Psychological Assessment

Methodology of Performance Scoring in the d2 Sustained-Attention Test: Cumulative-Reliability Functions and Practical Guidelines

Michael B. Steinborn, Robert Langner, Hagen C. Flehmig, and Lynn Huestegge

Online First Publication, April 13, 2017. http://dx.doi.org/10.1037/pas0000482

Psychological Assessment

1040-3590/17/$12.00 http://dx.doi.org/10.1037/pas0000482

Methodology of Performance Scoring in the d2 Sustained-Attention Test: Cumulative-Reliability Functions and Practical Guidelines

Michael B. Steinborn
University of Wuerzburg

Robert Langner
Heinrich Heine University Düsseldorf and Research Centre Jülich, Jülich, Germany

Hagen C. Flehmig
Saxon Hospital Großschweidnitz, Großscheidnitz, Germany

Lynn Huestegge
University of Wuerzburg

We provide a psychometric analysis of commonly used performance indices of the d2 sustained-attention test, and give methodical guidelines and recommendations, based on this research. We examined experimental effects of repeated testing on performance speed and accuracy (omission and commission errors), and further evaluated aspects of test reliability by means of cumulative reliability function (CRF) analysis. These aspects were also examined for a number of alternative (yet commonly used) scoring techniques and valuation methods. Results indicate that performance is sensitive to change, both differentially within (time-on-task) and between (test–retest) sessions. These effects did not severely affect test reliability, since perfect score reliability was observed for measures of speed (and was even preserved with half the test length) while variability and error scores were more problematic with respect to reliability. Notably, limitations particularly hold for commission but less so for omission errors. Our recommendations to researchers and practitioners are that (a) only the speed score (and error-corrected speed score) is eligible for highly reliable assessment, that (b) error scores might be used as a secondary measure (e.g., to check for aberrant behavior), that (c) variability scores might not be used at all. Given the exceptional reliability of performance speed, and (d) test length may be reduced up to 50%, if necessary for time-economic reasons, to serve purposes of population screening and field assessment.

Public Significance Statement
This study found that some commonly used scoring techniques to assess sustained-attention performance are problematic with regard to measurement accuracy (test reliability). It further advances the use of cumulative reliability function analysis for purposes of comparably evaluating tests or scoring techniques. The recommendation to practitioners is that only scores of average performance speed (but none of variability) are eligible for highly reliable psychometric assessment of basic cognitive functioning.

Keywords: psychometric testing, visual search, sustained attention, concentration

According to Cronbach (1975), the ultimate goal of any psychometric test is to globally predict relevant criteria. This is especially important to practitioners in applied fields such as school psychology or in clinical contexts such as neuropsychological assessment and rehabilitation, being dependent on commercially available tests and assessment batteries. However, because there are a myriad of validity criteria and areas of application, demonstrating predictive validity to one among many criteria are not eligible as a means to judge a test with respect of its general predictive value. Instead, as Cronbach (1975) argues, a performance test should in the first instance be judged by its ability to reproduce an observed pattern of individual differences in a standard target sample, and thereafter might be adapted to rather specific contexts to probe and optimize criterion validity. In other words, a performance test should initially be judged by its measurement accuracy as revealed from a high-quality sample, which forms the basis for research on validity in clinical sample populations. Reproducibility of measurement is usually assessed by means of the correlations between individuals' performance in parallel test forms or in repeated test applications, administered either simultaneously (split-half model reliability) or at successive

Michael B. Steinborn, Methods and Applied Cognitive Psychology, University of Wuerzburg; Robert Langner, Institute of Clinical Neuroscience and Medical Psychology, Heinrich Heine University Düsseldorf, and Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany; Hagen C. Flehmig, Department of Forensic Psychiatry, Saxon Hospital Großschweidnitz, Großscheidnitz, Germany; Lynn Huestegge, Methods and Applied Cognitive Psychology, University of Wuerzburg.

Correspondence concerning this article should be addressed to Michael B. Steinborn, Department of Psychology III, University of Wuerzburg, Roentgenring 11, 97070 Wuerzburg. E-mail: michael.steinborn@uni-wuerzburg.de; michael.b.steinborn@gmail.com

occasions (test–retest model reliability). Because reliability coefficients are known to increase with the number of measurement units, any psychometric analysis of speeded test performance should also elucidate the functional relation between measurement accuracy (reliability) and test length (Miller & Ulrich, 2013).

