Psychological Assessment
2015, Vol. 27, No. 3, 755–762

1040-3590/15/$12.00 http://dx.doi.org/10.1037/pas0000106

The Utility of Set-Loss Error Scores in the General Population

Ivan Ropovik and Monika Bobakova
University of Presov

Jan Ferjencik
University of Pavol Jozef Safarik

Marta Filickova and Iveta Kovalcikova
University of Presov

Miriam Slavkovska
University of Pavol Jozef Safarik

Although the measurement of cognitive performance usually relies on achievement sum scores, a growing body of research suggests that the analysis of errors made may have a predictive validity beyond that provided by the number of items correct. This study examined the validity related to one such kind of error scores—the set-loss errors—in the general population of 8- to 11-year-old children. Set-loss errors (also called rule violations) can be conceptualized as a breakdown in the adherence to task-specific rules, and in clinical populations, the propensity to make these errors has shown some specificity for identifying disorders connected with frontal lobes dysfunction. The results, however, indicate that set-loss errors derived from distinct tests could not be effectively explained by a single latent dimension; hence, they do not tap a single construct that could be called set loss or the ability to maintain set. At the same time, there were only few weak associations between various kinds of error scores as well as between the set-loss error scores and relevant constructs such as the ability to learn, attentional control, working memory, fluid and crystallized intelligence, and executive functions–related real-world behaviors, indicating an overrepresentation of construct-irrelevant variance in these kinds of scores. These indications were further accentuated by the analysis of sensitivity and specificity where any elevated number of set-loss error scores was unable to classify individuals on theoretically relevant constructs beyond chance levels. The evidence thus speaks against the use of set-loss error scores in the general population of 8- to 11-year-old children.

Keywords: set loss, rule violation, error score, ability to maintain set, attentional control

Despite the enormous progress in psychometrics that we have been witnessing in recent decades, total achievement scores keep proving their somehow surprising utility. Summing the item scores as in a number-correct score is the most primitive way to "measure" psychological attributes, yet it usually beats the more sophisticated scaling models due to its simplicity and robustness (Nunnally & Bernstein, 1994). That is why the achievement sum scores remain the most frequently used type of measure. Here, the majority of scoring procedures that are employed in cognitive assessment provide a measure of the target construct by summing only the items correct while ignoring the route to solution and the nature of errors. However, by assuming that all the errors reflect just the opposite aspect of the same construct as do correct answers, we may miss to identify and differentiate among the possibly important factors impeding the effective functioning of cognition. This is the information that is not present in the correct answers alone.

In the field of clinical neuropsychology, it has repeatedly been shown that total achievement scores may not be sensitive to specific cognitive or executive impairments, which has led to the rise of the so-called Boston process approach (Ashendorf, Swenson, & Libon, 2013). Part of this approach focuses on the analysis of errors made. Several studies have demonstrated the clinical utility of various error scores in clinical adult populations (see Possin & Kramer, 2013), that is, the interpretation of error scores was reported to possess incremental validity for differential diagnosis above and beyond total of achievement scores and so may point to a deficit in a specific neural region even when total achievement scores fall within the normal range.

When it comes to predictive validity of interpreting error scores in clinical populations, probably the most frequently interpreted type of these scores are the set-loss errors (and their subtype, the rule-violation errors, if the rule is regarded a part of the current cognitive set). Set-loss error scores reflect a deficit in the ability to maintain set or, more generally, attentional control (Figueroa & Youmans, 2013). The attentional control is one of the most important explanatory constructs in cognitive assessment, especially in more complex tasks that require high-order cognitive control.

This article was published Online First March 2, 2015.

Ivan Ropovik and Monika Bobakova, Department of Preschool and Elementary Education and Psychology, Faculty of Education, University of Presov; Jan Ferjencik, Department of Psychology, Faculty of Arts, University of Pavol Jozef Safarik; Marta Filickova and Iveta Kovalcikova, Department of Preschool and Elementary Education and Psychology, Faculty of Education, University of Presov; Miriam Slavkovska, Department of Psychology, Faculty of Arts, University of Pavol Jozef Safarik.

This work was supported by the Slovak Research and Development Agency (APVV-0281-11).

Correspondence concerning this article should be addressed to Ivan Ropovik, Faculty of Education, University of Presov, 17. Novembra 15, 080 01 Presov, Slovakia. E-mail: ivan.ropovik@unipo.sk

Because weak performance in these kinds of tasks may reflect impairment in quite different cognitive and executive functions, the possible presence of set-loss issue should be one of the baseline considerations prior to the interpretation of higher order cognitive constructs.

Although the exact operating mechanism of set loss is yet unknown, it can be assumed that working memory (WM) is the crucial element here. Hypothetically, once a correct cognitive set is established and kept in some of the storage components of WM (Baddeley, 2000), the executive attentional control component takes over to maintain that set undistorted while elaborating and performing on the task. When this system breaks down (due to distraction- or capacity-related issues), the subject loses the current cognitive set, and that is replaced by a new set that can be a distorted or incomplete rendition of the previous one.

