HHS Public Access

Author manuscript

Child Neuropsychol. Author manuscript; available in PMC 2020 February 01.

Published in final edited form as:

Child Neuropsychol. 2019 February ; 25(2): 198–216. doi:10.1080/09297049.2018.1433156.

Item Response Theory Analyses of the Delis-Kaplan Executive Function System Card Sorting Subtest

Mercedes Spencer,

Vanderbilt University, 230 Appleton Place, Vanderbilt University, Nashville, TN 37203; (615) 322-8240

Sun-Joo Cho, and

Vanderbilt University, PMB 407817, 2301 Vanderbilt Place, Vanderbilt University, Nashville, TN 37240; (615) 322-8409

Laurie E. Cutting

Vanderbilt University, 230 Appleton Place, Vanderbilt University, Nashville, TN 37203; (615) 322-8240

Abstract

In the current study, we examined the dimensionality of the 16-item Card Sorting subtest of the Delis-Kaplan Executive Functioning System assessment in a sample of 264 native English-speaking children between the ages of 9 and 15 years. We also tested for measurement invariance for these items across age and gender groups using item response theory (IRT). Results of the exploratory factor analysis indicated that a two-factor model that distinguished between verbal and perceptual items provided the best fit to the data. Although the items demonstrated measurement invariance across age groups, measurement invariance was violated for gender groups, with two items demonstrating differential item functioning for males and females. Multigroup analysis using all 16 items indicated that the items were more effective for individuals whose IRT scale scores were relatively high. A single-group explanatory IRT model using 14 non-differential item functioning items showed that for perceptual ability, females scored higher than males and that scores increased with age for both males and females; for verbal ability, the observed increase in scores across age differed for males and females. The implications of these findings are discussed.

Keywords

item response theory; exploratory factor analysis; executive function; differential item functioning; multigroup analysis

Executive function (EF) is an umbrella term that encompasses multiple cognitive skills believed to be integral for planning, problem solving, and maintaining/updating information (Miyake et al., 2000; Miyake & Friedman, 2012). Specific cognitive skills thought to comprise EF include working memory, metacognition, inhibitory control, cognitive

Please address correspondence to the corresponding author at laurie.cutting@vanderbilt.edu.

Disclosure of interest: The authors report no conflicts of interest.

flexibility, and attention shifting (Gioia, Isquith, Retzlaff, & Espy, 2002; St Clair-Thompson & Gathercole, 2006). The importance of EF emerges from studies showing that EF skills predict future social functioning and academic performance (Biederman et al., 2004; Blair & Razza, 2007; Brock, Rimm-Kaufman, Nathanson, & Grimm, 2009; Clark, Pritchard, & Woodward, 2010; Eisenberg, Fabes, Guthrie, & Reiser, 2000; Miller & Hinshaw, 2010). More specifically, deficits in EF have been implicated in reading and math difficulties (Blair & Razza, 2007; Bull, Espy, & Wiebe, 2008; Locascio, Mahone, Eason, & Cutting, 2010; Sesma, Mahone, Levine, Eason, & Cutting, 2009) and weaknesses in social competence (Razza & Blair, 2009). Moreover, the observed correlated deficits between EF and academic performance likely arise from the fact that deficits in EF are related to attention-deficit hyperactivity disorder (e.g., Willcutt, Doyle, Nigg, Farone, & Pennington, 2005), which has been implicated in both reading and math disabilities (Mayes, Calhoun, & Crowell, 2000; Willcutt et al., 2013).

In light of the growing interest in the role that EF plays in cognitive development, several measures of EF have been created. In particular interest to the current investigation is the Delis-Kaplan Executive Function System (D-KEFS; Delis, Kaplan, and Kramer, 2001a), which is a standardized measure of EF. The D-KEFS has been utilized across both clinical and educational settings (e.g., Altemeier, Jones, Abbott, & Berninger, 2006; Baron, 2004; Heled, Hoofien, Margalit, Natovich, & Agranov, 2012; Homack, Lee, & Riccio, 2005; Kiefer & Tranel, 2013; Parrish et al., 2007; Savla et al., 2010, 2011; Swanson, 2005), and in some instances, has also been used as a tool for identifying individuals who may be at risk for executive dysfunction (e.g., Strong, Tiesma, & Donders, 2010). In the current investigation, we focused on the Card Sort subtest of the D-KEFS. This subtest contains both verbal and perceptual stimuli and requires participants to categorize card sets based on various different categorization rules. The Card Sort subtest purports to measure multiple aspects of EF, including verbal and non-verbal problem solving, cognitive flexibility, and inhibition (Swanson, 2005; Delis et al., 2001b).

Development of the D-KEFS

The D-KEFS (Delis et al., 2001a) is a norm-referenced assessment that is based on a nationally-representative stratified sample of 1,750 individuals between the ages of 8 and 89 and consists of nine subtests measuring different aspects of verbal and nonverbal EF (Delis et al., 2001b). The standardization sample was representative of all major geographical regions across the United States, contained relatively equal proportion of males and females across most age groups, and contained proportions of racial/ethnic groups that were stratified based on the 2000 U.S. Census (for exclusion criteria, see Delis et al., 2001b). In an effort to identify whether age effects were present across D-KEFS, test developers conducted linear and nonlinear regression analyses for each subtest. There was evidence to suggest that age effects were present across multiple subtests of the D-KEFS, including the Card Sort subtest; performance peaked later for verbal concept formation relative to nonverbal (i.e., perceptual) concept formation (see Delis et al., 2001b).

The Card Sort subtest of the D-KEFS is based on the California Card Sorting Test (for a review, see Delis et al., 2001b). This subtest consists of 16 items that require participants to

sort cards into two sets based on verbal and nonverbal (i.e., perceptual) categorization rules. Following each sort, the participant is then asked to describe the sorting rules used for each set (free sort condition) based on any of the eight acceptable categorization rules for each of the two card sets (see Delis, Kaplan, & Kramer, 2001a for details). Following this, the examiner then categorizes the cards based on these eight categorization rules and requires the participant to identify which of the eight rules is being applied for each set (recognition condition). We chose to focus on the Card Sort subtest because card sorting tasks are prominently used measures of EF (e.g., Wisconsin Card Sorting Test; Romine et al., 2004).

Reliability for D-KEFS was established using a smaller sample relative to the original standardization sample for alternate-forms reliability (n = 286) and test-retest reliability (n= 101; Delis et al., 2001b). For the free sort version of the Card Sort subtest, internal consistency was based on the correlation between card set 1 and card set 2 and was estimated using the Spearman-Brown correction of the Pearson correlation coefficient (Delis et al. 2001b). Alternate-forms reliability was estimated using correlations based on alternate forms that were calibrated using Angoff's (1984) design II-B linear calibration (Delis et al. 2001b). According to test developers, internal consistency was moderate to high across all ages for the free sorting condition (r = .55 – .86) and sort recognition condition (r = .62 – .81); internal consistency for 8- to 15-year-olds ranged from .62 to .74 for sort recognition and .55 to .82 for the free sorting condition. Alternate-form reliability was moderate across the free sorting (r = .59), sort recognition (r = .72) and total confirmed sorts conditions (r = .60) for all ages. Test-retest reliability was moderate when averaged across all ages for the free sort condition (r = .51); however, test-retest reliability was lower when the 8- to 19-year-olds were examined separately (r = .49), which may have been due to the small sample size for this age group (n = 28).

Previous Studies

There have been several investigations examining the theoretical and practical utility of the D-KEFS (Crawford, Garthwaite, Sutherland, & Borland, 2011; Floyd, Bergeron, Hamilton, & Parra, 2010; Floyd et al., 2006; Savla et al., 2010). However, there appeared to be an even greater interest in examining the psychometric properties of this assessment, particularly with respect to its validity and reliability across various populations (Crawford, Sutherland, & Garthwaite, 2008; Delis, Kramer, Kapland, Holdnack, 2004; Parmenter et al., 2007; Latzman & Markon, 2010; Mitchell & Miller, 2008; Strong et al., 2010). Despite this interest, however, no known study to date has examined the item characteristics for any of the D-KEFS subtests using item response theory (IRT). Item response models are statistical models specifying that an individual's probability of endorsing an item depends on the individual's ability level and item characteristics such as item location (or threshold) and discrimination (Lord, 1980). Compared to classical item analysis, a major advantage of IRT item analysis is that item location information can be interpreted along with ability levels because they are obtained simultaneously in the same model (e.g., Embretson & Reise, 2000). When the item location and the ability levels are matched, the most precise (IRT scale) scores can be obtained (controlling for the other factors which may affect the precision). In addition, IRT scale scores provide an estimate of a individual's true ability, which is assumed to be free of measurement error. Furthermore, the IRT scale scores

represent estimates of true ability that differentially weight the contribution of the individual items. For example, individuals who got items having high item discriminations correctly can have higher IRT scale scores than those who got items having low item discrimination correctly.

