model of the BDEFS-SF is warranted. Hierarchical models include direct paths from the general (higher-order) factor onto uncorrelated domain-specific (lower-order) factors. This allows a hierarchical model to assess the degree to which covariation among domain-specific factors is captured by the general factor (Brown, 2015). Importantly, a hierarchical model cannot test the assumption that items from domain-specific factors provide unique information beyond the higher-order construct. In contrast, bifactor modeling can be used to examine the degree to which items are representative of the general factor versus domain-specific factors and can be used to determine whether the domain-specific factors are meaningfully distinct from the general factor (Reise, 2012). This is accomplished by allowing indicators to simultaneously load onto the general factor and lower-order domains while fixing factor correlations to zero. Thus, bifactor modeling isolates the unique contributions of the domain-specific and general factor and allows us to determine whether use of the total score is sufficient, or whether the subscale scores provide unique information beyond the total score, thus warranting their continued use.

The hierarchical model is nested within the bifactor model (Chen et al., 2006; Yung et al., 1999), thus allowing us to use chi-square difference testing to quantify the degree of difference in model fit. Some have suggested that the value of comparing bifactor models to alternative models is limited because bifactor models often provide better fit to the data because of their inherent qualities (e.g., Bonifay et al., 2017). As such, we also examined several additional indices developed for use with bifactor modeling that provide information central to the aims of this study (e.g., determining whether the domain-specific factors have value beyond the general factor, determining the stability and replicability of the factors; Rodriguez et al., 2016). Through bifactor modeling, our study was designed to explicate the degree to which using the subscale scores of the BDEFS-SF is justified and make recommendations about future use of the measure (Reise et al., 2013).

An additional benefit of bifactor modeling is that the relation between domain-specific factors and criterion variables can be examined while holding the general factor constant (Brown, 2015). Thus, as a preliminary investigation, we used bifactor modeling to examine the incremental utility (i.e., unique variance) of the BDEFS-SF domain-specific factors in predicting two criterion measures (i.e., executive attention and health anxiety) after accounting for the general factor (i.e., EF deficits). Executive attention was selected as a criterion variable because a) data suggest that it relies on some of the same cognitive process as the cognitive and behavioral manifestations of executive dysfunction identified by Barkley (Spruijt et al., 2018) and b) executive attention (assessed via the same measure used in the present study) has exhibited medium-sized associations with the BDEFS-SF in prior research (Clauss & Bardeen, 2020). Health anxiety was used as the criterion variable because (a) small to medium-sized correlations have been observed between the BDEFS and several self-report measures of anxiety (e.g., Franklin et al., 2018 [BDEFS-SF]; Jarett, 2016 [BDEFS-LF]) and (b) health anxiety was a construct of interest in the larger study from which the data for this study were drawn. Using executive attention and health anxiety as criterion variables should allow for an adequate initial test of the incremental utility of the BDEFS-SF subscales beyond the total score.

We hypothesized that the total score of the BDEFS-SF would exhibit a significant positive relationship with executive attention, such that greater EF deficits would be associated with relatively worse executive attention. We hypothesized that there would be a significant positive relationship between the total score of the BDEFS-SF and health anxiety, such that greater EF deficits would be associated with greater health anxiety symptoms. There is relatively little empirical evidence to support making specific hypotheses about the subscales scores of the BDEFS-SF and how they might relate to these criterion variables. As such, we did not make specific hypotheses regarding potential differences in the strength of associations between the BDEFS-SF subscales and health anxiety or executive attention. However, because each of the BDEFS-SF subscales are supposed to represent the same overarching construct, we predicted that the subscales would relate to the the criterion variables in the same direction as the BDEFS-total score.

Method

Participants and procedure

Participants were community adults (N = 1,255) recruited via Amazon’s Mechanical Turk (MTurk). MTurk is an online labor market where adults from the general population can be recruited to complete questionnaires in exchange for payment. In comparison to American undergraduate samples, MTurk samples tend to be more demographically diverse and a number of studies support the quality of data collected via MTurk (Behrend et al., 2011; Buhrmester et al., 2011; Shapiro et al., 2013; Paolacci & Chandler, 2014). Recruitment was limited to MTurk users who were fluent in English, located within the United States, and between the ages of 18–65. In addition, as a quality control measure, only participants with at least a 95% approval rating from requesters in Mturk and who had completed at least 50 past human intelligence tasks were allowed to participate in this study. Quality control measures such as these have been shown to improve data quality (Peer et al., 2014). One-hundred and thirty-five participants (10.76% of the larger sample) did not complete any items of the BDEFS-SF, and thus, were removed from the final sample.