So far, set-loss (or rule-violation) errors have been shown to be useful in identifying various disorders connected with frontal lobes dysfunction (e.g., Carey et al., 2008; Cattie et al., 2012; Marton, 2008; McDonald, Delis, Norman, Tecoma, & Iragui, 2005; Possin et al., 2009; Reske, Delis, & Paulus, 2011; Yochim, Baldo, Kane, & Delis, 2009). Apart from showing frontal-anatomic specificity, with prefrontal cortex recruited to prevent set-loss errors, they were reported to reflect a unitary construct, that is, various types of set-loss errors loaded on a single latent dimension (Possin et al., 2009).

The Present Study

Although there are some indications of specific utility in clinical adult populations, the psychological meaning of set-loss error scores in normal and especially the children's population is not at all clear. The aim of the present study was to explore the predictive and construct validity of set-loss errors derived from the Delis-Kaplan Executive Function System (D-KEFS) test battery (Delis, Kaplan, & Kramer, 2001) in a population of 8- to 11-year-old children. It was expected that the set-loss errors might possess relevant qualities for predicting some important cognitive or behavioral domains that are educationally relevant to this age.

The design of the study aimed at clarifying three conceptual issues—namely, the identity, meaning, and utility of set-loss error scores as follows. (a) Regarding the identity of set-loss error scores (i.e., questions related to the validity of measurement), the following research questions were examined. First, does a unitary set-loss dimension exist? The assertion that there is a unitary factor underlying the occurrence of these types of error scores requires that a single latent variable accounts for the relationships among set-loss error scores derived from distinct measures (Bollen, 1989; Borsboom, Mellenbergh, & van Heerden, 2003; Markus & Borsboom, 2013). Second, are set-loss errors diverse from other types of errors? Specifically, the focus was on perseverance (or repetition) errors or omission errors. (b) With respect to the psychological meaning, are set-loss errors connected with the overload of WM capacity? If so, is that effect of a direct nature, or is it mediated by deficits in attentional control? Given that less efficient functioning of WM leads to its overload, especially in complex tasks (Sweller, 1988), it can be directly associated with set loss or, possibly, set loss does not have to be a capacity-related issue but can be caused by executive attentional control breakdown. (c) Finally, concerning the utility of set-loss error scores, do these scores have any diagnostically relevant predictive power, and can they be sensibly interpreted in the general population of children? What is their relationship with such theoretically relevant constructs as the ability to learn, WM, attentional control, fluid intelligence, crystallized intelligence, and executive functions-related real-world behaviors (across the entire between-subjects distribution of the propensity to set loss)? What are the predictive properties of linear constructs coined from distinct error scores? Does set loss hinder learning, and is it distinguishable from attentional control? What sensitivity and specificity to predict low functioning in those domains do the set-loss error scores possess? Are they suitable for any decision making at the level of an individual?

Nowadays, a number of cognitive tests included in test batteries such as the D-KEFS (Delis et al., 2001), The Wechsler Intelligence Scale for Children, Fourth Edition (WISC IV) Integrated (Wechsler et al., 2004), or Developmental Neuropsychological Assessment, Second Edition (NEPSY-II) (Korkman, Kirk, & Kemp, 2007) offer error scores that are normed on a representative (thus, in fact, normal) population. But should these scores be brought into play just in case the subject's performance falls within the tail of the distribution, or do they really have some meaning across the entire distribution? What kind of interpretation, if any, do some of the error scores also have in the general population?

Method

Participants and Procedure

The sample comprised 250 Caucasian (Slovak) children, 147 girls and 103 boys, aged from 8 to 11 years (M = 9.7 years; SD = 0.4; interquartile range (IQR) = 9.3–10.2). The children were attending the last (fourth) grade of elementary school. A cluster sampling technique was used to select subjects for the sample. Based on 2011 census data, 15 classes (clusters) of elementary schools were selected, proportionally stratified by the size of residence (into three levels). The mean size of a cluster was 18.1 children. Informed consent for the child's participation in the study was obtained from parents.

Every child was tested individually on a battery of 18 tests by trained psychologists. Sixteen of these tests were used in this study. Testing took place before noon in a quiet room on three occasions, lasting approximately 180 min in total. Moreover, a subset of children (n = 106) was rated by their class teachers. All the teachers had been working with the children for at least 3.5 years on a daily basis.

Measures

D-KEFS (Delis et al., 2001). The D-KEFS is a test battery designed to assess executive functions. The battery is composed of 10 stand-alone tests, of which 6 were used for this study. Following the guidelines of the International Test Commission (2005), the battery was adapted for language and cultural equivalence (Ferjenčík, Bobáková, Kovalčíková, Ropovik, & Slavkovská, 2014).

Verbal Fluency Test. In three subtests mirroring phonemic, semantic, and switching fluency, the subject was required to generate as many words as possible—within a 60-s time limit and under restricted search conditions (words beginning only with a certain letter or belonging to a defined category). The total number of responses violating the task rules provided a measure of set loss.

Design Fluency Test. In the Design Fluency Test, a measure of fluency in the nonverbal domain, the task was to produce as many novel abstract designs in 60 s as possible by connecting dots. The total number of designs violating errors across the three subtests served as the dependent measure of set loss.