As another limitation of previous studies, although measurement invariance for age groups has been examined in the D-KEFS using multigroup modeling (Latzman & Markon, 2010)[1], no investigation has used differential item functioning (DIF) analysis to determine the extent to which various items are measuring the same (or different) constructs across groups. This absence of research is surprising given that there is evidence to suggest that age-related differences do exist across EF tasks (e.g., De Luca et al., 2003; Huizinga, Dolan, & van der Molen, 2006), including the D-KEFS (Delis et al., 2001b; Latzman & Markon, 2010). For instance, Anderson, Anderson, Northam, Jacobs, and Catroppa (2001) compared the performance of 138 11- to 17-year-olds on several measures of EF (attentional control, cognitive flexibility, and goal setting behaviors). The results indicated that older children tended to score higher than younger children across multiple measures. Yet, despite such evidence, only one study to date has examined age-related measurement invariance for the D-KEFS (Latzman & Markon, 2010). Furthermore, this study was conducted on the original normative sample, thus potentially replicating some of the age effects previously described by Delis and colleagues (2001b). Along this same vein, there has been no examination of gender-related measurement invariance or DIF for the D-KEFS.

The Current Study

Given the general paucity of studies investigating the item characteristics of the D-KEFS, the current investigation sought to address this by examining items on the D-KEFS Card Sort subtest using factor analysis and IRT. The aims of the current study were threefold. First, we extended Latzman and Makon's (2010) and Salva et al.'s (2010) investigations of dimensionality by (a) applying a more fine-grain approach to examining dimensionality based on the Card Sort *items* and (b) examining dimensionality a more diverse typically-developing sample of adolescents using factor analysis. Second, we tested measurement invariance across groups based on age and gender using IRT DIF analyses. Third, we obtained IRT item characteristics, scale scores, and reliability information for D-KEFS.

Method

Participants and Procedure

The original sample included 274 adolescents between the ages of 9 and 15 years. We excluded one participant for whom gender information was missing, nine participants who had missing data across all 16 items, and one participant who had missed the last 7 items. Thus, a total of 4% of the data was excluded due to missingness; the final analysis sample consisted of 264 adolescents (M = 11.66 years, SD = 1.36 years). Within this sample, there were six participants (2.27%) who had missing data for one out of the 16 items, and these

[1]Measurement invariance was investigated using the original standardization sample of the D-KEFS only. Dimensionality of EF was investigated using the original standardization sample and an independent sample of 174 male adolescents.

missing data points were treated as missing at random in all subsequent analyses. Participants were identified as 67.0% Caucasian, 11.4% African American, 4.9% multi-racial, 1.9% Asian, and 14.8% did not specify; 52.27% of the sample was female.

These data were part of a larger reading comprehension study. Testers were trained research assistants who were required to attain at least 90% administration accuracy prior to administering tests to participants. Due to the length of the testing battery, all participants were assessed across 1.5 days, with the two testing days scheduled no more than two weeks apart. Study approval was granted by the university's Institutional Review Board prior to conducting the study. Parents provided written consent and child participants gave assent prior to participating in the study.

Measures

The D-KEFS Card Sort subtest is a 16-item individually-administered assessment of verbal and non-verbal problem solving, cognitive flexibility, and inhibition (Swanson, 2005; Delis et al., 2001b). Points were awarded for correct sorts, and up to 36 sorts were allowed across two sets. Repeated sorts could only be counted once. Raw scores (i.e., item responses) were used in the analyses.

Analysis Outline

For IRT analyses, it is important to select an item response model with appropriate assumptions, dimensionality and local independence. Thus, as the first step, the number of dimensions and the dimensionality structure were investigated using an exploratory factor analysis (EFA). Based on the findings for the number of dimensions and the dimensionality structure, an item response model was chosen for measurement invariance analyses regarding gender and age groups. If there was no concern about measurement invariance, a single-group IRT analysis would serve to provide item characteristics, IRT scale scores, and their precision and would be used to detect group differences on the IRT scale scores. However, if measurement invariance was violated (i.e., DIF was present) then these DIF items should be deleted or modelled for group comparisons (i.e., gender or age groups) on the same scale. In this study, two DIF treatments – deleting DIF items and multigroup IRT (Bock & Zimowski, 1997) – were considered. All IRT analyses were conducted using maximum likelihood estimation in Mplus software (Version 7.11; Muthén & Muthén, 1998-2012).

The frequency of scores across all items is present in Table 1. As shown in Table 1, there were no or few individuals who got Score 1 or Score 2, which can result in unstable threshold estimates of item response models (e.g., large standard errors, convergence problems in estimation). Thus, Score 1 and Score 2 were combined as a partial score to avoid such an estimation problem. Scoring for subsequent IRT analyses was Score 0 (incorrect; Score 0 in Table 1), Score 1 (partially correct; Score 1 and Score 2 in Table 1), or Score 2 (correct; Score 3 in Table 1). For multigroup analysis based on age, the continuous age variable was dichotomized: Group 1 included children who were younger than 12 years of age (56.82% of the sample) and Group 2 included children who were 12 years of age or older (43.18% of the sample). We chose this cut-off for the age groups because there is

evidence to suggest that maturation of executive function skills occur around 12 years of age (e.g., Anderson, 2002). Age for Group 1 ranged from 9 and 11.92 years; age for Group 2 ranged from 12 to 14.83 years.

Results

Number of Dimensions and Dimensionality Structure

A series of EFAs using correlations were conducted using polychoric correlations for categorical outcomes in Mplus, extracting 1-4 factors.[2] Fit indices were compared across models regarding the different number of factors: the root mean square error of approximation index (RMSEA; Steiger & Lind, 1980), the standardized root mean square residual (SRMR), the comparative fit index (CFI; Bentler, 1990), and the Tucker-Lewis index (TLI; Tucker & Lewis, 1973). According to empirical guidelines by Hu and Bentler (1999), a model fits well if RMSEA is smaller than 0.06, SRMR is less than 0.08, and the CFI and TLI are larger than 0.95. The dimensionality structure was examined based on the patterns of factor loadings[3] and their interpretability according to item designs, such that the extracted factors could be meaningfully conceptualized (e.g., in this instance, factors represented verbal and perceptual abilities).

Table 2 presents the four fit indices for one-, two- and three-factor models with GEOMIN, which is an oblique rotation.[4] Results of a four-factor model were not reported in Table 2 because there was a convergence problem. The one-factor model did not provide a good fit to the data based on all four fit indices. Both fit indices and interpretability of factor loading patterns suggested that the two-factor model was the most compelling. Table 3 shows the item characteristics and the GEOMIN rotated factor loadings for the two-factor model. According to the factor loadings, items were mainly clustered with respect to target sort. With the exception of Item 16, all “Perceptual” items (Items 1, 3, 5, 7, 8, 9, 11, 12, and 14) loaded highly on Factor 1 and all “Verbal” items (Items 2, 4, 6, 10, 13, and 15) were loaded highly on Factor 2. Two items (Items 4 and 9) cross-loaded onto both factors; Item 4 required participants to sort cards based on whether the displayed items were single- or multisyllabic words (shorter vs. longer words), and Item 9 required participants to sort cards based on whether or not shapes displayed on the card were tightly grouped. Item 16 was categorized as tapping perceptual skills but loaded onto the Verbal factor. This item required participants to distinguish between shapes that were outlined versus filled in. The two factors were moderately related ($r = 0.523$), which suggests that the two factors are distinct constructs. Taken together, these results suggest that the two dimensions representing “Perceptual” and Verbal” domains could be modelled in the subsequent IRT analyses.

[2]In this paper, we use the term of factor to refer to latent variables in factor analyses and the term dimension to refer to latent variables in IRT analyses.

[3]We used 0.32 as a good rule of thumb for the minimum loading of an item to consider it important/salient, which equates to approximately 10% overlapping variance with the other items in that factor in EFA.

[4]Quartimin, Oblimin, and Varimax rotation methods were also tried. The same conclusions were reached in terms of the number of dimensions and the patterns of factor loadings.