Theoretical Considerations: Psychometric Assessment

Psychometric performance tests aiming to assess sustained attention and concentration ability have been widely used since the earliest times in the history of psychology. Since the pioneering work of Kraepelin (1902), these tests have been conceptualized as self-paced tests, which require individuals to engage in continuous responding to a series of target items that are arranged one after another. His work curve test required individuals to perform as many operations (addition of two digits in adjacent columns) as they could within 90 min, with average speed, accuracy, and persistence (variability) as the main performance dimensions. Another historically important psychometric test was developed by the French psychologist Benjamin Bienaimé Bourdon (1860–1934), where individuals were required to cancel out target items (4 dots) among distracters (less or more than 4 dots) on a one-page paper sheet, within a certain time-limit. Generally, these tests are often conceptualized as either letter cancellation, mental arithmetic, or coding tasks, although the particular item characteristic is not of primary importance (Vanbreukelen et al., 1995). Common to all these tests is that they are administered in a self-paced mode, either as a paper-and-pencil based or computerized version, which requires individuals to continuously monitor their current performance speed and accuracy (and to adjust it, whenever necessary) over the testing period.

A specific feature of these tests is the opportunity to simultaneously consider several performance aspects, such as average performance speed, error rate, and variability (Pieters, 1983, 1985; Vanbreukelen et al., 1995). Some authors prefer a measure of throughput, which is the rate of work in a given time frame (cf. Szalma & Teo, 2012; Thorne, 2006). According to Flehmig et al. (2007), the exact utility of either performance speed or error rate depends on several aspects of test construction, that is, on its measurement intention, its complexity, response mode, or test length (Schweizer, 1996, 2001). Manual guidelines for commercially available psychometric tests usually recommend to combine measures of speed and accuracy into a single compound dimension (Brickenkamp, 1962, 1966), although this proposal has also been criticized (Westhoff, 1985; Westhoff & Lemme, 1988). In fact, one of the problems is that speed and error rate are often considered distinct, equipotent, and psychometrically reliable dimensions of ability as reasoned from small (or absent) correlative relationships between both measures (Bates & Lemay, 2004). It is often ignored, however, that error scores (being rare events) usually lack test reliability, which severely limits correlative relationships to other performance indices or validity criteria (Hagemeister, 2007). Further, there is some disagreement as to the correct calculation instruction, because different calculation rules lead to different performance outcomes, which might engender confusion among clinical and neuropsychological practitioners as to whether a chosen scoring technique is being correctly applied. This is very problematic and essentially formed the motivational basis of the present work.

To judge the psychometric quality of speeded tests, one has to demonstrate the measurement accuracy of the given test, which, in classical test theory, is defined as the correlation of the same test at repeated (i.e., parallel-test, split-half, and test–retest model) administrations (Cronbach, 1975, pp. 190–214; Lord & Novick, 1968, p. 61). As alluded to above, reliability is well known to increase with the number of trials administered in a test, such that the reliability coefficient is often low ($r < .50$) for very small numbers but quite rapidly increases to a sufficient degree ($r > .85$) with the lengthening of the test (Miller & Ulrich, 2013, p. 824). Although many standard textbooks extensively cover reliability issues, the fundamental questions of what exactly determines test reliability and of how these correlations are related to the underlying mental processes have rarely been addressed. In fact, psychometricians regard reliability issues as a purely statistical, not a psychological issue, to be resolved simply by means of true-score accumulation methods. Spearman (1910) and Brown (1910), for example, were the first to develop a method to predict the reliability of a test after changing (i.e., after shortening) its length. According to their reasoning, there is a nonlinear relationship of reliability with test length because as test length increases, true-score variance quadrupled with a doubling of error variance. Yet, the formula is agnostic with respect to psychological processes, and predictions will not be strictly accurate in empirical data (Miller & Ulrich, 2013).

Theoretical Considerations: Cognitive Processes

Miller and Ulrich (2013) conducted a formal and systematic investigation regarding the question of what aspects of mental processing affect reliability of speeded performance. According to their view, speed scores (e.g., RT mean) can be considered as being composed of four components: person-specific general processing time (i.e., ability-induced true-score variance, e.g., general processing speed), task-specific processing time (i.e., experimentally induced true-score variance), residual time, and measurement error. In some ways, the model resembles intuitive considerations that have been made earlier by psychometricians, although none of the previous authors has formally explicated reliability components related to mental processing. According to Cronbach (1975, p. 35), for example, ". . . *the distinguishing feature of any ability test is that the test taker is encouraged to earn the best score he can. The goal of the test constructor should be to bring out the person's best possible performance*. . . ." This statement bears a mechanism that taps into what Miller and Ulrich (2013) term person-specific processing component, which includes maximal motivation to perform well, ensured by the test administrator's instruction or by the consequences of performance (cf. Stuss, Meiran, Guzman, Lafleche, & Willmer, 1996).