The majority of the final sample (N = 1,200) self-identified as White (79.7%), followed by Black (8.8%), Asian (8.7%), Other (1.8%), and American Indian/Alaska Native (1.3%). Additionally, 7.8% of the sample reported their ethnicity as Hispanic or Latino. The majority of participants reported their sex as female (58.8%). The average age of
participants was 35.48 years \((SD = 10.84, \text{range} = 18 - 65)\). The majority of participants were married or cohabitating \((57.9\%), 31.7\% \text{ identified as "single, never married," and the remaining} 10.4\% \text{ of the sample identified as either "widowed," "divorced," or "separated."} There was a great deal of variability in annual income; 19.2\% of the sample report an annual income of between $0 and $25,000, 29.3\% reported an income of between $25,000 and $50,000, 25\% reported an income of between $50,000 and $75,000, and the remaining 26.5\% reported an annual income greater than $75,000. Level of education varied with 11.8\% of participants having completed high school or obtained their GED, 34.3\% attending business school, technical school, or obtaining an associate degree, 40.1\% earning a bachelor’s degree, 11.3\% earning a master’s degree, and 1.6\% earning a doctoral degree.

Study procedures were approved by the local institutional review board. Informed consent and self-report measures were administered via a secure online survey program. Participants were able to complete study questionnaires from any device with internet access. Participants completed demographics and the BDEFS-SF as part of a larger study. Upon study completion, participants were debriefed and paid $1.50 as compensation for their time, an amount consistent with precedence for paying MTurk workers in similar studies (Buhrmester et al., 2011).

**Measures**

The 20-item BDEFS-SF is scored on a 1 \((\text{never or rarely})\) to 4 \((\text{very often})\) rating scale. Higher scores indicate greater deficits in executive functioning over the past 6 months. The BDEFS-SF includes five subscales \((\text{subscale scores ranged from} 4-16)\): self-management to time (time; \(M = 7.56, SD = 3.16\)), self-organization/problem solving (organization; \(M = 6.55, SD = 3.00\)), self-restraint (restraint; \(M = 6.06, SD = 2.58\)), self-motivation (motivation; \(M = 6.06, SD = 2.64\)), and self-regulation of emotion (emotion; \(M = 6.75, SD = 3.28\)). The total score exhibited adequate internal consistency in the present study \((\alpha = .96, M = 32.98, SD = 12.67, range = 20-80)\) and skewness and kurtosis were within acceptable limits \((\text{skew} = .91, \text{kurtosis} = 0.06)\). The subscale scores of the BDEFS-SF also exhibited adequate internal consistency \((\alpha \text{ s from} .84 \text{ to} .92)\). All inter-item correlations were significant \((p < .01)\) and positive.

The short form of the Attentional Control Scale \((\text{ACS}; Derryberry \& Read, 2002)\) is a 12-item measure of executive attention that consists of items assessing attentional focusing and shifting. Items are scored on a 4-point scale \((1 = \text{almost never to} 4 = \text{always})\). Items on the ACS were coded so that higher scores were indicative of relatively worse executive attention. ACS scores have exhibited acceptable internal consistency in past research (Judah et al., 2014) and in the current study \((\alpha = .85; M = 26.32, SD = 6.77, range = 12-46)\). The ACS was administered to a subset of the total sample in the current study \((n = 597)\), and thus, the results of the structural regression described below, in which executive attention serves as the criterion variable, is confined to this smaller sample.

The Whiteley Index-6 \((\text{WI-6}; \text{Welch et al., 2009})\) is a brief self-report measure of health anxiety derived from the original Whiteley Index (Pilowsky, 1967). Participants rate six statements \((\text{e.g., "Do you often worry about the possibility that you have got a serious illness?"})\) on a five-point scale \((1 = \text{not at all to} 5 = \text{a great deal})\). WI-6 scores have exhibited high internal consistency in past research \((\alpha = .87; \text{Welch et al., 2009})\) and the current study \((\alpha = .92; M = 13.41, SD = 6.21, range = 6-30)\).

