Published in final edited form as:

Clin Neuropsychol. 2020 January ; 34(1): 243–258. doi:10.1080/13854046.2019.1571634.

When theory met data: factor structure of the BRIEF2 in a clinical sample

Lisa A. Jacobson^{a,b}, Luther G. Kalb^{a,c}, E. Mark Mahone^{a,b}

^aDepartment of Neuropsychology, Kennedy Krieger Institute, Baltimore, MD, USA

^bJohns Hopkins University School of Medicine, Baltimore, MD, USA

^cJohns Hopkins School of Public Health, Baltimore, MD, USA

Abstract

Objective: The BRIEF2 is the recent revision of a frequently employed measure of executive behaviors; however, no research has yet addressed the validity of the new measure's theoretical design.

Method: The present study examined the factor structure of the BRIEF2 in 5212 clinically referred youth (66% male, 5–18 years) via exploratory (EFA) and confirmatory (CFA) factor analyses of item-level responses.

Results: Results from the EFA suggested the BRIEF2 has fewer factors than would be suggested by the nine theoretically derived scales. While the theoretical CFA model, that omitted item-level information, demonstrated the best fit, when the item-level information was employed there was a decrement in model fit statistics and several extremely high loadings suggested scale-level redundancy in measurement. When the scales were omitted, and the items were loaded directly onto the indices, there was very little change in item-level factor loadings.

Conclusions: Findings suggest fewer than nine scales are needed and that clinical interpretation of the BRIEF2 may be more appropriate at the index, rather than scale, level.

Keywords

Executive function; attention; self-regulation; rating scales; exploratory factor analysis; confirmatory factor analysis

Introduction

Executive function is a multifactorial construct reflecting skills that develop over time, corresponding to the developmental maturation and myelination of frontal-striatal brain systems (Best & Miller, 2010; Brocki & Bohlin, 2004; Welsh & Pennington, 1988).

Executive functions are thought to reflect control processes required for goal-directed

CONTACT Lisa A. Jacobson, jacobson@kennedykrieger.org, Department of Neuropsychology, Kennedy Krieger Institute, 1750 E. Fairmount Avenue, Baltimore, MD 21231, USA.

Disclosure statement

No potential conflict of interest was reported by the authors.

actions and for managing affective reactions to guide appropriate behavior. The umbrella of executive functions also includes affective aspects of self-regulation (e.g. Prencipe et al., 2011), although multiple subcomponent skills are often included in the construct. Although findings appear sensitive to sample and measure characteristics, factor analytic studies of executive function tasks have typically found approximately three interrelated but dissociable factors, often labeled as inhibition, working memory, and flexibility/shifting (Cirino et al., 2018; Hughes, Ensor, Wilson, & Graham, 2009; Lehto, Juujärvi, Kooistra, & Pulkkinen, 2003; Miyake et al., 2000; Senn, Espy, & Kaufmann, 2004). Importantly, the structure of executive skills (e.g. number of factors derived from task performance) appears generally stable across childhood and adolescence (Huizinga, Dolan, & van der Molen, 2006).

Development of the BRIEF

There are several rating scales designed to assess children's executive behaviors as applied in everyday settings, one of which is the Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000). The BRIEF is a commonly used informant-report measure of executive behaviors in children and adolescents thought to reflect application of executive function skills to daily life tasks or "real world" situations (Gioia, Kenworthy, & Isquith, 2010). The original measure consisted of 86 items designed to measure eight "theoretically and empirically derived" (BRIEF Professional Manual, Gioia et al., 2000, p. 1) executive constructs, which in turn were conceptualized as representing two broad domains of executive function: behavioral regulation and metacognition. Items were assigned to nine domains of executive function, and thus, to nine scales, based upon the following steps: (1) agreement of ratings from "expert" neuropsychologists and the scale authors; (2) item-scale correlation analyses and domain-based exploratory principal factor analysis (e.g. nine analyses examining clustering of items within their assigned scales); and finally, (3) examination for redundant content via inter-correlation matrix across scales (Gioia et al., 2000). Ultimately, eight clinical scales were selected by the test authors, labeled Inhibit, Shift, Emotional Control, Initiate, Working Memory, Plan/Organize, Organization of Materials, and Monitor.

Exploratory factor analytic studies were conducted using principal factor analysis of the standardization sample ($N = 1419$; principal factor analysis) from the parent-report version of the BRIEF. Results presented in the test manual (Gioia et al., 2000) indicate that "solutions beyond two factors produced factors with single variables" (BRIEF Professional Manual, Gioia et al., 2000, p. 61) and the authors reported that due to "high correlations between variables and high communalities" (Gioia et al., 2000, p. 61), empirical methods for identifying underlying factors were "overridden in favor of theoretical considerations" (p. 61). Notably, exploratory factor analyses of parent ratings in the standardization sample were conducted at the *scale* rather than the *item* level, thus making the assumption that the items accurately reflect the scale to which they were assigned. Based on this methodology, the scales loading on the first factor (in order of strength of association) included Plan/Organize, Working Memory, Initiate, Organization of Materials, Monitor, and Inhibit. Of concern, the Inhibit scale was reported to load on both factors, with a stronger loading on the first factor; however, the test authors ultimately elected to assign this scale to the second

factor (defined by the Shift and Emotional Control scales). The measure scoring structure was subsequently designed such that three scales (Inhibit, Shift, and Emotional Control) comprise the Behavior Regulation Index (BRI) and the remaining five (Plan/Organize, Working Memory, Initiate, Organization of Materials, and Monitor) form the Metacognition Index (MI). Sums of both indices contribute to an overall Global Executive Composite (GEC).