Actually, to attain and maintain an optimal performance level in sustained-attention tests, individuals are required to constantly deploy effort and cognitive resources to the task at hand (cf. Langner & Eickhoff, 2013; Vanbreukelen et al., 1995). With respect to reliability and construct validity, the specific task form is of less importance than the self-paced mode (Hagemeister, 2007; Krumm, Schmidt-Atzert, & Eschert, 2008). According to Flehmig et al. (2007), attentional fluctuations in self-paced speedtests are reflected in greater performance variability, as indexed by the coefficient of variation, but less so in an increase in error rate

(Steinborn, Flehmig, Westhoff, & Langner, 2008, 2009). In a letter cancellation test, the situation is somewhat different with respect to response mode, because target and distracter items compete for limited capacity during serial search, and individuals must selectively attend to target items while filtering out irrelevant stimuli in a rapid manner (cf. Desimone & Duncan, 1995; Miller & Schröter, 2002). This task form is specific with respect to the processing of visual material that requires rapid perceptual discriminations, and the response mode, since two kinds of (omission and commission) errors are possible. Thus, fluctuations in attentional efficiency might not only induce an increase in intraindividual variability but also in the rate of omission errors (Gropel, Baumeister, & Beckmann, 2014). Thus, these tests stand somewhat between classical speed and vigilance tests (Helton & Russell, 2011a, 2011b).

Brickenkamp (1962) constructed the test d2 in the style of the classic Bourdon principle. The original d2 sustained-attention test is a paper-and-pencil cancellation test, where individuals have to scan for target items (d2) among distracter items through a series of consecutively ordered objects consisting of 14 rows with 47 characters each. The participants are instructed to cancel out as many target symbols as possible, by moving through on the page in a reading-like manner from left to right with a time limit of 20 s per trial (i.e., per row of objects) without a break (with performance being measured as the amount of work carried out within the given time). The overall testing time is 4 min and 40 s. Brickenkamp (1962, 1966) suggested to exclude the first and the last trial because he believed that warm-up and end-spurt (final sprint) effects might pose a danger to measurement accuracy. He further recommended to interpret speed and error rate in a trait-like manner as quantity and quality dimensions of ability (i.e., information-processing speed and diligence) and to interpret the range (i.e., difference between worst and best performance in each of the 14 trials: Max–Min) to index persistence control (cf. Table 1). Others have referred to omission/commission errors as indicating careless/confused attention, and to Max/Min as indicating impulsive/lapsing tendencies (Bates & Lemay, 2004; Pieters, 1985). However, we cannot but note that none of the aforementioned criteria to judge a test's basic psychometric quality were ever examined with sufficient detail (cf. Bornstein, 2011; Cronbach, 1947; Miller & Ulrich, 2013). Lack of psychometric evaluation is in itself problematic, which is why we considered an in-depth investigation of basic psychometric properties necessary.

Present Study

According to Miller and Ulrich (2013), crucial to the understanding of what determines test reliability is to know (a) how situational preconditions reveal true-score (ability-related) variance, (b) how demand characteristics evoke true-score (task-related) variance, (c) to determine factors that tap into residual variance, and (d) to reduce the influence of any variable that produces coincidental effects and thus increases error variance. We here follow this reasoning, although our focus is on the practical use of commonly accepted performance indices (cf. Bates & Lemay, 2004). To this end, we examined three conceptual aspects of test-evaluation criteria, test stability, and measurement accuracy (reliability) at the same occasion (i.e., within a session) and at a repeated occasion (i.e., between sessions). First, we examined experimental effects of repeated testing (test vs. retest) and time-on-task (Trials 1–14) on performance speed and accuracy. This includes, of course, a thorough reexamination of hypothesized warm-up and end-spurt (final sprint) effects. Second, we examined the effects of repeated testing on performance stability, by examining indices of performance variability including the performance speed cumulative distributive function (CDF) of trials. This was done to evaluate whether the observed effects of retesting on average performance speed originate from a global processing speed-up or alternatively from a rather selective speed-up of only the slower CDF percentiles (Miller, 2006; Steinborn, Langner, & Huestegge, 2016).