Selected Item Response Model

Based on findings on the dimensionality, we chose a two-dimensional version of the (logistic) graded responses model (Samejima, 1969) for subsequent IRT analyses. In the model, all "Perceptual" items were loaded on Dimension 1 and all "Verbal" items were loaded on Dimension 2. Using the two-dimensional graded response model for three possible scores (incorrect, partially correct, correct), four kinds of item parameters for each item and IRT scale scores were obtained. The four item parameters for each item included: (a) item discriminations for Dimension 1 (denoted by "a1"); (b) item discriminations for Dimension 2 (denoted by "a2"); (c) the first threshold (denoted by "b1"); and (d) the second threshold (denoted by "b2"). Item discrimination parameters describe how well an item can differentiate between individuals and indicate the degree of association between an item and ability. The larger the item discrimination estimates, the better an item differentiates between individuals and the more related that the item is to ability.

The first threshold parameter reflects the transition point from a score of 0 (incorrect) to a score of 1 (partially correct) or 2 (correct; i.e., 0 vs. 1 or 2). The second threshold parameter reflects the transition points from a score of 0 (incorrect) or 1 (partially correct) to a score of 2 (correct; i.e., 0 or 1 vs. 2). When the thresholds cover a wide range of the latent ability scale, the precision of the IRT scale scores increases. The threshold estimates and the IRT scale scores can be interpreted as a *z*-score metric on the logit scale.

Measurement Invariance Analyses

Measurement invariance analyses were implemented for gender and age groups, respectively, using the two-dimensional graded response model. Based on results of dimensionality analysis, we first tested if the two-dimensional graded response model fits well to each subgroup of age and gender groups, respectively. In testing measurement invariance, we first tested whether item discriminations or thresholds differ for all items (i.e., a global test). If the measurement invariance assumption is violated in the global test, we would then test whether item discriminations or thresholds differ for each item (i.e., DIF analysis; Suh & Cho, 2014).

For the global test, three nested invariance models with different constraints were compared to test measurement invariance (e.g., Meredith, 1993): (a) a *configural* invariance model, in which all item parameters were free to be estimated across groups under the same factor structure across groups; (b) a *weak* invariance model, in which only discrimination parameters were constrained to be equal across groups; and (c) a *strong* invariance model, in which all item parameters were constrained to be equal across groups. By comparing the difference in the chi-square values between the configural and weak invariance models, we can test whether measurement invariance assumption is violated for item discriminations. In addition, by comparing the difference in the chi-square values between the weak and strong invariance models, we can test whether measurement invariance assumption is violated for item thresholds. Each test statistic approximately follows a chi-square distribution with degrees of freedom (*df*) equal to the difference in the number of parameters to be estimated between the two models compared.

Wald DIF test was used for DIF analyses.[5] Specifically, taking gender as an example, we tested the null hypotheses that an item discrimination parameter for females was the same as an item discrimination for males and that the two thresholds for females was the same as the two thresholds for males, respectively. For the metric anchoring, we employed the IRT-based likelihood ratio (LRT-LR) DIF procedure (Thissen, Steinberg, & Wainer, 1993). That is, in detecting DIF for each item discrimination or a set of the two thresholds of an item (called a studied item), all items other that the studied items were set as anchor items, which have the same item parameters between the groups.

Age groups—Two age groups (Age Group 1 ranged from 9 to 11.92 years; Age Group 2 ranged from 12 to 14.83 years) were used to compare the three invariance models. As shown in Table 4, the more parsimonious weak invariance model was preferred to the configural invariance model. Thus, there was evidence that item discriminations were the same between the two age groups. Furthermore, the more parsimonious strong invariance model was preferred to the weak invariance model, which means that item thresholds were the same between the two age groups. These results imply that the measurement invariance assumption can be made for the two age groups.

Gender groups—We then compared the three invariance models for males and females. Based on the chi-square difference tests presented in Table 4, the configural model provided a better fit to the data than the weak and strong invariance models. This result indicated that item discriminations and thresholds may differ between males and females.

Because global test results indicated that measurement invariance was violated, DIF analysis were conducted using Wald DIF tests (see Table 5). Item 6 (Verbal) exhibited DIF for threshold parameters and Item 12 (Perceptual) exhibited DIF for both the item discrimination and threshold parameters. Item 6 required participants to sort cards based on whether the displayed words were commonly found in the air versus on land; Item 12 required participants to distinguish between cursive and print words. As shown in Figure 1, thresholds for females were higher than those for males in Item 6, which suggested that Item 6 was harder for females than for males. For Item 12, item discrimination was higher for males than for females and thresholds were higher for females than for males. This means that Item 12 is more discriminating among males and is easier for males than for females (see Figure 1).

IRT Analyses for Gender

Based on finding for measurement invariance tests for comparisons of males and females, two IRT analyses were considered: multigroup IRT analysis including all 16 items and a single-group IRT analysis including 14 items (omitting Items 6 and 12, which demonstrated DIF). We used multigroup IRT analysis for within-group comparisons and a single-group explanatory IRT analysis for between-group comparisons (e.g., group mean comparisons). We discuss this in greater detail below.

[5]Wald DIF test is asymptotically equivalent to LRT-LR DIF procedure.

Multigroup IRT analysis—In a multigroup two-dimensional graded response model, item parameters for males and females were estimated simultaneously and the item parameter estimates for the two groups were connected to a common (latent) metric using the 14 non-DIF items (i.e., anchor items). That is, the same item parameters were estimated for the 14 non-DIF items and the unique item parameters were estimated for the two DIF items (i.e., the first and second thresholds for Item 6 and the item discriminations and first and second thresholds for Item 12).

Item characteristics: Item parameter estimates of the multigroup IRT analysis are reported in Table 6. As indicated by DIF analyses, Item 12 discriminated individuals more for males than for females. Both Items 6 and 12 were easier for males than for females. For the perceptual dimension, item discrimination estimates were between 0.923 and 2.323 for females, whereas they were between 0.923 and 1.462 for females. For the verbal dimension, item discrimination estimates were between 1.045 and 1.665 for males and females. These item discrimination estimates were medium to large values according to empirical guidelines (e.g., Baker, 2001). Varying item discrimination estimates suggested that the strength of item-ability association varied across items. The threshold estimates covered a wide range of the latent ability scale; estimates ranged from −1.562 to 2.932 for the first threshold and from −1.387 to 4.495 for the second threshold for males and females, respectively. This outcome was ideal because it indicated that these items measured a wide range of ability levels.

Distribution of IRT scale scores and reliability: Mean and variance estimates of the ability distributions for males (as the reference group) were set to 0 and 1, respectively, to identify the model. In this model, the covariance estimates were actually correlation estimates due to variance constraints. The covariance between the perceptual and verbal dimensions was 0.712 (*SE*=0.086), which indicated that higher perceptual IRT scale scores had correspondingly high verbal IRT scale scores. Mean estimates of the ability distribution for females were 0.289 (*SE*=0.143) and −0.016 (*SE*=0.160) for the perceptual and verbal dimensions, respectively. The presence of DIF items implied that the meaning of the dimensions was not equivalent between the two groups. Thus, the mean estimates were not meaningful for the comparison purpose. Variance estimates of the ability distribution for females were 0.827 (*SE*=0.263) and 0.755 (*SE*=0.262) for the perceptual and verbal dimensions, respectively, and the covariance estimate of the two dimensions was 0.600 (*SE*=0.178). These results suggest that there was smaller variability in IRT scale scores and a weaker relation of the two dimensions for females compared to males.

One advantage of IRT is that it allows the precision of IRT scale scores to vary as the function of ability levels (i.e., test information curves). These test information curves depict precision as the function of ability levels. It clarifies how well a test measures particular ranges of the ability levels. That is, the higher the test information curve over a given range of ability values, the more reliable the IRT scale scores correspond to the ability level. Figure 2 shows test information curves for each dimension of each group. As seen in Figure 2, the precision differed regarding gender subgroups for perceptual ability: It was relatively high for individuals whose ability levels ranged between −0.5 and 0.5 for males and between

0 and 1 for females. For verbal ability, precision was relatively high for individuals whose ability levels ranged between −0.5 and 0.5 for both gender subgroups. For both perceptual and verbal abilities, precision decreased for individuals whose true ability level was markedly worse or better.