Subsequent analyses of the parent version of the BRIEF by the scale authors suggested a potentially better fit for a three-factor solution (Gioia, Isquith, Retzlaff, & Espy, 2002) that reorganized the eight scales into nine by splitting the Monitor scale (Task-Monitor and Self-Monitor). Specifically, given this scale structure, examination of four competing models via maximum likelihood confirmatory factor analyses (CFA) within a mixed clinical sample of 374 children indicated better fit (relative to one-and two-factor models) for a three-factor model including latent factors labeled behavioral regulation (BRI), emotional regulation (ERI), and metacognition (MI; Gioia et al., 2002).

Independent factor analytic studies of the BRIEF

Several independent examinations of the parent-report BRIEF internal structure have been published in clinical samples, albeit with generally small sample sizes. Donders, DenBraber, and Vos (2010) conducted maximum likelihood CFA in a small sample ($N = 100$) of children undergoing evaluation due to traumatic brain injury, examining relative fit of one-factor and two-factor models based upon the original eight clinical scales. Within these data, a two-factor model (BRI, MI) fit the data better than a one-factor (GEC) model, although consistent with data from the test manual, the Inhibit scale exhibited significant cross loadings on the first factor (MI; Donders et al., 2010). Similarly, Slick, Lautzenhiser, Sherman, and Eyrl (2006) found that a two-factor solution (BRI, MI) was preferred to either a one-factor or three-factor (BRI, MI, ERI) solution in scale-level data from the BRIEF parent version in 80 children with epilepsy. Principal factor analysis (direct oblimin rotation) produced a single factor, although measures of goodness of fit favored a forced two-factor solution in these data. Substantial cross loadings were found for most scales, with the Inhibit and Monitor scales demonstrating the highest cross loadings (Slick et al., 2006). Among 181 youth with ADHD, Lyons Usher, Leon, Stanford, Holmbeck, and Bryant (2016) CFA analyses suggested that, using the clinical scales, a two-factor model fit data better than a one-factor model. Although the overall fit was unacceptable, it significantly improved by permitting the Monitor scale to load on both factors (BRI, MI) and to correlate with the Inhibit scale (Lyons Usher et al., 2016).

Two more recent examinations have compared the internal structure of the parent-report BRIEF based upon both the original eight clinical scales and the revised nine-scale arrangement. Using data gathered from parents of 158 Norwegian children, Egeland and Fallmyr (2010) found the clinical scales were generally well correlated with each other and thus queried whether the identified division into indices “is at all relevant” (p. 332). They subsequently examined the fit of the eight-scale and nine-scale arrangements, testing competing one-factor, two-factor, and three-factor models. Fit indices favored a three-factor model (BRI, ERI, MI) in the nine-scale version. Notably, even the best-fitting model showed

only a moderate fit to these data (e.g. three-factor solution Comparative Fit Index [CFI]=.96, Adjusted Goodness-of-Fit Index [AGFI]=.81, Root Mean Squared Error of Approximation [RMSEA]=.14). More recently, Fournet and colleagues (2015) similarly examined fit of serial CFA models in a larger sample of typically developing French children ($N=951$) used to provide normative data for the translated parent-report BRIEF. Consistent with findings from Egeland and Fallmyr, the nine-scale, three-factor solution (BRI, ERI, MI) in these data appeared to be the best-fitting model relative to nine-scale one-factor and two-factor models and eight-scale models with one to three latent factors (Fournet et al., 2015). Of critical importance is that all of the factor analytic analyses applied to the BRIEF were conducted at the scale level. This approach assumes all of the items uniquely load onto their assigned scales.

Development of the revised BRIEF2

Recently, the original BRIEF was modestly revised, reducing the number of items to 63, splitting the Monitor scale into two scales consistent with earlier analyses (Gioia et al., 2002; Slick et al., 2006), renaming the Metacognitive Index, and adding the Emotional Regulation Index. Specifically, the Behavior Rating Inventory of Executive Function, Second Edition (BRIEF2; Gioia, Isquith, Guy, & Kenworthy, 2015), now contains nine scales and three factors (consistent with Egeland & Fallmyr, 2010; Fournet et al., 2015), with the Inhibit and Self-Monitor scales comprising the Behavior Regulation Index (BRI), the Shift and Emotional Control scales comprising the new Emotional Regulation Index (ERI), and the Initiate, Working Memory, Plan/Organize, Task-Monitor, and Organization of Materials scales comprising the Cognitive Regulation Index (CRI; the renamed Metacognitive Index). The removal of the Shift and Emotional Control scales from the BRI and addition of the Self-Monitor scale to this index suggests that prior work examining the internal structure of the measure no longer applies and new examinations of the measure are needed to support utility of the revised measure and its proposed scale alignment and interpretive guidelines. Notably, the authors asserted that the goal of the revision was to “enhance its utility for clinical and research purposes without substantially altering the scales” (Gioia et al., 2015, p. 62), although given the high inter-correlations noted in independent examinations of the original measure, the uniqueness of the clinical scales has not yet been fully demonstrated.