Third, we examined the within-session reliability of performance speed by means of a multiple split-half model, separately for both the first and the second testing session. Note that in the context of speeded tests where items are per definition homogeneous, this measure gives information about the consistency of test performance on a series of trials. This type of reliability is particularly important during the construction phase, because it is unbiased from time-related changes in mental state and/or mental

Table 1
Description and Calculation of Performance Indices in the d2 Sustained-Attention Test

Score	Calculation method	Description
1. Speed	Number of items worked through	Speed of processing
2. $Speed_C$	Speed minus number of all errors	Speed of processing (error-corrected)
3. Error rate (%)	Percentage of errors	Accuracy of processing (overall)
4. Error (omission)	Number of omitted targets	Processing accuracy (carelessness)
5. Error (commission)	Number of cancelled distractors	Processing accuracy (confusion error)
6. Variability (*SD*)	Standard deviation of speed	Processing variability
7. Variability (CV)	Coefficient of variation of speed	Processing variability
8. Max	Maximum performance speed	Best performance (impulsivity)
9. Min	Minimum performance speed	Worst performance (lapsing)
10. Range (%)	Relative Max-Min difference	Processing variability

Note. The speed-score (and the error-corrected speed score) can be computed both as average performance speed (per trial) and as aggregated performance speed (across all trials). $Speed_C$ = error-corrected measure of speed defined as the average number of correctly cancelled items per row; variability (CV) = coefficient of variation of performance speed; variability (SD) = reaction time standard deviation of performance speed. Note that because these scores are equivalent with respect to descriptive characteristics and correlational relationships, they will not explicitly be distinguished further.

ability (cf. Humphreys & Revelle, 1984, pp. 158–169; Miller & Ulrich, 2013, pp. 821–825). Fourth, we examined the between-session reliability of performance via computing the test–retest reliability coefficient. This type of reliability is notably of practical importance to estimate the reproducibility of measurement, referring to the question of whether the test measures the same construct at different occasions. An important aspect is that we evaluated reliability as a function of test length, by means of cumulative reliability function (CRF) analysis. Although it is held a truism among theoreticians that long tests are more reliable than short ones, empirical findings are often at odds with this view (Ackerman & Kanfer, 2009; Stuss et al., 1996). Most often, this aspect is simply neglected (e.g., Maloney, Risko, Preston, Ansari, & Fugelsang, 2010; Stolz, Besner, & Carr, 2005; Waechter, Stolz, & Besner, 2010). Yet, our understanding of what determines reliability of cognitive processes might partly depend on CRF analysis.

Method

Participants

The study was conducted at the Dresden University of Technology. The objective was to have a diversified student-based sample with an equalized gender ratio (thus including participants not only from the faculty of psychology but also from the humanities, natural sciences, as well as technical studies), which is ideally suited to study reliability issues and other basic psychometric characteristics (cf. Cronbach, 1947; Miller & Ulrich, 2013). The sample consisted of originally 103 individuals, being tested twice within a retest interval of 1 week and under similar conditions (at the same place and at about the same time), three of which did not appear to the retesting session. Thus, data of 100 participants (50 female, 50 male; mean age = 26.6 years, SD = 7.3 years) entered the analysis. Participants had normal or corrected-to-normal vision and reported to be in normal health condition.

Ethical Statements

Informed consent was obtained from the participants regarding their agreement with their participation in this research. Our study was in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All authors declare that there are no conflicts of interests.

Task Description

The d2 test (Brickenkamp, 1962, 1966) is a Bourdon-style paper-and-pencil letter cancellation test that is widely used in both research and applied contexts in Europe. The test can be used as complete package (i.e., results can be obtained and interpreted according to handbook guidelines, etc.), which means that neither specific expertise in cognitive-experimental and psychometric-testing research nor programming skills are required. The test consists of 14 trials. Each trial is a row with 47 "p" and "d" characters being disposed adjacent to one another. The characters have one to four dashes that are configured individually or in pairs above and/or below each letter. The target symbol is a "d" with two dashes (either two dashes above the "d," two dashes below the "d," or one dash above and one dash below the "d"). Thus, all other items are distracters. The participants' task is to cancel out as many target symbols as possible, moving from left to right, with a time limit of 20 s per trial. No pauses are allowed between trials.