One-Group Explanatory IRT Analysis without the two DIF items—The 14 non-DIF items were also used to fit a single-group model. An explanatory version of two-dimensional graded response model (De Boeck & Wilson, 2004) was applied to obtain item characteristics, the IRT scale scores, and the effects of three covariates (age, gender, and their interaction) on the latent ability scale. The explanatory model was a combination of the two-dimensional graded response model (i.e., a measurement model) and a regression model where the response variables were latent abilities (perceptual and verbal abilities) and the three covariates were added for each dimension. All parameters in the explanatory model were estimated simultaneously. For the measurement model of the explanatory model, the two-factor model fit the data well. Items were clustered by perceptual and verbal domains in EFA results using the 14 non-DIF items. In addition, none of the items were detected as DIF items when DIF analyses were re-conducted using the 14 items between age and gender groups, respectively.

Item characteristics: Table 7 presents the item parameter estimates of the explanatory model. Patterns of item characteristics for the 14 items in the single-group analysis were similar to those in the multigroup analysis. The item discriminations were relatively high, and thresholds covered a wide range of ability levels.

Distribution of IRT scale scores and reliability: Means and variances of perceptual and verbal abilities were set to 0 and 1, respectively, to identify the model. The correlation of the two abilities was 0.776, which indicated a strong linear relation between the two dimensions. According to the test information curves for each dimension, the precision of scores was high for individuals whose ability levels were between 0.5 and 1.5 for perceptual ability and between 1.0 and 2.0 for verbal ability (see Figure 3). The marginal IRT reliability (which can range from 0 to 1; Green, Bock, Humphreys, Linn, & Reckase, 1984) was 0.79 and 0.74 for perceptual ability and verbal ability, respectively, which suggests that the reliability was satisfactory.

Mean differences across gender and age groups on the perceptual and verbal dimensions: Using the explanatory model, mean differences on the two dimensions were tested. Prior to including gender and age within the explanatory models, we effect coded gender so that the male grouping was equal to −0.5 and the female grouping was equal to 0.5, and we centered age at the mean ($M = 11.66$).[6] For perceptual ability, the main effects of gender and age were significant (Estimate = 0.343, $t = 2.124$, $p = 0.034$ for Gender; Estimate = 0.132, $t = 2.295$, p=0.022 for Age); however, there was no interaction effect of gender and age (Estimate = −0.141, $t = -1.266$, $p = 0.205$). These results for perceptual ability indicate that the mean level for females was 0.343 higher than males (on the

[6]Intercepts in the regression model of the explanatory model were fixed to 0 to identify the model.

standardized ability scale) and there was 0.132 score increase with one-year change in age. For verbal ability, the main effect of gender was positive but was not significant (Estimate = 0.035, $t = 0.200$, $p = 0.841$). This result indicated that although the level for females was higher than for males, the difference on average was not significant. The main effect of age was significant (Estimate = 0.193, $t = 2.583$, $p = 0.010$), which suggests that there is 0.193 score increase with one-year change in age on average. Additionally, the interaction of gender and age was significant (Estimate = −0.336, $t = -2.221$, $p = 0.026$), which indicates that the slope of age (i.e., change in score with change in age) is different for males and females.

Discussion

Within the current investigations, we examined the psychometric properties of the D-KEFS due to its relatively wide use in both clinical and educational practice. The present study adds to what is known about the D-KEFS by including a sample that is separate from the original standardization sample (Delis et al., 2001b) and one that is more diverse than those included within previous investigations of measurement invariance (e.g., Latzman & Markon, 2010). Results of the EFA indicated that a two-factor model provided the best fit to the data, with verbal and perceptual items loading highly onto their respective verbal and perceptual factors.

Comparisons of invariance models suggested that measurement invariance held based on age, with a strong invariance model being supported by the data. Although our findings differed from those of Latzman and Markon (2010) who showed that the D-KEFS demonstrated partial factorial invariance across age, there are several notable differences between the current study and this previous investigation. First, Latzman and Markon examined dimensionality for all subtests of the D-KEFS whereas we examined dimensionality for the Card Sort items only. Second, the age groups compared in the current study (comparing two groups of children aged 9 to 11.92 years and 12 to 14.83 years) were more narrow in range that those in Latzman and Markon's study (comparing three groups of individuals aged 8 to 19 years, 20 to 49 years, and 50 to 89 years).

When we compared measurement invariance models based on gender we found that measurement invariance was violated; the weak invariance model provided the best fit to the data. This suggested that some items varied in their discrimination and threshold parameters. Further exploration indicated that two items, Item 6 and Item 12, demonstrated DIF. Both items were easier for males (i.e., variability in threshold parameters), and one item (Item 12) had higher differentiation for males than females (i.e., variability in the discrimination parameter). This finding is important because DIF is problematic given its potential for bias towards particular subgroups (Lord, 1980), and it is often advised that items demonstrating DIF be modified or removed from an assessment in given their potential to impact the accuracy of measurement (e.g., Teresi, Ramirez, Jones, Choi, & Crane, 2012).

The current investigation additionally adds to what is known about the psychometric properties of the Card Sort subtest of the D-KEFS for adolescents across varying ability levels. Overall, both verbal and perceptual items measured a wide range of ability levels;

however, the precision (i.e., reliability) of these scores was best for individuals whose IRT scale scores were relatively high (e.g., between 1 and 2 on the *z*-score scale). Consequently, any ability level falling outside of this range tended to be estimated with much less accuracy. This outcome has implications for the use of the Card Sort subtest of the D-KEFS in the assessment of individuals with impaired EF (e.g., Heled et al., 2012), as the current findings suggest the possibility of imprecise measurement for these individuals. Turning to scale score performance, results showed that older children tended to attain higher scores compared to younger children for both verbal and perceptual sorts. This outcome was not entirely unexpected given the age effects described in the D-KEFS technical manual (Delis et al., 2001b) and previous studies showing that performance on EF tasks tends to increase with age (De Luca et al., 2003; Huizinga et al., 2006).

Gender-based comparisons of scores indicated that females tended to attain higher scores on the perceptual card sorts than males. Such findings may have implications for investigations of EF that include gender-imbalanced treatment and control groups (e.g., Huizinga et al., 2006; Parrish et al., 2007), as it may affect observed scores. The finding that scores increased with age across verbal and perceptual sorts for both males and females was expected given that EF skills tend to develop well into adulthood (e.g., Blakemore & Choudhury, 2006; Zelazo & Carlson, 2012). For verbal ability, there was an interaction between age and gender that indicated that females attained higher scores with age relative to males. This outcome may be explained by the fact that the development of EF is highly stable and that individuals who exhibit lower performance on measures of EF (males, in this instance) may subsequently exhibit lower gains in EF across age (e.g., Biederman et al., 2007). However, this potential explanation must be approached cautiously given that these data were not longitudinal and also the fact that the mean difference in performance between males and females was not statistically significant.

In summary, the current investigation adds greatly to what is known about the psychometric properties of the Card Sort subtest of the D-KEFS. However, it is important to acknowledge that the Card Sort subtest is but one of nine subtests that comprise the D-KEFS. Therefore, future studies should use IRT to examine the other subtests of the D-KEFS in order to construct a more global psychometric evaluation of this assessment. Future investigations should also include a wider age range as a way to obtain item characteristics for younger children as well as adults. The application of such methods can lead to the subsequent improvement of the D-KEFS assessment so that individuals can be accurately assessed across the entire continuum of EF ability.

Acknowledgments

This work was supported by the National Institute of Child Health and Human Development under grant numbers R01 HD 044073, U54 HD 083211 and R01 HD 044073-14S1 and the National Center for Advancing Translational Sciences under grant number UL1 TR000445.

References

Altemeier L, Jones J, Abbott RD, Berninger VW. Executive functions in becoming writing readers and reading writers: Note taking and report writing in third and fifth graders. Developmental Neuropsychology. 2006; 29(1):161–173. DOI: 10.1207/s15326942dn2901_8 [PubMed: 16390292]

Anderson P. Assessment and development of executive function (EF) during childhood. Child Neuropsychology. 2002; 8(2):71–82. DOI: 10.1076/chin.8.2.71.8724 [PubMed: 12638061]

Anderson VA, Anderson P, Northam E, Jacobs R, Catroppa C. Development of executive functions through late childhood and adolescence in an Australian sample. Developmental Neuropsychology. 2001; 20(1):385–406. DOI: 10.1207/S15326942DN2001_5 [PubMed: 11827095]

Angoff WH. Scales, norms, and equivalent scores. Educational testing service; Princeton, NJ: 1984.

Baker F. The Basics of item response theory. Second. College Park: MD: ERIC Clearinghouse on Assessment and Evaluation; 2001.