**Data analytic plan**

**Confirmatory factor analysis**

Model comparisons were made using confirmatory factor analysis \((\text{CFA})\) in MPlus (Muthén & Muthén, 2015). Mean and variance-adjusted weighted least squares \((\text{WLSMV})\) estimation was used to test all models because BDEFS-SF item responses are ordered categories (Asparouhov, 2005). The first model was a one-factor model; all 20 items loaded onto the same EF factor. The second model was a correlated five-factor model, with four items loading onto each factor. The third model was a hierarchical model where the indicators loaded onto their respective lower-order factors, which were subsumed under an overarching general EF factor. In order to create a metric for the hierarchical model, the direct path from the general factor to the first lower-order factor was fixed to zero. Finally, the fourth model was a bifactor model where all 20 items were allowed to load onto the general EF factor, as well as their respective lower-order factors, and the correlations between all factors were fixed to zero.

**Model estimation and comparison**

Model fit was evaluated using three commonly recommended fit indices (Brown, 2015; Kline, 2016): Root Mean Square of Approximation, Tucker-Lewis Fit index \((\text{TLI})\) and the Comparative Fit Index \((\text{CFI})\). RMSEA values < .05 indicate excellent fit, .05 – .08 indicate adequate fit (Browne & Cudeck, 1993), .08 – .10 indicate mediocre to acceptable fit, and > .10 indicates inadequate fit (Meyers et al., 2006). For TLI and CFI, values greater than .90 suggest adequate fit (Bentler, 1990; Meyers et al., 2006).

Scaled chi-square difference testing was conducted to compare nested models (using the method outlined by Satorra & Bentler, 2001). However, chi-square difference tests are strongly influenced by sample size, often indicating a significant difference when actual differences are trivial in magnitude (Cheung & Rensvold, 2002). As such, we also compared models using RMSEA 90% confidence intervals \((\text{CIs})\) (Brown, 2015; Kline, 2016). Differences in model fit are considered non-significant when models have overlapping 90% RMSEA CIs, (Wang & Russell, 2005).

**Bifactor model evaluation**

Further evaluation of the bifactor model was conducted using the following indices (Dueber, 2016; Rodriguez et al., 2016). OmegaH \((\omega_H)\) reflects the proportion of variance in the total score that can be attributed to the general factor.
OmegaHS ($\omega_{HS}$) reflects the proportion of variance attributable to each domain-specific factor after removing the variance due to the general factor. Explained common variance (ECV) serves as a measure of unidimensionality and is calculated as the proportion of common variance that is accounted for by the general factor (Dueber, 2016). Item-level explained common variance (I-ECV) is the amount of variance for each item that is attributable to the general factor. I-ECV values greater than .80 to .85 are indicative of unidimensionality at the item level (Gorsuch, 1983; Stucky & Edelen, 2015). Percentage of uncontaminated correlations (PUC) indicates the percentage of item correlations contaminated by variance attributed to the general and domain-specific factors. PUC is often interpreted in combination with ECV. When both are greater than .70, common variance within a model can be regarded as essentially unidimensional. Average relative parameter bias (ARPB) indicates the average bias across parameters if items are forced into a unidimensional, versus multidimensional, structure. ARPB less than 0.10 or 0.15 suggests that the multidimensionality within a measure is not substantial enough to preclude a unidimensional solution (Muthén et al., 1987; Rodriguez et al., 2016). The correlation between factors and factor scores (i.e., factor determinacy [FD]) serves as an indicator of the degree to which factor scores are of practical value and should be used in measurement models (i.e., FD > .90; Gorsuch, 1983). Finally, construct replicability (H) reflects the degree to which a factor is well defined by its indicators. H values greater than .80 suggest that a latent variable will demonstrate sufficient stability across studies (Hancock & Mueller, 2001).