Trail Making Test. The Trail Making Test, consisting of five subtests, was designed to measure visual attention, set shifting, and motor speed. The tasks relevant for this study included connecting encircled numbers or letters in proper order (Subtests 2 and 3) and linking numbers with letters in alternating order (Subtest 4). The time taken to complete Subtest 2 (Number Sequencing) was used as an indicator of attentional control. The raw number of set-loss errors for Subtests 2, 3, and 4 was totaled to provide a measure of set-loss for the entire test.

Tower Test. The Tower Test is a complex task that required the subject to inhibit prepotent responses, devise a solution plan, hold it in WM, and monitor performance. The task was to move disks across three pegs according to rules to reach a given goal state with as few moves as possible. The number of rule violation moves per administered item was used as the dependent measure.

Word Context Test. In the Word Context Test, a measure of abstract thinking, deductive reasoning, and cognitive flexibility, the subject was required to infer the meaning of made-up words by means of contextual clues that became increasingly more specific after every guess. The set loss was indicated by the raw number of correct-to-incorrect errors.

Color-Word Interference Test. In the Color-Word Interference Test, a variation on the Stroop test, the color-naming condition (Subtest 1) score was used as one of the indicators of attentional control.

Woodcock-Johnson International Editions (W-J IE; Ruef, Furman, & Muñoz-Sandoval, 2003). The W-J IE, a shortened and adapted version of the Woodcock-Johnson (WJ)–Revised (Woodcock, Johnson, & Mather, 1989), consists of 10 tests that aimed to measure seven of the Cattell-Horn-Carroll broad abilities. The raw scores of Picture Vocabulary, Synonyms, Antonyms, and Verbal Analogies subtests were used as the indicators of crystallized intelligence (*Gc*). Similarly, Memory for Names, Spatial Relations, Quantitative Reasoning, Visual Matching, and Numbers Reversed subtests were used to indicate fluid intelligence (*Gf*).

AnimaLogica (Stevenson, Hickendorff, Resing, Heiser, & De Boeck, 2013). The AnimaLogica measure, a nonverbal, fully computerized adapted dynamic test measuring the ability to learn in the domain of figural analogies, employed a pretest-intervention-posttest design, in which pretest and posttest (20 items each) were designed as isomorphic measures with no help provided. The intervention (teaching) phase followed the graduated-prompt procedure (Campione & Brown, 1987) that was based on a series of five hints (from metacognitive through cognitive to solution-constructing prompts), progressively revealing the solution in each of the 10 items. Within a 2 × 2 matrix, the subject was required to place the missing animal figures to complete the analogy. There were between two to eight variations of the animal figures according to their number, size, color, orientation, and position. The AnimaLogica posttest score, reflecting initial performance as well as the effects of learning, was used to derive the ability to learn.

Behavior Rating Inventory of Executive Function–Teacher Form (Gioia, Isquith, Guy, Kenworthy, & Ptáček, 2011). The Behavior Rating Inventory of Executive Function–Teacher Form, an 86-item rating scale for teachers, was designed to assess children's executive functions as manifested in everyday behavior. The 3-point Likert-scaled items gave rise to eight clinical scales: Initiate, Working Memory, Plan/Organize, Organization of Materials, Monitor, Inhibit, Shift, and Emotional Control. The raw sum scores of each scale were used as a measure of the respective behavioral domain.

Results

Preliminary Analyses

In the first step, the data were screened for missing or improbable values and univariate outliers. There were almost no missing data (<0.1%), and the multiple imputation method was used to handle those few. As a general rule, given that it fell within probable bounds, no value was regarded an outlier for the error scores. To screen for severely outlying values in variables with expected normal distribution, a matrix of *z* scores was created to check that no more than three excessive values appeared ($x > M \pm 2\, SD$). For outlying cases, a raw score was assigned that was one unit larger (or smaller) than the next most extreme score in the distribution of the offending variable (Tabachnick & Fidell, 2007). No nonlinear transformations were performed because error scores tend to be highly positively skewed and leptokurtic by nature, and so there is no need to enforce any shift toward normal distribution.

All variables were tested for age, gender, and group effects. Regarding the age (two levels split by median age), significant differences were found for *Gf* and WM, however, of small effect sizes: $r = .20$ and $r = .16$, respectively. No significant gender differences were found in the studied variables. Similarly, there were no significant classroom differences (with Bonferroni correction) on any of the variables.

In the second step, the following variables that were needed for further analyses were derived:

Ability to learn. To statistically isolate the ability to learn such that the variance in the AnimaLogica posttest accounted for by analogical reasoning was removed, we computed an unstandardized residual score by regressing the AnimaLogica posttest (partial-credit score) onto the Verbal Analogies subtest of the W-J test battery. For that purpose, the pretest score was not used due to the overrepresentation of error variance in the raw gain score in the case of large pretest-posttest correlation, which was the case here ($r = .70$).