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107(2):238–246. DOI: 10.1037/0033-2909.107.2.238 [PubMed: 2320703]

Biederman J, Monuteaux MC, Doyle AE, Seidman LJ, Wilens TE, Ferrero F, Faraone SV. Impact of executive function deficits and attention-deficit/hyperactivity disorder (ADHD) on academic outcomes in children. Journal of Consulting and Clinical Psychology. 2004; 72(5):757–766. DOI: 10.1037/0022-006X.72.5.757 [PubMed: 15482034]

Biederman J, Petty CR, Fried R, Doyle AE, Spencer T, Seidman LJ, Faraone SV. Stability of executive function deficits into young adult years: a prospective longitudinal follow-up study of grown up males with ADHD. Acta Psychiatrica Scandinavica. 2007; 116(2):129–136. DOI: 10.1111/j.1600-0447.2007.01008.x [PubMed: 17650275]

Blair C, Razza RP. Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. Child Development. 2007; 78(2):647–663. DOI: 10.1111/j.1467-8624.2007.01019.x [PubMed: 17381795]

Blakemore SJ, Choudhury S. Development of the adolescent brain: implications for executive function and social cognition. Journal of Child Psychology and Psychiatry. 2006; 47(3–4):296–312. DOI: 10.1111/j.1469-7610.2006.01611.x [PubMed: 16492261]

Bock RD, Zimowski MF. Multiple group IRT. In: van der Linden WJ, Hambleton RK, editorsHandbook of modern item response theory. New York, NY: Springer; 1997. 433–448.

Brock LL, Rimm-Kaufman SE, Nathanson L, Grimm KJ. The contributions of 'hot'and 'cool'executive function to children's academic achievement, learning-related behaviors, and engagement in kindergarten. Early Childhood Research Quarterly. 2009; 24(3):337–349. DOI: 10.1016/j.ecresq.2009.06.001

Bull R, Espy KA, Wiebe SA. Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. Developmental Neuropsychology. 2008; 33(3):205–228. DOI: 10.1080/87565640801982312 [PubMed: 18473197]

Clark CA, Pritchard VE, Woodward LJ. Preschool executive functioning abilities predict early mathematics achievement. Developmental Psychology. 2010; 46(5):1176–1191. DOI: 10.1037/a0019672 [PubMed: 20822231]

Crawford JR, Garthwaite PH, Sutherland D, Borland N. Some supplementary methods for the analysis of the Delis–Kaplan Executive Function System. Psychological Assessment. 2011; 23(4):888–898. DOI: 10.1037/a0023712 [PubMed: 21574720]

Crawford JR, Sutherland D, Garthwaite PH. On the reliability and standard errors of measurement of contrast measures from the D-KEFS. Journal of the International Neuropsychological Society. 2008; 14(06):1069–1073. DOI: 10.1017/0S1355617708081228 [PubMed: 18954487]

De Boeck P, Wilson M. Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer; 2004.

De Luca CR, Wood SJ, Anderson V, Buchanan JA, Proffitt TM, Mahony K, Pantelis C. Normative data from the CANTAB. I: development of executive function over the lifespan. Journal of Clinical and Experimental Neuropsychology. 2003; 25(2):242–254. DOI: 10.1076/jcen.25.2.242.13639 [PubMed: 12754681]

Delis DC, Kaplan E, Kramer JH. The Delis-Kaplan executive function system: Examiner's manual. San Antonio: The Psychological Corporation; 2001a.

Delis DC, Kaplan E, Kramer JH. The Delis-Kaplan Executive Function System: Tech. San Antonio. TX: Pearson, Inc; 2001b.

Delis DC, Kramer JH, Kaplan E, Holdnack J. Reliability and validity of the Delis-Kaplan Executive Function System: an update. Journal of the International Neuropsychological Society. 2004; 10(02):301–303. DOI: 10.1017/0S1355617704102191 [PubMed: 15012851]

Eisenberg N, Fabes RA, Guthrie IK, Reiser M. Dispositional emotionality and regulation: their role in predicting quality of social functioning. Journal of Personality and Social Psychology. 2000; 78(1): 136.doi: 10.1O37//0022-3514.78.1.136 [PubMed: 10653511]

Embretson SE, Reise SP. Item response theory for psychologist. Mahwah, NJ: Erlbaum; 2000.

Floyd RG, Bergeron R, Hamilton G, Parra GR. How do executive functions fit with the Cattell–Horn–Carroll model? Some evidence from a joint factor analysis of the Delis–Kaplan executive function system and the Woodcock–Johnson III tests of cognitive abilities. Psychology in the Schools. 2010; 47(7):721–738. DOI: 10.1002/pits.20500

Floyd RG, McCormack AC, Ingram EL, Davis AE, Bergeron R, Hamilton G. Relations between the Woodcock-Johnson III clinical clusters and measures of executive functions from the Delis-Kaplan Executive Function System. Journal of Psychoeducational Assessment. 2006; 24(4):303–317. DOI: 10.1177/0734282906287823

Gioia GA, Isquith PK, Retzlaff PD, Espy KA. Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF) in a clinical sample. Child Neuropsychology. 2002; 8(4): 249–257. DOI: 10.1076/chin.8.4.249.13513 [PubMed: 12759822]

Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement. 1984; 21(4):347–360.

Heled E, Hoofien D, Margalit D, Natovich R, Agranov E. The Delis–Kaplan executive function system sorting test as an evaluative tool for executive functions after severe traumatic brain injury: A comparative study. Journal of Clinical and Experimental Neuropsychology. 2012; 34(2):151–159. DOI: 10.1080/13803395.2011.625351 [PubMed: 22114911]

Homack S, Lee D, Riccio CA. Test review: Delis-Kaplan executive function system. Journal of Clinical and Experimental Neuropsychology. 2005; 27(5):599–609. DOI: 10.1080/13803390490918444 [PubMed: 16019636]

Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 1999; 6(1):1–55. DOI: 10.1080/10705519909540118

Huizinga M, Dolan CV, van der Molen MW. Age-related change in executive function: Developmental trends and a latent variable analysis. Neuropsychologia. 2006; 44(11):2017–2036. DOI: 10.1016/j.neuropsychologia.2006.01.010 [PubMed: 16527316]

Keifer E, Tranel D. A neuropsychological investigation of the Delis-Kaplan executive function system. Journal of Clinical and Experimental Neuropsychology. 2013; 35(10):1048–1059. DOI: 10.1080/13803395.2013.854319 [PubMed: 24236952]

Latzman RD, Markon KE. The factor structure and age-related factorial invariance of the Delis-Kaplan Executive Function System (D-KEFS). Assessment. 2010; 17(2):172–184. DOI: 10.1177/1073191109356254 [PubMed: 20040723]

Locascio G, Mahone EM, Eason SH, Cutting LE. Executive dysfunction among children with reading comprehension deficits. Journal of learning disabilities. 2010; 43(5):441–454. DOI: 10.1177/0022219409355476 [PubMed: 20375294]

Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence; 1980.

Mayes SD, Calhoun SL, Crowell EW. Learning disabilities and ADHD overlapping spectrum disorders. Journal of Learning Disabilities. 2000; 33(5):417–424. DOI: 10.1177/002221940003300502 [PubMed: 15495544]

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993; 58(4):525–543. DOI: 10.1007/BF02294825

Miller M, Hinshaw SP. Does childhood executive function predict adolescent functional outcomes in girls with ADHD? Journal of Abnormal Child Psychology. 2010; 38(3):315–326. DOI: 10.1007/s10802-009-9369-2 [PubMed: 19960365]

Mitchell M, Miller LS. Prediction of functional status in older adults: The ecological validity of four Delis–Kaplan Executive Function System tests. Journal of Clinical and Experimental

Neuropsychology. 2008; 30(6):683–690. DOI: 10.1080/13803390701679893 [PubMed: 18608647]

Miyake A, Friedman NP. The nature and organization of individual differences in executive functions: Four general conclusions. Current Directions in Psychological Science. 2012; 21(1):8–14. DOI: 10.1177/0963721411429458 [PubMed: 22773897]

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive Psychology. 2000; 41(1):49–100. DOI: 10.1006/cogp.1999.0734 [PubMed: 10945922]

Muthén LK, Muthén BO. Mplus User's Guide. 7th. Los Angeles, CA: Muthén & Muthén; 1998–2012.