### Structural regression model
Structural regression was used to determine the extent to which the domain-specific factors of the BDEFS-SF relate to executive attention and health anxiety when holding the general factor constant. The structural regression models consisted of simultaneously modeling the bifactor model of the BDEFS-SF described above and a one-factor model of the WI-6 or a higher-order model of the ACS. The general and domain-specific factors of the BDEFS-SF were regressed onto the WI-6 or ACS. Path coefficients from the general and remaining domain-specific factors to the WI-6 or ACS were freely estimated.

### Results

#### Model estimation and comparison
Fit statistics for all four models are presented in Table 1. The one-factor model was the only model that did not provide adequate fit to the data. For the correlated five-factor, hierarchical, and bifactor models, RMSEA, CFI, and TLI were all within specified guidelines. Goodness-of-fit statistics suggested that the bifactor model provided the best fit to the data, and scaled chi-square difference testing was conducted to more definitively make this determination (Satorra & Bentler, 2001).

First, we compared the one-factor model to the correlated five-factor model. The correlated five-factor model provided significantly better fit to the data, as evidenced by a significant difference test ($\chi^2[10] = 802.29, p < .001$) and non-overlapping RMSEA 90% CIs. For the correlated five-factor model, correlations between latent factors ranged from .74 to .88. Second, we compared the hierarchical model to the correlated five-factor model. Although the chi-square difference test was significant in favor of the hierarchical model ($\chi^2[5] = 52.26, p < .001$) and non-overlapping RMSEA 90% CIs were overlapping, suggesting that difference in model fit might not be meaningful. Examination of the hierarchical model revealed that all lower-order factors exhibited large magnitude standardized factor loadings on the higher-order factor ($\text{Time} = .88$, Organization = .93, Motivation = .95, Restraint = .87, Emotion = .85, all $p < .001$). Third, we compared the hierarchical model to the bifactor model. Chi-square difference testing indicated that the bifactor model provided significantly better fit to the data than the hierarchical model ($\chi^2[15] = 286.25, p < .001$), and RMSEA 90% CIs did not overlap. As such, the bifactor model was retained for further evaluation.

The standardized factor loadings from the bifactor model are provided in Table 2. All items exhibited significant positive loadings onto the general factor (all $p < .001$) and significant positive loadings onto their respective domain-specific factors (all $p < .01$). Item loadings on the domain specific factors tended to be substantially smaller than on the general factor.

### Table 1. Goodness of fit statistics and model comparisons.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>RMSEA 90% CI</th>
<th>LL</th>
<th>UL</th>
<th>CFI</th>
<th>TLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Factor</td>
<td>3014.603</td>
<td>170</td>
<td>.122</td>
<td>.118</td>
<td>.126</td>
<td>.941</td>
<td>.934</td>
</tr>
<tr>
<td>Correlated Five-Factor</td>
<td>921.489</td>
<td>160</td>
<td>.065</td>
<td>.061</td>
<td>.069</td>
<td>.984</td>
<td>.981</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>938.771</td>
<td>165</td>
<td>.065</td>
<td>.061</td>
<td>.069</td>
<td>.984</td>
<td>.982</td>
</tr>
<tr>
<td>Bifactor</td>
<td>611.383</td>
<td>150</td>
<td>.052</td>
<td>.048</td>
<td>.057</td>
<td>.990</td>
<td>.988</td>
</tr>
</tbody>
</table>

### Table 2. Standardized factor loadings for the bifactor model.

<table>
<thead>
<tr>
<th>Item</th>
<th>General</th>
<th>Time</th>
<th>Organization</th>
<th>Restraint</th>
<th>Motivation</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.638</td>
<td>.538</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.789</td>
<td>.285</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.782</td>
<td>.451</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.816</td>
<td>.399</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.844</td>
<td>.234</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.824</td>
<td>.261</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.830</td>
<td>.301</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.824</td>
<td>.416</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.854</td>
<td>.135</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.699</td>
<td>.428</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.729</td>
<td>.530</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>.779</td>
<td>.486</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>.823</td>
<td>.157</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>.802</td>
<td>.155</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>.831</td>
<td>.465</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>.837</td>
<td>.205</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>.778</td>
<td>.517</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>.801</td>
<td>.447</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>.776</td>
<td>.477</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.844</td>
<td>.390</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. All factor loadings were statistically significant at $p < .001$. 