Attentional control. A linear composite of weighted observed variables was created out of the three indicators of attentional control—namely, the Visual Matching test (W-J IE), the Number Sequencing condition from the Trail Making Test, and the Color Naming condition from the Color-Word Interference Test. The intercorrelations ranged from $r_s = .38$ to $r_s = .44$ in absolute value (irrespective of scaling). The principal component analysis yielded a single component (eigenvalue = 1.8) accounting for 62% of the overall variance. The marginally acceptable value of the Kaiser-Meyer-Olkin measure (.67) and a significant Bartlett's test ($p <$

.001) implied the adequacy of such extraction. The variable was scaled inversely.

WM. WM was operationally defined as the performance in the Numbers Reversed subtest (W-J IE).

***Gc*.** A factor analysis using the maximum likelihood fit function was performed. A unitary factor model proved to fit the data well, with $\chi^2(2) = 0.3$, $p = .87$. The single latent variable accounted for 47% of the variance in its respective indicators (see the "Measures" section), with loadings ranging from .58 to .76. A regression factor score was estimated and checked for normality to serve as the dependent measure of *Gc*.

***Gf*.** In the same way, a one-factor model proved to be consistent with the data, given the maximum likelihood $\chi^2(5) = 2.8$ with associated probability $p = .73$. This latent variable explained 41% of the variance in five *Gf* indicators (see the "Measures" section), with factor loadings ranging from .52 to .79. A regression factor score was estimated, checked for normality, and further used as a dependent measure of *Gf*.

Identity of Set-Loss Error Scores

Distribution and dimensionality. As expected, all the set-loss error scores were positively skewed and leptokurtic with *z-skew* and *z-kurtosis* significant at $p < .001$. Very large skew and kurtosis values were found especially for the Trail Making Test and Tower Test with skewness ($SE = .15$) of 3.4 and 3.2, respectively, and kurtosis ($SE = .31$) of 16.3 and 13.6, respectively. Generally, there was a high proportion of subjects making no errors (50%, 13%, 68%, 38%, and 46%, respectively, for the Verbal Fluency Test, Design Fluency Test, Trail Making Test, Tower Test, and Word Context Test, respectively); however, children this age still make more errors than did adults (see Azuma, 2004, for review).

Regarding the dimensionality of these error scores, a confirmatory factor analysis (CFA) was carried out where all the set-loss error scores loaded on a single latent variable. The analysis was conducted using M*plus* 6.12 (Muthén & Muthén, 1998/2011). To account for the high proportion of zeros (i.e., highly nonnormal, positively skewed distributions of the set-loss error scores), we used a mean- and variance-adjusted weighted least squares fit function to analyze the covariance matrix, with an adjusted χ^2 similar to that of the robust Yuan-Bentler scaled $T2^*$ test statistic. The variables were specified as being censored from below at zero and analyzed using a Tobit CFA model (see Muthén, 1989). At first, some convergence problems were caused by the Word Context Test indicator because all the pairwise correlations of that test were statistically not different from zero. After that indicator had been dropped, fitting of the model to the sample covariance matrix converged to an admissible solution. The single latent variable reproduced the observed covariances well, with $\chi^2 = 0.3$, $df = 2$, $p = .85$, favorable approximate fit indices (comparative fit index = 1.0; root mean square error of approximation = .00; 90% confidence interval [.00, .08]; weighted root mean square residual = .12), and nonsignificant standardized residuals.

However, despite the technical unidimensionality of the set-loss error scores, such a model was not practically interpretable. As can be seen in Table 1, all the factor loadings were rather low and not significant. In fact, it was very easy for the factorial model to account for the intercorrelations because of their very low magnitude in general (see Table 2; reported Spearman's rho coefficients were adjusted for tied ranks by Kendall's correction). Given the $N = 250$ and $\alpha = .05$, there was sufficient power of $(1 - \beta) = .80$ to detect an effect as low as $r = .16$, but only 3 of 10 correlations reached significance. Although Bartlett's test was still significant, the correlation matrix came close to an identity matrix. Moreover, rescaling the variables in ordinal discrete categories (aggregating the high scores) did not bring about any significant change. It can be concluded that although it is technically possible to identify a set-loss error dimension, it is represented by the observed set-loss scores in a very limited way and is practically inconsequential.

Table 1
Factor Loadings

Set-loss measured by	λ (R^2)	*SE*	*p*
Verbal Fluency Test	.74 (.55)	.46	.11
Design Fluency Test	.34 (.12)	.21	.10
Trail Making Test	.21 (.04)	.14	.14
Tower Test	.06 (.00)	.17	.71

Note. λ = factor loading; *SE* = standard error.

Divergence of error scores. The other question concerned whether the set-loss errors are diverse from other types of errors, specifically from perseverance (or repetition) errors, sequencing errors, omission errors, or uncorrected interference errors. The results (see Table 3) show that they are highly distinct given that the variance overlap (R^2) ranged from .00 to .07. The number of committed set-loss errors thus seems to say very little about the subject's propensity to make other kinds of errors.