Parmenter BA, Zivadinov R, Kerenyi L, Gavett R, Weinstock-Guttman B, Dwyer MG, Benedict RH. Validity of the Wisconsin card sorting and Delis–Kaplan executive function system (DKEFS) sorting tests in multiple sclerosis. Journal of Clinical and Experimental Neuropsychology. 2007; 29(2):215–223. DOI: 10.1080/13803390600672163 [PubMed: 17365256]

Parrish J, Geary E, Jones J, Seth R, Hermann B, Seidenberg M. Executive functioning in childhood epilepsy: parent-report and cognitive assessment. Developmental Medicine & Child Neurology. 2007; 49(6):412–416. DOI: 10.1111/j.1469-8749.2007.00412.x [PubMed: 17518924]

Razza RA, Blair C. Associations among false-belief understanding, executive function, and social competence: A longitudinal analysis. Journal of Applied Developmental Psychology. 2009; 30(3): 332–343. DOI: 10.1016/j.appdev.2008.12.020 [PubMed: 20161159]

Romine CB, Lee D, Wolfe ME, Homack S, George C, Riccio CA. Wisconsin Card Sorting Test with children: a meta-analytic study of sensitivity and specificity. Archives of Clinical Neuropsychology. 2004; 19(8):1027–1041. DOI: 10.1016/j.acn.2003.12.009 [PubMed: 15533695]

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph. 1969; 34(Suppl. 1):1–97. DOI: 10.1007/BF03372160

Savla GN, Twamley EW, Delis DC, Roesch SC, Jeste DV, Palmer BW. Dimensions of executive functioning in schizophrenia and their relationship with processing speed. Schizophrenia Bulletin. 2010; :760–768. DOI: 10.1093/schbul/sbq149 [PubMed: 21163899]

Savla GN, Twamley EW, Thompson WK, Delis DC, Jeste DV, Palmer BW. Evaluation of specific executive functioning skills and the processes underlying executive control in schizophrenia. Journal of the International Neuropsychological Society. 2011; 17(01):14–23. DOI: 10.1017/S1355617710001177 [PubMed: 21062522]

Sesma HW, Mahone EM, Levine T, Eason SH, Cutting LE. The contribution of executive skills to reading comprehension. Child Neuropsychology. 2009; 15(3):232–246. DOI: 10.1080/09297040802220029 [PubMed: 18629674]

St Clair-Thompson HL, Gathercole SE. Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. The Quarterly Journal of Experimental Psychology. 2006; 59(4):745–759. DOI: 10.1080/17470210500162854 [PubMed: 16707360]

Steiger JH, Lind JC. Statistically-based tests for the number of common factors; Paper presented at the annual Spring meeting of the Psychometric Society; Iowa City, IA. 1980 May.

Strong CAH, Tiesma D, Donders J. Criterion validity of the Delis-Kaplan Executive Function System (D-KEFS) fluency subtests after traumatic brain injury. Journal of the International Neuropsychological Society. 2010; 17(2):230–237. DOI: 10.1017/S1355617710001451 [PubMed: 21122190]

Sue Baron I. Delis-Kaplan executive function system. Child Neuropsychology. 2004; 10(2):147–152. DOI: 10.1080/09297040490911140

Suh Y, Cho S-J. Chi-square difference tests for detecting differential functioning in a multidimensional IRT model: A Monte Carlo study. Applied Psychological Measurement. 2014; 38(5):359–375. DOI: 10.1177/0146621614523116

Swanson J. The Delis-Kaplan Executive Function System: A Review. Canadian Journal of School Psychology. 2005; 20(1–2):117–128. DOI: 10.1177/0829573506295469

Teresi JA, Ramirez M, Jones RN, Choi S, Crane PK. Modifying measures based on differential item functioning (DIF) impact analyses. Journal of Aging and Health. 2012; 24(6):1044–1076. DOI: 10.1177/0898264312436877 [PubMed: 22422759]

Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holland PW, Wainer H, editorsDifferential item functioning. Hillsdale, NJ: Erlbaum; 1993.

Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika. 1973; 38(1):1–10. DOI: 10.1007/BF02291170

Willcutt EG, Doyle AE, Nigg JT, Faraone SV, Pennington BF. Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. Biological Psychiatry. 2005; 57(11):1336–1346. DOI: 10.1016/j.biopsych.2005.02.006 [PubMed: 15950006]

Willcutt EG, Petrill SA, Wu S, Boada R, DeFries JC, Olson RK, Pennington BF. Comorbidity between reading disability and math disability: Concurrent psychopathology, functional impairment, and neuropsychological functioning. Journal of Learning Disabilities. 2013; 46(6):500–516. DOI: 10.1177/0022219413477476 [PubMed: 23449727]

Zelazo PD, Carlson SM. Hot and cool executive function in childhood and adolescence: Development and plasticity. Child Development Perspectives. 2012; 6(4):354–360. DOI: 10.1111/j.1750-8606.2012.00246.x

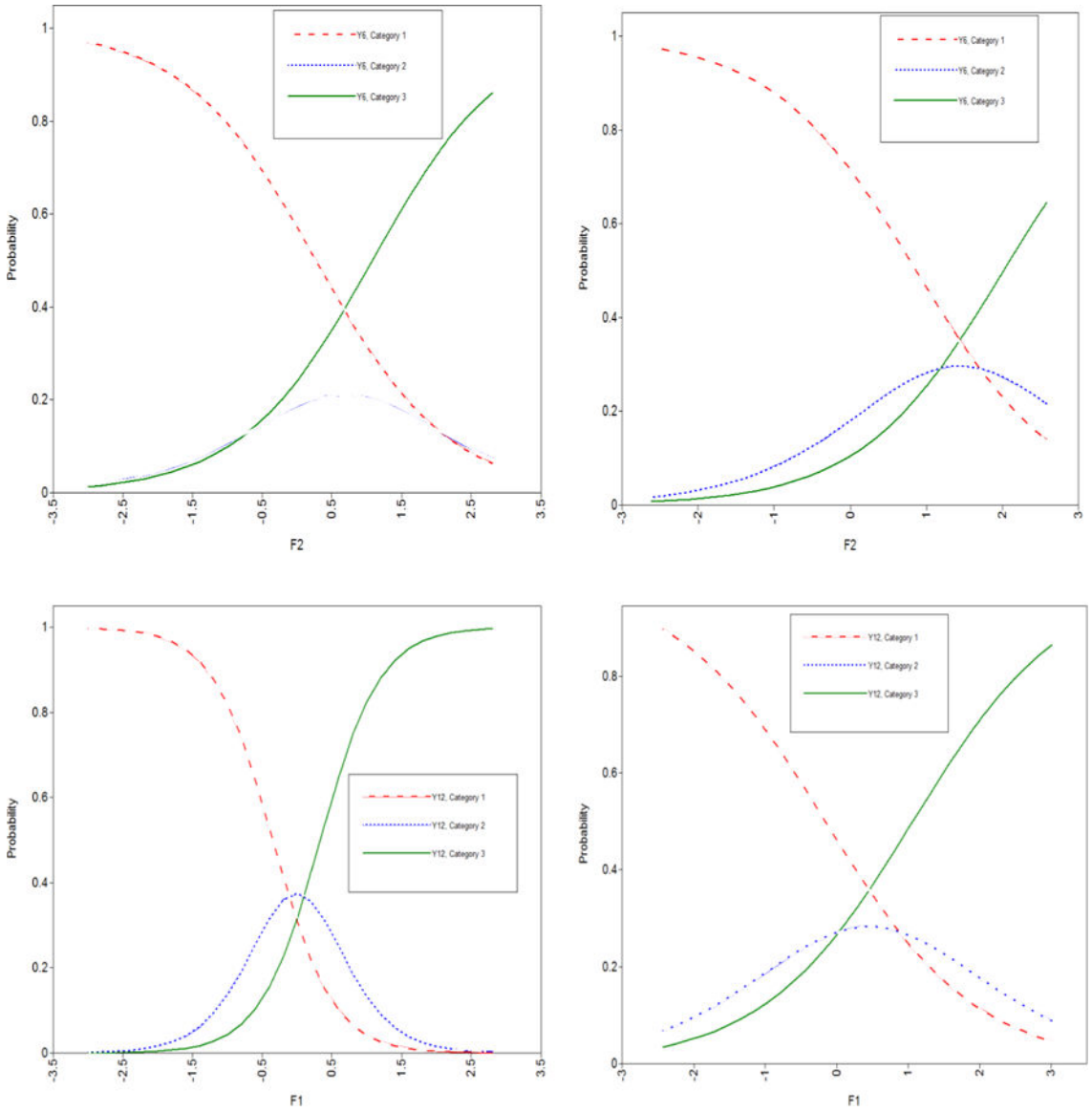

Figure 1.