---

**References**

Table 3. Bifactor evaluation indices.

<table>
<thead>
<tr>
<th></th>
<th>General</th>
<th>Time</th>
<th>Organization</th>
<th>Restraint</th>
<th>Motivation</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega_{0.5}$</td>
<td>.984</td>
<td>.926</td>
<td>.936</td>
<td>.928</td>
<td>.923</td>
<td>.958</td>
</tr>
<tr>
<td>$\omega_{H}/\omega_{HS}$</td>
<td>.943</td>
<td>.217</td>
<td>.110</td>
<td>.195</td>
<td>.075</td>
<td>.236</td>
</tr>
<tr>
<td>ECV</td>
<td>.809</td>
<td>.241</td>
<td>.123</td>
<td>.234</td>
<td>.102</td>
<td>.248</td>
</tr>
<tr>
<td>$H$</td>
<td>.974</td>
<td>.485</td>
<td>.306</td>
<td>.485</td>
<td>.270</td>
<td>.521</td>
</tr>
<tr>
<td>FD</td>
<td>.976</td>
<td>.824</td>
<td>.737</td>
<td>.842</td>
<td>.782</td>
<td>.867</td>
</tr>
</tbody>
</table>

Note. $\omega$ = omega; $\omega_{0.5}$ = omega subscale; $\omega_{H}$ = omega hierarchical; $\omega_{HS}$ = omega hierarchical subscale; ECV = explained common variance; $H$ = construct replicability; FD = factor determinacy.

Bifactor model evaluation

Bifactor models often provide better fit to the data because of their inherent qualities (Bonifay & Cai, 2017). As such, it is important to go beyond mere model fit and consider additional statistical indices to identify the most appropriate model. Additional indices derived from the bifactor model are presented in Table 3. Results suggest acceptable reliability of the general and domain specific factors (total $\omega = .98$, subscale $\omega$ from .92 to .96). An overwhelming amount of the reliable variance in BDEFS-SF scores was attributable to the general factor ($\omega_H = .94$). Relatively little reliable variance in the domain-specific factors was independent of the general factor ($\omega_H = .94$). Relatively little reliable variance in the domain-specific factors was independent of the general factor ($\omega_H = .94$). Relatively little reliable variance in the domain-specific factors was independent of the general factor ($\omega_H = .94$). Relatively little reliable variance in the domain-specific factors was independent of the general factor ($\omega_H = .94$).

Moreover, the general factor accounted for the overwhelming majority of variance in BDEFS-SF scores and encourages use of the total score. In addition, FD values for the general factor suggest adequate factor determinacy (.97), whereas the FD values for all five domain-specific factors suggest that these factors may not be suitable for use as summed subscale scores and as latent variables in a SEM framework (FDs ranging from .74 to .88; Gorsuch, 1983). This hypothesis is further supported by domain-specific H values that suggest inadequate construct replicability (Hs from .27 to .52). The general factor exhibited acceptable construct replicability (H = .97).

Structural regression model

The structural regression with executive attention as the criterion variable provided adequate fit to the data: $\chi^2(436) = 1206.53$, $p < .001$; RMSEA = .05 (90% CIs = [.051, .058]); CFI = .97; TLI = .97. The general factor accounted for the overwhelming majority of variance in executive attention ($R^2 = .95$, $\beta = .76$, $p < .001$). After accounting for the general factor, Emotion accounted for a significant, although small, amount of variance in executive attention ($\beta = .14$, $p = .003$). Significant effects were observed between executive attention and both Restraint and Motivation, but the effects were not in the expected direction ($\beta = -.21, p < .001$ and $\beta = -.43, p = .004$, respectively). Time and Organization did not account for variance in executive attention beyond the total score ($\beta = .04, p = .49$ and $\beta = .12, p = .06$, respectively).