Psychological Meaning of Set Loss

The second research question addressed whether the set-loss errors are connected with the overload of WM capacity and, if so, whether that effect is mediated by deficits in attentional control. Despite low variability in those kinds of scores, three of five set-loss error scores correlated negatively with the measure of WM negatively, as expected (see Table 4). Moreover, controlling for WM caused all but one partial correlation (between Verbal and Design Fluency Tests) to become essentially zero (all correlations *ns*, mean $r < .09$), which means that it might be the WM that accounts for the relationships (albeit weak) between set-loss error scores. Here, two indirect effects were detected. First, the relationship between the Trail Making Test set-loss errors and WM ($r_s = .13$, $p < .05$) was found to be moderated by attentional control. Second, the relationship between the Tower Test set-loss score and WM ($r_s = .28$, $p < .01$) was, on the other hand, fully mediated by attentional control.

Predictive Utility of Set-Loss Error Scores

The third and final framework question tackled the issue of whether the set-loss error scores possess any diagnostically relevant predictive power in the general population of children. Table 4 shows that none of the set-loss scores could predict the ability to learn. Almost no evidence of predictive validity for *Gc* was found, but there were some relationships especially with attentional con-

Table 2
Set-loss Error Intercorrelations

	Verbal Fluency	Design Fluency	Trail Making Test	Tower Test	Word Context Test
Verbal Fluency	1.3 (1.0) [1.8]				
Design Fluency	.17**	3.1 (2.0) [3.2]			
Trail Making Test	.10	.13*	0.5 (0.0) [1.0]		
Tower Test	.18**	.06	.08	0.3 (0.1) [0.5]	
Word Context Test	.07	−.02	.00	−.01	0.8 (1.0) [0.9]

Note. Spearman's rho (adjusted for tied ranks). Means (medians) [standard deviations] are presented on the diagonal.
* $p < .05$ two-tailed. ** $p < .01$, two-tailed.

trol, *Gf*, and WM, as mentioned. Although several of the correlations reached significance, the effect sizes were of low magnitude.

Within the question of predictive utility, one of the exploratory goals was to establish whether there are some relationships with executive functioning–related behavioral manifestations. As shown in Table 5, only 3 of 40 intercorrelations of set-loss error scores and eight Behavior Rating Inventory of Executive Function scales reached significance. Yet, even without considering the "crud factor" (Meehl, 1990), with an $\alpha = .05$, one can always expect 2 of 40 correlations to be significant just by chance.

One of the most frequent justifications for the absence of predictive power is the low reliability of observed scores. This problem is usually solved by creating sum scores or parcels that are supposed to be more reliable (Nunnally & Bernstein, 1994). This was not the case with set-loss error scores, simply because such a construct was not internally consistent and truly unidimensional, as shown by the results of the CFA (i.e., the zero factor loadings). Here, the rank correlations were −.03, −.08, −.18**, −.01, and −.11 (** $p < .01$, two-tailed), respectively, for the ability to learn, attentional control, WM, *Gc*, and *Gf*, respectively. At the same time, creating a formative construct (whether weighted or unweighted) out of diverse kinds of error scores does not do much better. The inclusion of repetition, sequencing, or omission errors in various combinations increased the predictive power, however, only by a small margin, with up to 6% and 5% of explained variance in WM and *Gf*, respectively.

Now, because any type of correlation is undefined in the absence of variability, it is almost certain that the relationships found across the entire sample were attenuated by a high proportion of subjects stacked on the very left side of the set-loss error score distributions, making none or almost no errors. But if the subject commits such errors, does it tell anything? Does only some specific elevated number of errors have the predictive quality of indicating the low level of some other focal variables, just like in clinical populations?

For the purpose of finding out whether some discrimination threshold of error commissions can be used to classify subjects by the presence of low functioning in a focal cognitive construct, a receiver operating characteristic (ROC) curve analysis was employed. At first, the subjects were classified according to their score on each target construct (ability to learn, attentional control, WM, *Gc*, and *Gf*) into a "below −1 *SD* group" and "above −0.5 *SD* group" (to eliminate inconsequential differences). The results showed that most of the ROC curves were not very different from the reference line (45-degree diagonal), which indicates a pure chance-level prediction. Exactly speaking, only 5 of 25 set-loss error score prediction curves were able to significantly identify (by the significance of the area under the curve parameter) subjects with a poor level of some of the target cognitive constructs as such. One of those five predictors was even significantly counterperforming.

However, although a beyond-chance level prediction of four set-loss error scores was detected (Design Fluency Test predicting attentional control; Verbal Fluency Test, Trail Making Test, and Tower Test predicting WM), there was no threshold level with sufficient sensitivity and specificity for classifying the subjects. The overall best-performing set-loss error score was that derived from the Trail Making Test for classifying subjects with poor WM (with the cutoff score > 1), but still, the sensitivity and specificity of that cutoff were only 56% and 74%, respectively. Accordingly,

Table 3
Correlations Between Types of Errors

Set-loss errors	VF repetition errors	DF repetition errors	TMT sequencing errors	TMT omission errors	CWIT errors
Verbal Fluency	.21**	.18**	.09	.05	.15*
Design Fluency	.20**	.22**	−.08	−.04	.16*
Trail Making Test	.12	.03	.26**	.07	.05
Tower Test	.17**	.19**	.18**	.00	.12
Word Context Test	.07	.05	.12	−.01	−.01