To date, no independent examination of the BRIEF2 factor structure has been published. Based on the analyses that were conducted in the standardization sample of the BRIEF2 ($N=1400$), the items loaded adequately onto the nine clinical scales of the parent version; however, the nine-scale model did not fit as well as the three-factor (index level) model (BRIEF2 Professional Manual, Gioia et al., 2015). Given the use of the original measure with a variety of clinical populations (e.g. Donders et al., 2010; Gioia et al., 2002; Mahone et al., 2002; Mahone, Zabel, Levey, Verda, & Kinsman, 2002; Slick et al., 2006), and initial data suggesting potential utility of the BRIEF2 in discriminating among clinically referred youth (Jacobson, Pritchard, Koriakin, Jones, & Mahone, 2017), independent examination of the internal structure of the BRIEF2 in large clinical samples is sorely needed. For a measure to demonstrate clinical utility, there must be an accurate understanding of the measure's

internal structure. Otherwise, the clinical interpretations produced from that information will be erroneous.

In sum, the original BRIEF and its revision in the BRIEF2 reflect a measure that has been theoretically derived. A number of independent factor analytic studies have resulted in reshuffling or reinterpretation of the scale to index level of the measure. However, most – if not all – of the published factor analytic investigations, including those presented in the manual, were completed at the scale level, assuming that the item to scale correspondence is appropriate.

The overarching goal of this study was to test whether the theoretical derivation of the executive function construct underlying the BRIEF2 is empirically supported when the full hierarchical structure of the measure (item-scale-index) is appreciated, using methods that are agnostic to (exploratory factor analysis) and driven by (confirmatory factor analysis) theory. Specifically, the present study asked the following questions of the revised measure: (1) At the item level, how many latent factors best account for the measure's variance (1a) across the entire BRIEF2 as well as (1b) within the 3 indices? (2) Does the higher-order nine-scale, three-factor model proposed in the BRIEF2 manual provide an adequate fit to an independent clinical sample? And, (3) how does the model proposed in the test manual compare to (3a) a model that appreciates the full structure of the BRIEF2, including the item-level loadings, and (3b) a model that completely omits the scales by loading the items directly to the overall indices? To our knowledge, this is the first study to address these questions. Again, none of the prior published work examining factor structure of the original measure used item-level data; examination at the scale level presented in the test manual (Gioia et al., 2015) suggests the nine-scale model did not fit as well as the three-factor (index level) model. Thus, the working hypothesis for the current analyses proposed redundancy at the scale level.

Method

Sample

Data were gathered from 5212 youth who were seen for psychological or neuro-psychological assessment at a large, urban academic medical center. Youth ranged in age from 5 to 18 years (mean = 10.5, SD = 3.3), and 66% were male. A little less than half were Caucasian (47%), with the remainder being mostly Black/African American (28%). Only a few identified as Hispanic (2%), Asian (2%), or Other (5%); 16% were missing this variable. Approximately one-third (33%) of the sample was receiving medical assistance (or publicly funded insurance). With regard to reason for referral, the most common primary billing diagnoses for the sample were: Attention-Deficit/Hyperactivity Disorder (51%), adjustment disorders (7%), anxiety disorders (5%), encephalopathy (4%), oncologic diseases (4%), epilepsy (3%), and Oppositional Defiant Disorder or conduct problems (2%). As part of routine clinical practice, data from item-level scoring software programs for informant rating scales are extracted into a clinical database and are securely maintained by the hospital's Information Systems department. Following approval from the hospital Institutional Review Board, which includes review for compliance with ethical standards, the clinical database was queried, and a limited dataset was constructed of patients between 5 and 18 years of age

for whom complete item-level responses were available on the BRIEF2; 2.3% ($n = 126$) were missing item-level data. There were no exclusionary criteria beyond complete data on the measure of interest.

Using the full sample, the participants were randomly split into two subsamples of equal size (both $n = 2606$). The first of the two samples was used for initial analyses and the second for replication or cross-validation, as explained in the analysis plan below. No significant differences were found between the samples in terms of demo-graphics, billing diagnoses, or BRIEF scores, suggesting the randomization was successful ($p < .05$). The study design was cross-sectional and all observations were independent.

Measures

Behavior rating inventory of executive function, second edition, parent form (BRIEF2; Gioia et al., 2015)—The BRIEF2 is a revision of the BRIEF, a rating scale assessing everyday behaviors reflecting executive functions across the school-age span (ages 5–18). The BRIEF2 Parent-report form contains 63 items, 60 of which fall within nine theoretically derived clinical scales: Inhibit, Self-Monitor, Shift, Emotional Control, Initiate, Working Memory, Plan/Organize, Task-Monitor, and Organization of Materials (three items are validity measures and are not included in the clinical scales). The nine clinical scales form three Composite scores – the Behavior Regulation Index (BRI), Emotion Regulation Index (ERI), and Cognitive Regulation Index (CRI) – as well as the overall Global Executive Composite summary score. All Parent Form coefficient alpha values for index scores were reported to fall above .90, with coefficients for individual scales ranging from .80 (Monitor) to .91 (Emotional Control) in the standardization sample (Gioia et al., 2015, p. 101). The manual reported test-retest reliability estimates from a sample of 163 parent ratings; the mean test-retest correlation coefficient for clinical scales was .79 (range=.67–.92) over an average time interval of 2.9 weeks (Gioia et al., 2015, p. 111). For the present study, item-level data used for analysis were obtained via the original BRIEF parent rating form. Since all of the item content for the revised BRIEF2 is derived from existing items on the original BRIEF, the 60 clinical items present on the BRIEF2 were extracted from data gathered on the original BRIEF.