The structural regression with health anxiety as the criterion variable provided adequate fit to the data: $\chi^2(273) = 1406.48$, $p < .001$; RMSEA = .06 (90% CIs = [.058, .064]); CFI = .98; TLI = .98. The general factor accounted for the large majority of unique variance in health anxiety ($R^2 = .56$, $\beta = .71$, $p < .001$). After accounting for the general factor, Emotion accounted for a significant amount of unique variance in health anxiety, although the size of the effect was small ($\beta = .19$, $p < .001$). A significant effect was observed between Restraint and health anxiety, but the effect was small and not in the expected direction ($\beta = -.08$, $p = .03$). None of the other domain-specific factors accounted for variance in health anxiety beyond the general factor (Time: $\beta = -.001$, $p = .97$, Organization: $\beta = .03$, $p = .42$, Motivation: $\beta = -.12$, $p = .09$).

Discussion

To our knowledge, this study is the first to examine the factor structure of the BDEFS-SF since the measure was published in 2011. The bifactor model, in which all indicators simultaneously loaded onto a general factor and their respective orthogonal domain-specific factors, provided significantly better fit to the data than competing models (i.e., one-factor model, correlated five-factor model, hierarchical model). However, further evaluation of the bifactor model revealed that item loadings on the domain-specific factors were substantially attenuated and the majority of BDEFS-SF items exhibited content redundancy at the item level. Moreover, the general factor accounted for the overwhelming majority of variance in BDEFS-SF scores and none of the domain-specific factors exhibited adequate construct replicability or factor determinacy, thus indicating that these factor scores are of little practical value and should not be used in measurement models. Taken together, these results are indicative of a measure with a strong general factor that supports use of a total score, but not subscale scores.

A second aim of this study was to test the incremental utility of the BDEFS-SF domain-specific factors in predicting executive attention and health anxiety after accounting for the general factor. The general factor accounted for the overwhelming majority of variance in both executive attention and health anxiety. Both the Emotion and Organization domain-specific factors accounted for a small amount of unique variance in executive attention. The Emotion...
domain-specific factor also accounted for a small proportion of unique variance in health anxiety after accounting for the general factor. This suggests the possibility that the items of these domain-specific factors may provide enough unique variance beyond that of the general factor to be considered a separate domain. The Emotion subscale is comprised of items that were added after the initial psychometric investigation of the P-BDEFS to bolster the weakly represented self-activation/concentration scale. The self-activation/concentration scale was eventually dropped from the measure and the Emotion scale was retained. Items from this scale were developed based on Gross’s (1998) process model of emotion regulation, but the content of the items appears to represent the construct of emotion regulation self-efficacy (e.g., “I have trouble calming myself down once I am emotionally upset,” or “I remain emotional or upset longer than others”; Barkley, 2011). Previous research has shown that emotion regulation self-efficacy is inversely related to multiple dimensions of health anxiety, even after accounting for specific emotion regulation strategies and other emotion regulation deficits (Bardeen & Fergus, 2014). Similarly, emotion-regulation self-efficacy has been shown to be significantly associated with executive attention in prior research (O’Bryan et al., 2017). However, it is worth reiterating that even though the Emotion domain-specific factor accounted for a small amount of unique variance in both executive attention and health anxiety, it did not exhibit adequate construct replicability or factor determinacy.

The BDEFS-SF subscale scores have been used in previous research to examine differential patterns of EF deficits between groups (e.g., ADHD and control; Bender & Privitera, 2016). Observed differences have been interpreted as meaningful and then used to make recommendations for clinical practice. In clinical practice, Barkley (2011) recommends using the BDEFS-SF total score as an indicator of general EF deficits that require additional follow-up. In fact, the BDEFS manual provides normative data for the total score of the short form, but not for the subscale scores. Results from the present study support using the BDEFS-SF in this manner (i.e., using the total score as a screener). Clinicians should be mindful of this when using the BDEFS-SF in clinical settings and clinical research. For example, the total and subscale scores have been used to track changes in EF deficits before and after treatment for ADHD (Puente & Mitchell, 2016). Our data suggest that it would be most useful to interpret differences in the total score alone. Similarly, in assessment settings, elevations on the total score of the BDEFS-SF would need to be followed up with additional testing to establish specific deficits in one or more domains of EF.