Note. Spearman's rho (adjusted for tied ranks). VF = Verbal Fluency; DF = Design Fluency; TMT = Trail Making Test; CWIT = Color-Word Interference Test.
* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

Table 4
Intercorrelations Between the Set-Loss Error Scores and Relevant Cognitive Constructs

	Ability to learn	Attentional control	Working memory	*Gc*	*Gf*
Verbal Fluency	.03	.01	−.16*	−.02	−.13*
Design Fluency	−.02	−.15*	−.05	.06	.04
Trail Making Test	−.05	.13*	−.23**	−.04	−.18**
Tower Test	−.05	.30**	−.20**	−.03	−.13*
Word Context Test	.07	−.10	−.01	−.17**	−.08

Note. Spearman's *rho* (adjusted for tied ranks). *Gc* = crystallized intelligence; *Gf* = fluid intelligence.
* $p < .05$ two-tailed. ** $p < .01$, two-tailed.

Table 5
Correlations Between Set-Loss Errors and Real-World Behavior

	BRIEF (teacher rating scales)							
	Inhibit	Shifting	Emotional control	Initiate	Working memory	Plan	Organization	Monitor
Verbal Fluency	.10	.12	−.01	.04	.07	.16	.14	.05
Design Fluency	.21*	.05	−.01	.16	.00	.15	.22**	.18
Trail Making Test	.15	−.05	.01	.15	.16	.17	.07	.14
Tower Test	.08	.18	.19	−.04	.11	.00	.09	.09
Word Context Test	−.09	.09	.07	−.05	−.10	−.09	−.22**	−.12

Note. $n = 106$. Spearman's rho (adjusted for tied ranks). BRIEF = Behavior Rating Inventory of Executive Function.
* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

the positive and negative predictive values were .29 and .89, respectively. The ROC curves for executive functions-related real-world behavior looked very similar, that is, the tendency to commit set-loss errors did not help to identify children rated as subnormal on behavioral scales.

Discussion

The promise of the utility of the set-loss error scores in the general population of children seems to remain undelivered. To sum up, the systematic variance in every one of these scores seems to be of different causal origin, raising doubts about the existence of a set-loss dimension in the general population. Given only very weak correlations between the set-loss errors and theoretically relevant constructs as well as among the error scores themselves, it seems that there is very little reliable variance in the studied population (see Rabbitt, 1997). The problem is not solved by summing individual set-loss error scores since the reliable part of their variance is most probably of different causal origin. Moreover, the evidence (the analysis of specificity and sensitivity) suggests poor measurement and predictive qualities of the set-loss errors not just at the between-subjects level but also at the level of specific individual-related implications drawn from these kinds of scores.

Although set-loss errors proved to have some diagnostic utility in clinical samples (Carey et al., 2008; Cattie et al., 2012; Marton, 2008; McDonald et al., 2005; Possin et al., 2009; Reske et al., 2011; Yochim et al., 2009), there is very little evidence in favor of their utility in the general population. After accounting for their naturally skewed and censored character, the analyses showed that (even if technically possible to find) there was no meaningful latent dimension that would explain the variance across these kinds of scores, given that all the factor loadings were nonsignificant. Because association is one of the three general necessary preconditions for being able to speak in ontological terms about an underlying causal mechanism of set loss, the failure to find meaningful relationships is a blow to the measurement validity of set-loss error scores in the general population.

The lack of evidence for such construct in the general population might raise the question whether it was reasonable to expect concordance between such diverse measures at all. However, if we define set-loss error as the breakdown in the adherence to task-specific rules or attentional demands (i.e., maintaining vigilant focus and avoiding distractions), all the measures derived from the tests employed here conform to this conceptualization. Moreover, given a common set-loss label, a typical end user is justified to expect the existence of a single latent capacity preventing these errors from occurring. Almost no association between the set-loss error scores, however, leaves the psychologist in the dark regarding the meaning and interpretation of these scores. Thus, if a subject scores high on set loss in one test but commits almost no errors in the other, it is very difficult to form any judgment.

Apart from no or very little association between set-loss error scores, there were also only marginal associations between the error scores in general. These findings raise doubts whether parceling diverse set-loss error scores into a single sum score produces anything but a variable with a prevailing proportion of task-specific and random variance. However counterintuitive it may sound, summing several set-loss error scores will not make the composite error score more reliable but, to the contrary, essentially more random-like. Likewise, coining a formative construct out of diverse types of error scores did not lead to an increase in predictive power. In search for the validity related to set-loss errors, the data show rather a minimal or no overlapping variance with some of the focal cognitive constructs. Surprisingly, set loss, that is, the ability to maintain set as measured by error scores, was empirically shown not to be connected to the ability to learn, whereas other research using latent variable analysis demonstrated that there is a strong indirect effect of attentional control (a superordinate construct) on the ability to learn, mediated via WM (Ropovik, 2014). The predictive power regarding *Gf* and *Gc*, or the executive functioning–related real-world behaviors as measured by rating scales, turned out to be very weak as well. However, in line with expectations, some of the set-loss error scores were related to WM and attentional control (Figueroa & Youmans, 2013; Possin et al., 2009). The indirect effects between them lend some support to the hypothesis that set loss can be caused by the attentional control–related breakdown of WM that is holding a cognitive set “online” (see Miller, 2000; Miller & Cohen, 2001). This would fit with some of the prominent theories of executive functioning such as the supervisory attentional system (Norman & Shallice, 1986) and working memory model (Baddeley, 2000). To summarize, there are only a few insufficient indications regarding the measurement and predictive qualities of the set-loss errors, and even those few indications are too weak to provide for any solid conclusions about the identity of these error scores in the given population.