Analysis plan—The first step in the analysis was to examine the factor structure, at the item level, of the entire BRIEF2 using principal factor exploratory factor analysis (EFA). This agnostic, data-driven approach seeks to explain the relationship among a set of observed items using fewer, unobserved latent variables or constructs. EFA was employed specifically for the purposes of determining the optimal number of factors to retain. To accomplish this, a series of metrics was employed that included eigenvalues, visual inspection of a scree plot, and proportion of variance explained. The solution was not rotated since the goal was solely to identify the number of factors to retain – the first step in EFA – rather to understand the relations between items and factors. Typically, factors with eigenvalues < 1 are not retained (Ledesma & Valero-Mora, 2007), along with factors that explain a small proportion of variance.

In conjunction with the EFA, parallel analysis was employed to better understand the number of factors to retain. Parallel analysis is a statistical approach that generates a series of eigenvalues using a simulated dataset of random or uncorrelated variables mirroring the observed dataset in terms of sample size and number of variables (Ledezma & Valero-Mora, 2007). The cutoff, in terms of the number of factors to retain, is determined when the observed eigenvalue is greater than the eigenvalue based on the parallel analysis.

The second step in the analysis was to repeat the EFA within each index. The goal of this analysis was to better understand if the data supported the purported number of scales within each index. All EFA analyses were conducted in STATA 15.0 (StataCorp, 2017), including use of the *fapara* package for the parallel analysis, using a polychoric matrix.

The third and final analysis examined different models, specified a priori, of the BRIEF2 executive constructs using confirmatory factor analysis (CFA). In contrast to EFA, this top-down approach investigates the fit for a specific, or pre-specified latent variable structure underlying the data. A total of three models were evaluated (see Figure 1). The first “theoretical” model employed the previously published scale-index solution, omitting item-level indicators, where the nine scales served as the observed variables for the three higher factor indices. The three indices were correlated, along with their residual variances, as indicated in the manual (Gioia et al., 2015, p. 114). The second or “full” model loaded the items to their respective scales *and* each scale to its accompanying index. The third approach, now termed the “item to index” (ITI) model, omitted the scales and directly loaded the items onto the three higher-order indices. This model was developed for the purposes of contrasting the mean item-level loadings, as well as the fit statistics described below, between the ITI and Full models. Information about the fit of these models is novel since the model proposed in the BRIEF2 manual (the theoretical model) has not been validated in an independent clinical sample and previous research has omitted the lower order factor structure of the BRIEF2 regarding the item to scale factor loading. It was hypothesized that the inclusion of the intermediary scales (between the items and indices) in the Full model would better organize the data, which in turn would be reflected in higher item loadings and stronger fit statistics for the Full compared to the ITI model.

The CFA models were compared in two ways. First, mean item-level loading coefficients were compared across two models (Full and ITI). This analysis omitted the theoretical model since that approach did not include item-level information. The purpose of examining item-level loadings was to better understand if the scales were more proximal to the measurement of the items when compared to the indices. It was hypothesized that item-level factor loadings would be larger in the Full, compared to the ITI, model since the scales are assumed to organize the items into smaller and more organized constructs than the larger overall indices. Mean differences across item-level loading coefficients were interpreted as small (.10), medium (>.10 to .19), and large (.20). Second, a series of fit statistics were used to better understand how well the models fit the observed data. Fit statistics used in the present analysis included the Chi-square test (χ^2), Root Mean Squared Error of Approximation (RMSEA), Tucker Lewis Index (TLI), and Comparative Fit Index (CFI). A RMSEA of <.08 is considered an acceptable fit, and CFI/TLI values of >.90 are acceptable, with >.95 being optimal (Hu & Bentler, 1999). Non-significant or lower χ^2 values are

considered optimal since they indicate smaller differences between the expected and observed covariance matrices. These tests are almost always significant, however, when N is large (Hu & Bentler, 1999).

The CFI, TLI, and RMSEA values can be compared across models. However, the χ^2 values cannot be used to compare the theoretical with the other three models since it is not nested. On the other hand, the χ^2 difference test was used to examine if there were significant differences in fit between the ITI and the Full model. All CFA analyses were conducted in MPLUS 7.0 (Muthén & Muthén, 1998–2012) using poly-choric matrices.

Finally, a cross-validation strategy was employed. This procedure involved randomly splitting the sample in half and running the EFA and CFA analyses on the two separate samples. The purpose of cross-validation was to replicate the findings in an independent sample. Overall, there were no missing data, except for race.