Finally, regarding research practices, this study supports continued use of the BDEFS-SF total score. This is relevant to burgeoning research exploring executive functioning in the context of personality. More specifically, EF deficits have been implicated in personality pathology (i.e., antisocial personality disorder: Ogilvie, Stewart, Chan & Shum, 2011; borderline personality disorder: Unoka & Richman, 2016). As such, having psychometrically sound self-report measures of EF deficits with low participant burden is expedient in advancing research in this area. Moreover, some data suggest that self-reported EF, but not behavioral tests of EF, relate to personality dimensions (Buchanan, 2016). This further highlights the importance of having psychometrically sound self-report measures of EF.

Although this study provides important information about the factor structure of the BDEFS-SF, its limitations must be acknowledged. While research suggests that MTurk is capable of producing high quality data (Chandler & Shapiro, 2016), the samples obtained from MTurk tend to be more highly educated and younger than general population samples (Paolacci & Chandler, 2014). In addition, this study was conducted on an unselected sample, which is consistent with several published studies using the BDEFS-SF (Bender & Privitera, 2016; Feldman et al., 2013; Franklin et al., 2018), but does not preclude the need to replicate study findings in samples with relatively high levels of EF deficits. It is possible that the factor structure observed in this study might vary from that obtained in a clinical sample with a broader range of EF deficits (e.g., clinically diagnosed anxiety or ADHD). It is important to highlight the fact that findings from the present study only pertain to the short form of the BDEFS. Independent psychometric investigations of the BDEFS-LF have been conducted (see Kamradt et al., 2019), but are to date, limited to exploration of the correlated five factor model. As such, the results of these investigations cannot speak to the question of whether the subscale scores of the long form provide incremental utility beyond the total score. As such, it will be important in future research to examine the factor structure of the BDEFS-LF using a bifactor modeling approach to determine whether the domain-specific factors are meaningfully distinct from the general factor. Whereas the short form is intended to be a screener, the long form includes 89 items and is intended to be a more complete assessment of EF. It is possible that the full version has better coverage of the lower-order domains, making it capable of providing additional information beyond the total score.

In future research, it might also be important to use performance measures of EF (e.g., Wisconsin Card Sort Test; Heaton, 1981) rather than relying solely on self-report measures. Performance measures have the advantage of being objective; however, some argue that these measures have poor ecological validity (Barkley & Murphy, 2010). This position is supported by several studies showing stronger relationships between self-reported EF and related outcomes (e.g., occupational impairment, behavioral problems, depressive symptoms, and persistent ADHD symptoms) than between self-reported EF and performance measures of EF (Barkley & Fischer, 2011; Barkley & Murphy, 2010; Knouse et al., 2013). However, not all data suggest poor relations between performance measures of EF and self-reported EF deficits (Burgess et al., 1998). Moreover, the relatively stronger relationships observed between self-reported EF and self-reported impairment could be attributable to common method variance. Researchers have suggested that evaluating the psychometric properties of common assessments is an
important first step in bridging the gap between neuro-psychological and self-report methods (Snyder et al., 2015).

As noted previously, bifactor models often exhibit better fit to the data than less complex models due to their high flexibility (Bonifay et al., 2017; Reise, Kim, Mansolf, & Widaman, 2016). Because of this, parsimony, theory, and practicality should guide decisions regarding the selection of the most appropriate model (Brown, 2015). Results from the confirmatory factor analysis and bifactor model evaluation suggest that the one-factor model is the most appropriate model even though the bifactor model provided better fit to the data and only two of three fit indices suggested adequate fit of the one-factor model (CFI and TLI > .90; Bentler, 1990; Meyers et al., 2006). Although the Emotion domain-specific factor accounted for a small amount of unique variance in health anxiety in the structural regression analysis, it, as well as the four other domain-specific factors, did not exhibit adequate construct replicability or factor determinacy. Revising the BDEFS-SF by removing redundant and underperforming items may be worth considering in future research. Refining the BDEFS-SF in this manner may result in a more parsimonious measure with greater ease of interpretability and better unidimensional model fit. In the meantime, results from the present study support continued use of the BDEFS-SF total score, but not subscale scores.

Note. All chi-square differences were statistically significant at p < .001. $\chi^2$ = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CI = confidence interval; LL = lower limit; UL = upper limit; CFI = comparative fit index; TLI = Tucker-Lewis fit index.

ORCID
Joseph R. Bardeen http://orcid.org/0000-0002-4786-109X

References