Although the data speak against the notion of validity for measurement, that fact is usually of little concern for a practitioner psychologist provided that the score has some (incremental) validity for prediction and selection (Borsboom, Mellenbergh, & van Heerden, 2004). As mentioned, the set-loss error scores largely failed to reliably predict some of the theoretically relevant constructs across the entire distributions—that is, there was only very limited predictive power (and thus no incremental validity) for ranking the full range of subjects on the target construct by their propensity to set loss. The next step was to focus on the predictive power of an observed tendency to commit an elevated number of errors.

The ROC curve analysis showed that it was not possible with any of the set-loss error scores to establish a cutoff with sufficient sensitivity and specificity. Most of the ROC curves indicated that using set-loss error scores to pick up those with a low level of the target cognitive construct does no better than tossing a coin. Applying Bayes's theorem (Bayes & Price, 1763), it is not difficult to see that even a 56% sensitivity and a 74% specificity (the values of the best-classifying cutoff) are not enough for effective decision making at the level of an individual (see Lee, 2012). With a base rate of 16% (the proportion of subjects with the level of the target construct below -1 *SD*), exceeding the threshold level of errors slides the conditional probability of a case being a true positive from the prior probability of 16% to the posterior probability of just 29% (i.e., the positive predictive power). Hypothetically, when trying to detect a deficit of even lower incidence in the general population, say 5%, such sensitivity and specificity of set-loss error scores would produce a posterior probability as low as about 9%. Small bivariate correlations between the set-loss error scores and constructs such as attentional control, otherwise regarded as indicating at least some predictive power at the interindividual level, were thus shown to be insufficient to drive decisions and judgments regarding an individual.

Limitations

This study has some limitations that deserve mention. First, the findings and conclusions regarding the utility of set-loss errors apply only to the general population of 8- to 11-year-old children. Given the dynamic character of the child's development of executive functions, which is marked by substantial qualitative changes of their internal structure (Anderson, 2002; Huizinga, Dolan, & van der Molen, 2006), the conclusions regarding the utility of set-loss errors could potentially be more positive in older children, adolescents, or adults. Second, all the conclusions concerning predictive validity were based exclusively on concurrent evidence. It is thus still possible that the set-loss error scores might more reliably identify future underachievers, for instance. Last, although the measures used in this research are some of the most conventional ones, none were primarily designed to measure the ability to maintain set (Strauss, Sherman, & Spreen, 2006). The weak measurement and predictive properties of the set-loss error scores do not imply that the ability to maintain set is not important for effective cognitive functioning and that it is inconsequential as such. Instead, it may just mean that the set-loss error scores do not reflect the ability to maintain set well enough for them to be useful in the general population. For now, it has to be concluded that no empirical grounds would support the use of set-loss error scores in the general population of children.

References

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology, 8,* 71–82. http://dx.doi.org/10.1076/chin.8.2.71.8724

Ashendorf, L., Swenson, R., & Libon, D. J. (2013). *The Boston process approach to neuropsychological assessment: A practitioner's guide*. New York, NY: Oxford University Press.

Azuma, T. (2004). Working memory and perseveration in verbal fluency. *Neuropsychology, 18,* 69–77. http://dx.doi.org/10.1037/0894-4105.18.1.69

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423. http://dx.doi.org/10.1016/S1364-6613(00)01538-2

Bayes, T., & Price, M. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions of the Royal Society of London, 53,* 370–418. http://dx.doi.org/10.1098/rstl.1763.0053

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley. http://dx.doi.org/10.1002/9781118619179

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219. http://dx.doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82–109). New York, NY: Guilford.

Carey, C. L., Woods, S. P., Damon, J., Halabi, C., Dean, D., Delis, D. C., . . . Kramer, J. H. (2008). Discriminant validity and neuroanatomical correlates of rule monitoring in frontotemporal dementia and Alzheimer's disease. *Neuropsychologia, 46,* 1081–1087. http://dx.doi.org/10.1016/j.neuropsychologia.2007.11.001

Cattie, J. E., Doyle, K., Weber, E., Grant, I., Woods, S. P., & the HIV Neurobehavioral Research Program (HNRP) Group. (2012). Planning deficits in HIV-associated neurocognitive disorders: Component processes, cognitive correlates, and implications for everyday functioning. *Journal of Clinical and Experimental Neuropsychology, 34,* 906–918. http://dx.doi.org/10.1080/13803395.2012.692772

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System (D-KEFS)*. San Antonio, TX: The Psychological Corporation.