Results

Exploratory factor analyses

BRIEF2 total scale—Table 1 displays the statistics from the EFA, performed on the 60 BRIEF2 clinical items, for the first 10 factors generated. The observed eigenvalues for the first six factors were all >1.0 , with the first two factors being >6.0 . Shown in Figure 2, the scree plot demonstrates the steep drop-off in observed eigenvalues after the first two factors. While factors three through six had Eigenvalues of >1.0 , little additional variance was explained (6% or less) by the remaining factors. With regard to parallel analysis (see Table 1 and Figure 2), the estimated values were extremely low (.32 or less) and the observed values did not become smaller than the parallel analysis eigenvalues until the 32nd factor. Results were replicated, or nearly identical, across the cross-validation samples (see Table 1).

BRIEF2 indices—Table 2 and Figure 3 display the results from the EFA, parallel analysis, and scree plots conducted separately across each of the BRIEF2 indices, respectively. The results were somewhat similar across indices. Specifically, the proportion of variance explained was high (75% or greater) and the eigenvalue was large (5) for the first factor across each of the indices. For the BRI, no factor eigenvalue, other than the first, was >1 . For the CRI, the eigenvalues were around 1.4 for the second and third factor, respectively. The proportions of variance explained for the second and third factor of the CRI were equal and somewhat small (7%). For the ERI, the second factor was just over 1 for the second factor, which explained 10–12% of the variance depending on the sample. Visual inspection of the scree plots is shown in Figure 3, all of which demonstrate a large drop off in eigenvalues after the first factor. Taken together, these data support at minimum three factors and at most six factors in total. It is worth noting that the sum of the variance shown in Tables 1 and 2 is greater than 100. This is due to the negative variances that are produced for the lower order factors, which are not shown for the sake of brevity (e.g. this would require displaying up to 60 factors in Table 1, since there are 60 items).

Confirmatory factor analyses

Item and scale loadings—The average item-level loading coefficients for the Full and ITI models (see Figure 1 for depiction of the models) are shown in Table 3. In the first sample, when comparing the Full versus the ITI models, two of the nine coefficients had a difference $>.1$ (Initiate and Emotional Control). In the second sample, again only two of the nine mean item-level loadings had a difference $>.1$ (Initiate, Self-Monitor). The minimal difference in item-level loading coefficients can be seen in the overall averages, where the difference in ITI vs. Full model loadings was only .07 across samples. Shown in Table 4, differences in the scale to index loadings (calculated as the mean overall differences) were slightly higher (.07) for the full model, compared to the theoretical model, across samples. Notably, scale to index loadings were $>.90$ for five of nine scales when examining the Full model.

Fit statistics—Table 5 displays the fit statistics for each of the CFA models. For the theoretical model, the CFI and TLI values were high and the RMSEA was just below the cutoff of .08 (suggesting an adequate fit). The Full model demonstrated the second best fit, with acceptable CFI and TLI values and an identical RMSEA value to the theoretical model. The ITI model demonstrated acceptable RMSEA values, but the CFI/TLI values were poorest and suggested an ill-fitting model. The χ^2 difference test demonstrated the Full model fit the data significantly better than the ITI model ($p<.001$).

Discussion

The present study employed both exploratory and confirmatory methods to examine the factor structure of the revised BRIEF2 in a large, clinically referred sample. The use of such a large sample permitted a systematic, multiple step analytic approach and model replication of each step across equivalent sub-samples.

The first set of analyses examined the number of latent factors that best accounted for the measure's variance at the item level, across the entire BRIEF2 as a whole as well as within the three hypothesized indices (e.g. questions 1a and 1b). Results of item-level EFA across samples suggested that the items appear to primarily load onto three factors, with not more than six factors accounting for the majority of item-level variance. EFA within indices (e.g. scale-level analyses) likewise suggested not more than five to six factors (e.g. at most one factor within the BRI, two factors within the ERI, and two to three factors within the CRI). The parallel analyses were inconclusive as they suggested many more factors than would be reasonable to retain. Overall, results of the EFA suggest that there appear to be fewer factors within the measure than were theorized in the test manual; in other words, the variance at the item-level within a clinical sample appears best accounted for by no more than five to six factors, rather than nine (i.e. the number of theoretically identified scales).

Subsequent analyses examined competing higher-order CFA models, testing whether the theoretical model proposed in the test manual (e.g. the nine-scale, three-factor model) provides an adequate fit within an independent clinical sample (e.g. question 2), as well as how the fit of that proposed model compares to two competing models (e.g. the ITI and Full models; questions 3a and 3b) that acknowledge the full structure of the BRIEF2 by including

the (60) individual clinical items. The CFA results indicated that the Full model showed an acceptable fit, with a distinct difference from the theoretical model in that it incorporates item-to-scale loadings. Although fit statistics were best for the theoretical model, its fit may be somewhat inflated due to omission of item-level variance (e.g. item-to-scale loadings). This assertion is supported by the decrement in the fit indices, notably for the CFI and TLI, between the theoretical and Full model. The Full model is potentially a better representation of the existing data and the corresponding underlying structure of the measure, at least as observed within a clinical sample. Fit of the ITI model (which omits the scales altogether) was poorest.