Item characteristic curves for males on Item 6 (top; left) and for females on Item 6 (top; right); Item characteristic curves for males on Item 12 (bottom; left) and for females on Item 12 (bottom, right).

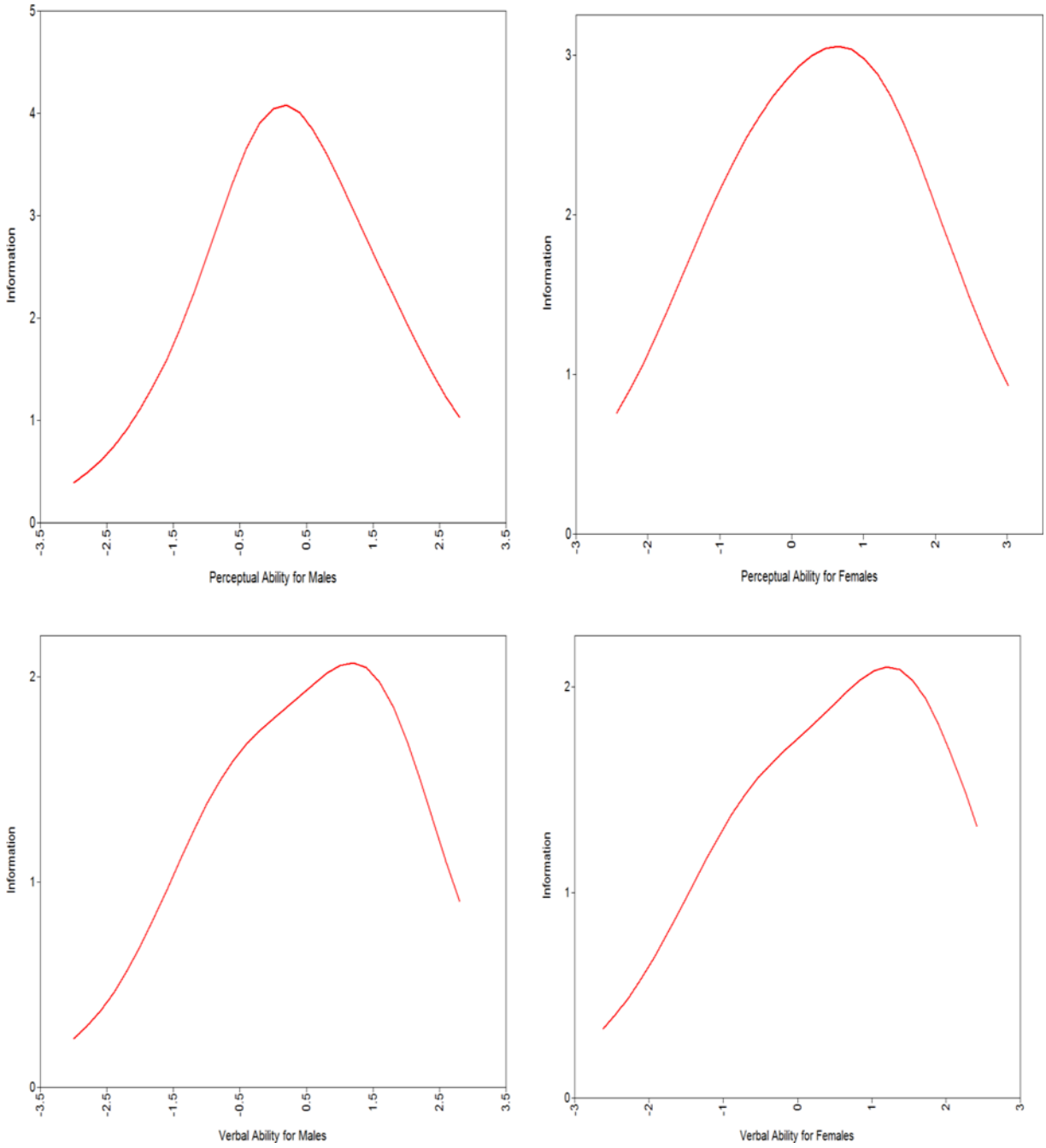

Figure 2.

Test information curves from the multigroup IRT analysis, holding the other dimensions constant at their means.

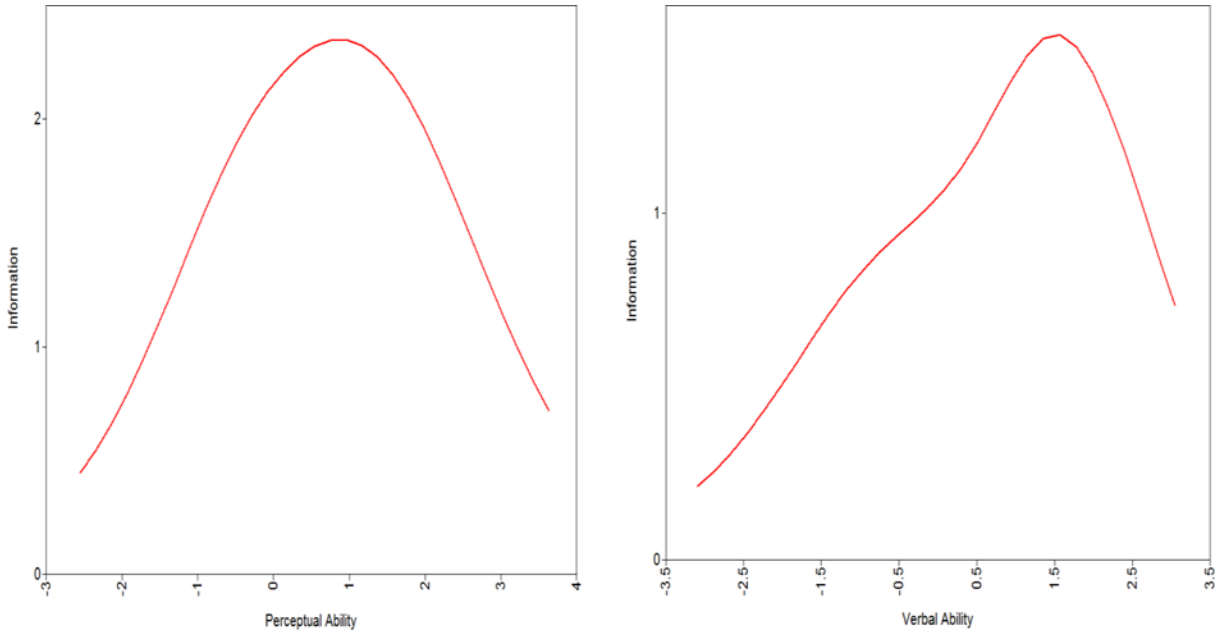

Figure 3.

Test information curves from the single-group IRT analysis, holding the other dimensions constant at their means.

Table 1

Frequency of Each Score Across Items for the Card Sort Subtest of the Delis-Kaplan Executive Function System (N=264)

	Scores				
Items	**0**	**1**	**2**	**3**	**Missing**
1	150	0	1	113	0
2	61	1	20	182	0
3	141	3	42	78	0
4	237	18	1	7	1
5	55	4	0	205	0
6	166	3	42	53	0
7	128	21	3	112	0
8	193	0	11	60	0
9	58	5	10	190	1
10	88	0	11	165	0
11	203	5	1	54	1
12	104	3	58	99	0
13	215	0	32	16	1
14	215	34	1	14	0
15	142	5	8	109	0
16	73	16	36	137	2

Table 2

Fit Indices for the One-, Two-, and Three-Factor Models

Model	RMSEA (90% CI)	SRMR	CFI	TLI
One-factor	0.040 (0.024, 0.054)	0.091	0.938	0.928
Two-factor	0.027 (0.000, 0.045)	0.074	0.976	0.968
Three-factor	0.023 (0.000, 0.044)	0.065	0.985	0.977

Note. RMSEA = root-mean-square-error of approximation; CI = Confidence interval; SRMR = standardized root mean square residual; CFI = Comparative fit index; TLI = Tucker-Lewis index.