Ferjenčík, J., Bobáková, M., Kovalčíková, I., Ropovik, I., & Slavkovská, M. (2014). Proces a vybrané výsledky Slovenskej adaptácie Delis-Kaplanovej systému exekutivnych funkcií D-KEFS [Process and selected results of Slovak adaptation of Delis-Kaplan System of Executive Functions D-KEFS]. *Československá Psychologie, 58,* 543–558.

Figueroa, I. J., & Youmans, R. J. (2013). Failure to maintain set: A measure of distractibility or cognitive flexibility? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57,* 828–832. http://dx.doi.org/10.1177/1541931213571180

Gioia, G. A., Isquith, P. K., Guy, S. C., Kenworthy, L., & Ptáček, R. (2011). *BRIEF—Hodnocení exekutivních funkcí u dětí* [Behavioral Rating of Executive Functions in Children]. Praha, Czech Republic: Hogrefe-Testcentrum.

Huizinga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia, 44,* 2017–2036. http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.010

International Test Commission. (2005). International Test Commission guidelines for translating and adapting tests. Retrieved from www.intestcom.org

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II: Administration manual*. San Antonio, TX: The Psychological Corporation.

Lee, P. M. (2012). *Bayesian statistics: An introduction*. NY: John Wiley.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Marton, K. (2008). Visuo-spatial processing and executive functions in children with specific language impairment. *International Journal of Language & Communication Disorders/Royal College of Speech & Language Therapists, 43,* 181–200. http://dx.doi.org/10.1080/16066350701340719

McDonald, C. R., Delis, D. C., Norman, M. A., Tecoma, E. S., & Iragui, V. J. (2005). Discriminating patients with frontal-lobe epilepsy and temporal-lobe epilepsy: Utility of a multilevel design fluency test. *Neuropsychology, 19,* 806–813. http://dx.doi.org/10.1037/0894-4105.19.6.806

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244. http://dx.doi.org/10.2466/pr0.1990.66.1.195

Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience, 1,* 59–65. http://dx.doi.org/10.1038/35036228

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202. http://dx.doi.org/10.1146/annurev.neuro.24.1.167

Muthén, B. O. (1989). Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology, 42,* 241–250. http://dx.doi.org/10.1111/j.2044-8317.1989.tb00913.x

Muthén, L. K., & Muthén, B. O. (1998–2011). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (pp. 1–18). New York, NY: Plenum.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Possin, K. L., Brambati, S. M., Rosen, H. J., Johnson, J. K., PA, J., Weiner, M. W., . . . Kramer, J. H. (2009). Rule violation errors are associated with right lateral prefrontal cortex atrophy in neurodegenerative disease. *Journal of the International Neuropsychological Society, 15,* 354–364. http://dx.doi.org/10.1017/S135561770909050X

Possin, K. L., & Kramer, J. H. (2013). Error analysis of the Delis-Kaplan Executive Function System. In L. Ashendorf, R. Swenson, & D. J. Libon (Eds.), *The Boston process approach to neuropsychological assessment: A practitioner's guide* (pp. 122–133). New York, NY: New York: Oxford University Press.

Rabbitt, P. (1997). Introduction: Methodologies and models in the study of executive function. In P. Rabbit (Ed.) *Methodology of frontal and executive function* (pp. 1–38). East Sussex, UK: Psychology Press.

Reske, M., Delis, D. C., & Paulus, M. P. (2011). Evidence for subtle verbal fluency deficits in occasional stimulant users: Quick to play loose with verbal rules. *Journal of Psychiatric Research, 45,* 361–368. http://dx.doi.org/10.1016/j.jpsychires.2010.07.005

Ropovik, I. (2014). Do executive functions predict the ability to learn problem-solving principles? *Intelligence, 44,* 64–74. http://dx.doi.org/10.1016/j.intell.2014.03.002

Ruef, M., Furman, A., & Muñoz-Sandoval, A. (2003). *Woodcock-Johnson: International editions: Administration manual: Slovak edition*. Nashville, TN: The Woodcock-Muñoz Foundation.

Stevenson, C. E., Hickendorff, M., Resing, W. C. M., Heiser, W. J., & De Boeck, P. A. L. (2013). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence, 41,* 157–168. http://dx.doi.org/10.1016/j.intell.2013.01.003

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. New York, NY: Oxford University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285. http://dx.doi.org/10.1207/s15516709cog1202_4

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson/Allyn & Bacon.

Wechsler, D., Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maerlander, A. (2004). *WISC-IV integrated*. San Antonio, TX: Harcourt Assessment.

Woodcock, R. W., Johnson, M. B., & Mather, N. (1989). *Woodcock-Johnson Psycho-Educational Battery–Revised*. Allen, TX: DLM Teaching Resources.

Yochim, B. P., Baldo, J. V., Kane, K. D., & Delis, D. C. (2009). D-KEFS Tower Test performance in patients with lateral prefrontal cortex lesions: The importance of error monitoring. *Journal of Clinical and Experimental Neuropsychology, 31,* 658–663. http://dx.doi.org/10.1080/13803390802448669

Received September 10, 2014
Revision received January 26, 2015
Accepted January 26, 2015 ■