With regard to BRIEF2 scale-level structure within a clinical sample, the evidence is variable but suggests the internal structure likely reflects fewer than the nine proposed scales. It is notable that the Full model, which includes item-to-scale variance, fits the data better than the ITI model, which omits the scales altogether. The significant chi-square test for nested models also provides some evidence for the utility of the scales. Upon closer inspection, however, item loading coefficients were very similar across the Full and ITI models. In fact, differences between the models, in terms of item as well as scale-level loadings, were minimal (a difference of only .07, on average). If the nine BRIEF2 scales did add meaningful organization among the items, these loadings would be expected to decrease substantially when the scales are omitted from the model (e.g. in the ITI model). Additionally, the scale-to-index loadings were very high for the hitherto unexplored Full model, with this pattern particularly noticeable for the Initiate, Working Memory, and Plan/Organize subscales (with loadings of .99, .92, and .97, respectively). The consistently high loadings suggest there may be redundancy in measurement at the scale level. In other words, loadings this high mean that the subscales show nearly complete overlap with CRI and there is little unique contribution to CRI from each individual scale.

The EFA results, item to index loadings, and item to scale loadings fail to fully support the theoretical structure of the BRIEF2's nine-scale arrangement within this clinical sample. Although the revised measure's structure appears to have expanded from eight to nine scales in response to prior EFA studies (e.g. splitting the Monitor scale; Gioia et al., 2002; Slick et al., 2006), and is described in the manual as representing an "improved internal structure" (Gioia et al., 2015, p. 3), the present study suggests that this degree of complexity in the scale-level structure may not be the best representation of the data – at least, within a clinical sample. As noted previously, a better representation of the data would be a more simple structure or fewer factors (i.e. the 5–6 factors at max suggested by the EFA). Importantly, the BRIEF2 is a clinical tool, designed to enable identification of "different profiles of executive function strengths and weaknesses" (Gioia et al., 2015, p. 6) and "problems with behaviors associated with specific domains of executive functioning" (p. 33). As such, it is critical that interpretations made for the purposes of clinical decision making and intervention planning be supported by the empirical evidence, including that obtained from clinical samples.

In developing clinical measurement tools, theory and associated clinical judgment represent a useful starting point. However, theory is often dynamic and evolves as new empirical evidence is obtained; interpretation of measurement tools must validly represent the existing data. Findings from the present study question whether the BRIEF2 demonstrates a

reasonable degree of valid differentiation among executive sub-skills, an assertion that may be further supported by the weak association between the original BRIEF scales and specific performance based measures of executive function (e.g. McAuley, Chen, Goos, Schachar, & Crosbie, 2010). Although it is possible that the measure's internal structure may differ between standardization and clinical samples, the present findings suggest that a more parsimonious, data-driven reorganization of the BRIEF2 scales is warranted in use with clinically referred youth. Until then, clinical interpretation of the individual BRIEF2 scales, particularly those with high levels of redundancy in the CRI, should be tempered.

This study should be interpreted in light of its strengths and limitations. With regard to strengths, the sample was large, permitting bifurcation into two comparable samples for model replication and confirmation. This study also provides several novel insights into the newest version of a widely employed clinical measure. Limitations of the present study include lack of detailed information regarding socioeconomic status or medication status of participants. Furthermore, item responses were taken from administration of the original BRIEF and transposed (by the test publisher's staff) into BRIEF2 items and scales, rather than derived from administration of the BRIEF2 itself. Notably, however, there are no differences in measure instructions to raters or item wording between the two versions. Finally, it is also important to note that findings are specific to a clinically referred sample and may differ in community samples; as noted previously, data supporting the proposed BRIEF2 structure (e.g. examination of the Full model) within standardization or other community samples have not been provided in the manual. However, given that the measure was revised to "enhance its utility for clinical and research purposes" (Gioia et al., 2015, p. 62), evidence supporting the proposed scale arrangement in clinical samples is most relevant within a clinical setting.

Further work is needed to replicate these findings, especially in non-clinical samples. If these findings are replicated, investigators might consider ways to reduce the length of the BRIEF2 with the goal of reducing rater burden, as there may be evidence for redundant items, particularly within the CRI. Examination of the utility of the revised measure for clinical decision-making and intervention planning is also needed, given the proposed purpose of the measure. Going forward, replication of these analyses is needed within other large, clinical samples to ensure that providers have the most valid and appropriate measures for clinical decision-making.

Acknowledgments

The authors would like to gratefully acknowledge the assistance of Peter Isquith, PhD, and Jennifer Greene at PAR, Inc., in conversion of BRIEF2 scores from original BRIEF item-level data.

Funding

This work was supported in part by the Intellectual and Developmental Disabilities Research Center under grant U54 HD079123 and the Johns Hopkins University School of Medicine Institute for Clinical and Translational Research, an NIH/NCRR CTSA Program.