Table 3

Item Characteristics and GEOMIN Rotated Loadings for the Two-Factor Model

Item	Card Set	Target Sort	Factor 1	Factor 2
1	1	Perceptual	**0.632**	−0.167
2	1	Verbal	0.035	**0.603**
3	1	Perceptual	**0.607**	0.098
4	1	Verbal	**0.324**	0.284
5	1	Perceptual	**0.692**	−0.075
6	1	Verbal	−0.149	**0.669**
7	1	Perceptual	**0.482**	0.050
8	1	Perceptual	**0.564**	0.008
9	2	Perceptual	**0.415**	**0.319**
10	2	Verbal	0.095	**0.489**
11	2	Perceptual	**0.560**	−0.010
12	2	Perceptual	**0.478**	0.200
13	2	Verbal	0.216	**0.443**
14	2	Perceptual	**0.476**	0.078
15	2	Verbal	−0.008	**0.619**
16	2	Perceptual	0.225	**0.445**

Note. Factor loading highly loaded on a factor (> 0.32 as a rule of thumb) is in bold. Factor 1 = Perceptual; Factor 2 = Verbal.

Table 4

Results of Measurement Invariance Tests for Age and Gender Groups

	Model Fit Statistics			Tests of Measurement Invariance			
Groups	**Models**	**Parameters**	**Log-likelihood**	**Comparisons**	**Chi-square**	***df***	***p*-value**
Age							
	Configural	98	−3174.142	Weak vs. Configural	13.006	14	0.526
	Weak	84	−3180.645	Strong vs. Configural	41.916	44	0.561
	Strong	54	−3195.099	Strong vs. Weak	28.908	30	0.522
Gender							
	Configural	98	−3164.253	Weak vs. Configural	30.486	14	0.007
	Weak	84	−3179.496	Strong vs. Configural	68.256	44	0.011
	Strong	54	−3198.381	Strong vs. Weak	37.77	30	0.156

Note. df = Degrees of freedom.

Table 5

Magnitudes for Differential Item Functioning Items Based on Gender

	Parameter Estimates		
Item	**a1 or a2**	**b1**	**b2**
1 *p*	0.202 (0.456)	−0.152 (0.341)	−0.191 (0.395)
2 *v*	2.486 (1.947)	−0.518 (0.364)	−0.626 (0.352)
3 *p*	0.276 (0.550)	−0.040 (0.327)	−0.108 (0.349)
4 *v*	0.358 (0.905)	−0.146 (0.562)	−1.445 (0.868)
5 *p*	−0.620 (0.428)	−0.282 (0.364)	−0.163 (0.350)
6 *v*	−1.055 (0.576)	**0.629 (0.305)**	**0.997 (0.357)**
7 *p*	−0.834 (0.435)	−0.024 (0.290)	−0.205 (0.285)
8 *p*	0.891 (0.530)	−0.927 (0.346)	−0.909 (0.378)
9 *p*	0.601 (0.551)	0.341 (0.375)	0.348 (0.365)
10 *v*	−0.036 (0.530)	0.013 (0.320)	−0.071 (0.322)
11 *p*	−0.019 (0.455)	−0.215 (0.351)	−0.388 (0.343)
12 *p*	**−1.270 (0.463)**	**0.694 (0.334)**	**0.669 (0.326)**
13 *v*	0.500 (0.910)	−0.096 (0.479)	0.359 (0.622)
14 *p*	0.268 (0.454)	0.123 (0.372)	−0.403 (0.627)
15 *v*	−0.733 (0.575)	−0.168 (0.320)	−0.020 (0.332)
16 *p*	0.561 (0.478)	0.104 (0.314)	0.336 (0.306)

Note. Standard errors reported in parentheses, and estimates demonstrating differential item functioning are in bold.

p Perceptual;

v Verbal; a1 = item discriminations for Dimension 1; a2 = item discriminations for Dimension 2; b1 = the first threshold; b2 = the second threshold.

Table 6

Item Parameter Estimates for Multigroup IRT Models Comparing Males and Females

	Estimates for Males			Estimates for Females		
Items	**a1 or a2**	**b1**	**b2**	**a1 or a2**	**b1**	**b2**
1 *P*	0.923 (0.236)	0.471 (0.171)	0.489 (0.172)	0.923 (0.236)	0.471 (0.171)	0.489 (0.172)
2 *V*	1.377 (0.320)	−1.562 (0.244)	−1.034 (0.224)	1.377 (0.320)	−1.562 (0.244)	−1.034 (0.224)
3 *P*	1.569 (0.329)	0.461 (0.229)	1.470 (0.252)	1.569 (0.329)	0.461 (0.229)	1.470 (0.252)
4 *V*	1.623 (0.529)	2.932 (0.473)	4.495 (0.580)	1.623 (0.529)	2.932 (0.473)	4.495 (0.580)
5 *P*	1.270 (0.290)	−1.500 (0.235)	−1.387 (0.229)	1.270 (0.290)	−1.500 (0.235)	−1.387 (0.229)
6 *V*	1.045 (0.265)	**0.294 (0.221)**	**1.146 (0.237)**	1.045 (0.265)	**0.920 (0.252)**	**2.141 (0.318)**
7 *P*	1.008 (0.238)	0.091 (0.177)	0.527 (0.185)	1.008 (0.238)	0.091 (0.177)	0.527 (0.185)
8 *P*	1.301 (0.264)	1.508 (0.254)	1.781 (0.270)	1.301 (0.264)	1.508 (0.254)	1.781 (0.270)
9 *P*	1.462 (0.294)	−1.466 (0.251)	−1.051 (0.231)	1.462 (0.294)	−1.466 (0.251)	−1.051 (0.231)
10 *V*	1.247 (0.285)	−0.871 (0.203)	−0.640 (0.200)	1.247 (0.285)	−0.871 (0.203)	−0.640 (0.200)
11 *P*	1.293 (0.295)	1.747 (0.272)	1.910 (0.269)	1.293 (0.295)	1.747 (0.272)	1.910 (0.269)
12 *P*	**2.323 (0.556)**	**−0.815 (0.344)**	**0.764 (0.342)**	**0.952 (0.285)**	**−0.152 (0.223)**	**1.013 (0.253)**
13 *V*	1.665 (0.536)	2.058 (0.371)	3.566 (0.489)	1.665 (0.536)	2.058 (0.371)	3.566 (0.489)
14 *P*	1.355 (0.324)	2.127 (0.311)	3.731 (0.446)	1.355 (0.324)	2.127 (0.311)	3.731 (0.446)
15 *V*	1.245 (0.325)	0.208 (0.192)	0.461 (0.201)	1.245 (0.325)	0.208 (0.192)	0.461 (0.201)
16 *P*	0.999 (0.198)	−0.939 (0.174)	0.090 (0.169)	0.999 (0.198)	−0.939 (0.174)	0.090 (0.169)

Note. Standard errors in parentheses; Items 6 and 12 were excluded for estimation; estimates are on the logit scale; and estimates differing between males and females are in bold.

P Perceptual;

V Verbal; a1 = item discriminations for Dimension 1; a2 = item discriminations for Dimension 2; b1 = the first threshold; b2 = the second threshold.

Table 7

Item Parameter Estimates for the Single-Group IRT Model Excluding Items Demonstrating Differential Item Functioning

		Parameter Estimates	
Items	**a1 or a2**	**b1**	**b2**
1 [p]	0.905(0.216)	0.825(0.299)	0.844(0.300)
2 [v]	1.140(0.304)	−1.419(0.351)	−0.908(0.345)
3 [p]	1.446(0.278)	1.004(0.415)	2.011(0.438)
4 [v]	1.416(0.466)	3.056(0.619)	4.626(0.678)
5 [p]	1.190(0.239)	−1.053(0.357)	−0.940(0.354)
6 [v]	–	–	–
7 [p]	0.983(0.215)	0.471(0.316)	0.913(0.332)
8 [p]	1.081(0.230)	1.841(0.401)	2.107(0.414)
9 [p]	1.370(0.294)	−0.949(0.374)	−0.536(0.369)
10 [v]	1.007(0.253)	−0.758(0.304)	−0.534(0.307)
11 [p]	1.081(0.263)	2.080(0.420)	2.238(0.418)
12 [p]	–	–	–
13 [v]	1.682(0.458)	2.364(0.608)	3.978(0.720)
14 [p]	1.233(0.287)	2.574(0.476)	4.178(0.586)
15 [v]	0.964(0.236)	0.281(0.293)	0.524(0.301)
16 [p]	0.905(0.216)	−0.579(0.281)	0.467(0.284)

Note. Standard errors in parentheses; Items 6 and 12 were excluded for estimation; and estimates are on the logit scale.

[p]Perceptual;

[v]Verbal; a1 = item discriminations for Dimension 1; a2 = item discriminations for Dimension 2; b1 = the first threshold; b2 = the second threshold.