References

- Best JR, & Miller PH (2010). A developmental perspective on executive function. *Child Development*, 81(6), 1641–1660. 10.1111/j.1467-8624.2010.01499.x [PubMed: 21077853]
- Brocki KC, & Bohlin G (2004). Executive functions in children aged 6 to 13: A dimensional and developmental study. *Developmental Neuropsychology*, 26(2), 571–593. 10.1207/s15326942dn2602_3 [PubMed: 15456685]
- Cirino PT, Ahmed Y, Miciak J, Taylor WP, Gerst EH, & Barnes MA (2018). A framework for executive function in the late elementary years. *Neuropsychology*, 32(2), 176–189. 10.1037/neu0000427 [PubMed: 29528682]
- Donders J, DenBraber D, & Vos L (2010). Construct and criterion validity of the Behaviour Rating Inventory of Executive Function (BRIEF) in children referred for neuropsychological assessment after paediatric traumatic brain injury. *Journal of Neuropsychology*, 4(2), 197–209. 10.1348/174866409X478970 [PubMed: 19930791]
- Egeland J, & Fallmyr O (2010). Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF): Support for a distinction between emotional and behavioral regulation. *Child Neuropsychology*, 16(4), 326–337. 10.1080/09297041003601462 [PubMed: 20209415]
- Fournet N, Roulin J, Monnier C, Atzeni T, Cosnefroy O, Le Gall D, & Roy A (2015). Multigroup confirmatory factor analysis and structural invariance with age of the Behavior Rating Inventory of Executive Function (BRIEF) – French version. *Child Neuropsychology*, 21(3), 379–398. 10.1080/09297049.2014.906569 [PubMed: 24754365]
- Gioia GA, Isquith PK, Guy SC, & Kenworthy L (2000). BRIEF: Behavior Rating Inventory of Executive Function Lutz, FL: Psychological Assessment Resources, Inc.
- Gioia GA, Isquith PK, Guy SC, & Kenworthy L (2015). BRIEF2: Behavior Rating Inventory of Executive Function (2nd ed.). Lutz, FL: Psychological Assessment Resources, Inc.
- Gioia GA, Isquith PK, Retzlaff PD, & Espy KA (2002). Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF) in a clinical sample. *Child Neuropsychology*, 8(4), 249–257. 10.1076/chin.8.4.249.13513 [PubMed: 12759822]
- Gioia GA, Kenworthy L, & Isquith PK (2010). Executive function in the real world: BRIEF lessons from Mark Ylvisaker. *Journal of Head Trauma Rehabilitation*, 25(6), 433–439. 10.1097/HTR.0b013e3181fbc272 [PubMed: 21076244]
- Hu L, & Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. 10.1080/10705519909540118
- Hughes C, Ensor R, Wilson A, & Graham A (2009). Tracking executive function across the transition to school: A latent variable approach. *Developmental Neuropsychology*, 35(1), 20–36. 10.1080/87565640903325691
- Huizinga M, Dolan CV, & van der Molen MW (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, 44(11), 2017–2036. 10.1016/j.neuropsychologia.2006.01.010 [PubMed: 16527316]
- Jacobson LA, Pritchard AE, Koriakin TA, Jones KE, & Mahone EM (2017). Initial examination of the BRIEF2 in clinically referred children with and without ADHD symptoms. *Journal of Attention Disorders*. Online First doi: 10.1177%2F1087054716663632.
- Ledesma RD, & Valero-Mora P (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12(2), 1–11.
- Lehto JE, Juujärvi P, Kooistra L, & Pulkkinen L (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology*, 21(1), 59–80. 10.1348/026151003321164627
- Lyons Usher AM, Leon SC, Stanford LD, Holmbeck GN, & Bryant FB (2016). Confirmatory factor analysis of the Behavior Rating Inventory of Executive Functioning (BRIEF) in children and adolescents with ADHD. *Child Neuropsychology*, 22(8), 907–918. 10.1080/09297049.2015.1060956 [PubMed: 26156531]

- Mahone EM, Cirino PT, Cutting LE, Cerrone PM, Hagelthorn KM, Hiemenz JR, ... Denckla MB (2002). Validity of the behavior rating inventory of executive function in children with ADHD and/or Tourette syndrome. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 17(7), 643–662. 10.1093/arclin/17.7.643 [PubMed: 14591848]
- Mahone EM, Zabel TA, Levey E, Verda M, & Kinsman S (2002). Parent and self-report ratings of executive function in adolescents with myelomeningocele and hydrocephalus. *Child Neuropsychology*, 8(4), 258–270. 10.1076/chin.8.4.258.13510 [PubMed: 12759823]
- McAuley T, Chen S, Goos L, Schachar R, & Crosbie J (2010). Is the behavior rating inventory of executive function more strongly associated with measures of impairment or executive function? *Journal of the International Neuropsychological Society*, 16(03), 495–505. 10.1017/S1355617710000093 [PubMed: 20188014]
- Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, & Wager TD (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100. 10.1006/cogp.1999.0734 [PubMed: 10945922]
- Muthén LK, & Muthén BO (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Prencipe A, Kesek A, Cohen J, Lamm C, Lewis MD, & Zelazo PD (2011). Development of hot and cool executive function during the transition to adolescence. *Journal of Experimental Child Psychology*, 108(3), 621–637. 10.1016/j.jecp.2010.09.008 [PubMed: 21044790]
- Senn TE, Espy KA, & Kaufmann PA (2004). Using path analysis to understand executive function organization in preschool children. *Developmental Neuropsychology*, 26(1), 445–464. 10.1207/s15326942dn2601_5 [PubMed: 15276904]
- Slick DJ, Lautzenhiser A, Sherman EMS, & Eyrl K (2006). Frequency of scale elevations and factor structure of the Behavior Rating Inventory of Executive Function (BRIEF) in children and adolescents with intractable epilepsy. *Child Neuropsychology*, 12(3), 181–189. 10.1080/09297040600611320 [PubMed: 16837394]
- StataCorp. (2017). *Stata Statistical Software: Release 15* College Station, TX: StataCorp LLC.
- Welsh MC, & Pennington BF (1988). Assessing frontal lobe functioning in children: Views from developmental psychology. *Developmental Neuropsychology*, 4(3), 199–230. 10.1080/87565648809540405

Figure 1.

Competing BRIEF2 models examined via Confirmatory Factor Analyses.

Note. Models included all items, scales, and indices as appropriate; the BRI and associated components are shown in the figure as an exemplar/for simplicity; item subscripts represent range of items assigned to specific scales; BRI: Behavior Regulation Index; In: Inhibit scale; SM: Self-Monitor scale.

Figure 2.
Scree Plots from Exploratory Factor Analyses and Parallel Analysis of the BRIEF2 within two randomly selected subsamples.

Samples 1 and 2

Figure 3.

Scree Plots of the BRIEF2 Indices within two randomly selected subsamples.

Note. BRI: Behavior Regulation Index, ERI: Emotional Regulation Index, CRI: Cognitive Regulation Index. Solid line represents factor analysis, dashed line represents parallel analysis

Table 1.

Exploratory factor analysis of the BRIEF 2 items.

Sample 1 (N = 2606)				Sample 2 (N = 2606)			
Factor	Observed EV	Parallel analysis EV	Proportion variance	Factor	Observed EV	Parallel analysis EV	Proportion variance
1	24.15	.32	59	1	24.60	.34	59
2	6.31	.30	15	2	6.06	.30	15
3	2.30	.26	06	3	2.32	.29	06
4	1.70	.26	04	4	1.61	.27	04
5	1.47	.25	04	5	1.39	.25	03
6	1.10	.23	03	6	1.10	.24	03
7	.91	.22	02	7	.93	.23	02
8	.79	.21	02	8	.88	.22	02
9	.71	.20	02	9	.69	.20	02
10	.66	.19	02	10	.65	.19	02

Note: EV = Eigenvalue.

Table 2.

Exploratory factor analysis of items within each of the BRIEF2 indices: BRI, ERI, CRI.

BRI				ERI			CRI		
Smaple 1 (N=2606)									
	EV	PA	%VE	EV	PA	%VE	EV	PA	%VE
1	7.23	.14	91	9.05	.15	87	15.1	.20	75
2	.79	.08	10	1.08	.09	10	1.49	.19	07
3	.21	.07	3	.61	.07	6	1.31	.16	07
4	.16	.05	2	.13	.07	1	.99	.14	05
5	.04	.03	<1	.10	.07	1	.73	.13	04
Smaple 2 (N=2606)									
	EV	PA	%VE	EV	PA	%VE	EV	PA	%VE
1	7.39	.14	91	8.84	.17	86	15.44	.23	76
2	.73	.08	10	1.23	.14	12	1.36	.21	7
3	.20	.07	3	.61	.07	6	1.35	.20	7
4	.12	.05	2	.16	.06	2	.97	.17	5
5	.04	.03	2	.10	.06	1	.74	.15	4

Note. BRI: Behavior Regulation Index; CRI: Cognitive Regulation Index; ERI: Emotional Regulation Index; EV: Eigenvalue; %VE: Proportion Variance Explained, PA: Parallel Analysis.

Table 3.

Average item loading coefficients.

	Full model		Item to Index (ITI) model	
	1	2	1	2
Sample				
Inhibit	.81	.82	.78	.80
Self-monitor	.84	.85	.80	.81
Shift	.75	.75	.73	.71
Emotional control	.84	.84	.81	.81
Initiate	.69	.70	.68	.68
Working memory	.78	.78	.74	.74
Plan/organize	.71	.71	.69	.69
Task-monitor	.79	.80	.64	.67
OrgMaterials	.81	.82	.70	.70
Overall mean	.78	.79	.73	.73

Note. OrgMaterials: Organization of Materials.

Table 4.

Average scale to index loading coefficients.

	Full model		Theoretical model	
	Sample 1	Sample 2	Sample 1	Sample 2
BRI				
Inhibit	.82	.84	.78	.79
Self-Monitor	.96	.96	.86	.88
ERI				
Shift	.82	.94	.87	.85
Emotional Control	.96	.81	.80	.80
CRI				
Initiate	.99	.99	.85	.85
Working Memory	.92	.92	.84	.85
Plan/Organize	.97	.98	.89	.90
Task-Monitor	.75	.78	.69	.72
OrgMaterials	.78	.77	.75	.75
Overall Mean	.88	.89	.81	.82

Note. BRI: Behavior Regulation Index; ERI: Emotional Regulation Index; CRI: Cognitive Regulation Index; OrgMaterials: Organization of Materials.

Table 5.

Results of confirmatory factor analysis and replication in two subsamples.

Model	df	χ^2	CFI	TLI	RMSEA
<i>Sample 1</i>					
Theoretical	33	331.90	.98	.96	.07
Item to Index	183	24359.33	.86	.85	.07
Full Model	195	16562.10	.91	.90	.06
<i>Sample 2</i>					
Theoretical	33	276.80	.98	.97	.07
Item to Index	183	24167.26	.86	.85	.07
Full Model	195	16504.04	.91	.90	.06

Note. CFI: Comparative Fit Index; TLI: Tucker Lewis Index; RMSEA: Root Mean Squared Error of Approximation.