Instruction

The standard instruction is, according to handbook guidelines, to perform the task with maximum efficiency, both in equal measures of speed and accuracy. Accordingly, the experimenter is held to inform the participant that he or she has to work both as fast and accurate as possible. At this point, we consider it vital to give the reader some further information related to this issue. First, despite established empirical effects of instructions (cf. Steinborn, Langner, & Huestegge, 2016, for an overview), it should be mentioned that speed–accuracy tradeoff policies are primarily dictated by the task, although within some limits, they might be modulated by (preset) instructions. In particular, task paradigms are commonly distinguished with respect to the origin of errors and, therefore, divided into *low-error domain* and *high-error domain* tasks (Luce, 1986, pp. 281–318). In low-error domain tasks, such is the d2 test, errors occasionally occur because of transient states of mental inefficiency (i.e., impulsivity, lapsing), and error rate is naturally low (2–10%) in these tasks (cf. Vanbreukelen et al., 1995). In high-error domain tasks, errors occur because of either being tricked by distractors activating the wrong response (conflict tasks) or by being confused by crosstalk between impending multiple-operation demands (multitasking-crosstalk paradigms), thus resulting in a relatively high (10–30%) error rate in these tasks (cf. Frings, Rothermund, & Wentura, 2007; Hommel, 1998; Huestegge & Koch, 2013; Huestegge, Pieczykolan, & Koch, 2014; Steinhauser & Hübner, 2006; Thomaschke & Dreisbach, 2015; Thomaschke, Hoffmann, Haering, & Kiesel, 2016).

Standard Scoring Techniques

A description of scoring techniques is displayed in Table 1. As already mentioned, the d2 test is a time-limit test and the basic unit of measurement is the amount of work carried out in a trial within 20 s (measured by the experimenter using a stopwatch). In contrast, work-limit tests usually measure the time needed to carry out a certain amount of work. Computerized test administrations in cognitive-experimental and neuropsychological research are usually work-limited and are referred to as RT tasks. As argued earlier, performance speed should serve as the principal dimension while error rate might serve as the secondary dimension, given that these measures prove to be psychometrically reliable (cf. Bates & Lemay, 2004; Flehmig, Steinborn, Langner, Scholz, et al., 2007). In the Brickenkamp d2 test, performance speed can be defined as the (average) number of items worked through in a particular trial (i.e., in one of the 14 rows, each containing 47 items). It might be appropriate to use an error-corrected measure of performance speed, where the number of errors is subtracted from the number of performed items in a trial (cf. Brickenkamp, 1962). The error score is defined as the (average) number of incorrect responses, which can be transformed into a measure of error percentage. Performance accuracy can be subdivided into two types of errors, omission errors and commission errors, as is the case in similar

scan-and-check paradigms (cf. Corcoran, 1966; Koriat & Greenberg, 1996; Wolfe et al., 2007). The former is defined as a missed response to a target (i.e., a passive error) while the latter is defined as a response to a distracter (i.e., an active error).

Alternative Scoring Techniques

There are good reasons, according to some authors (Graveson, Bauermeister, McKeown, & Bunce, 2016; Jensen, 1992; Leth-Steensen, Elbaz, & Douglas, 2000; Lim & Dinges, 2010; Schmiedek, Lovden, & Lindenberger, 2009; Stuss, Murphy, Binns, & Alexander, 2003; West, Murphy, Armilio, Craik, & Stuss, 2002), to consider performance variability as a third dimension with potential predictive value, though this view has also been questioned recently (Doebler & Scheffler, 2016; Miller & Ulrich, 2013). The precise computation of a performance variability index varies across studies, ranging from simple to more sophisticated statistical computations. First, Brickenkamp (1962, 1966) originally suggested to interpret the performance range between the worst and the best performance (of the 1–14 trials) as an index of variability. Second, the performance *SD* can be used to index variability (Jensen, 1992; Stuss et al., 2003). According to Flehmig et al. (2007), the performance coefficient of variation (CV) is a relativized index (dividing *SD* by the individual mean) that is independent of average performance speed and thus provides a less redundant measure of performance (Saville et al., 2011; Segalowitz & Frenkiel-Fishman, 2005). Here, we evaluated all three variability scores (cf., Steinborn, Langner, Flehmig, & Huestegge, 2016).

Design and Procedure

The design was two-factorial and contained the within-subject factors session (two levels: test vs. retest) and time-on-task (14 levels: Trials 1–14). The experiment took place in a noise-shielded room. The d2 sustained-attention test was administered twice within a test–retest interval of 1 week. Altogether, the task lasted about 4 min and 40 s per session.

Distribution Analysis

To analyze the entire distribution of performance (of the 1–14 trials), we computed the vincentized interpolated CDF of performance with 10 percentiles for each of the experimental condition (factor session, levels: test vs. retesting), according to Ulrich et al. (2007), and according to previous use of this method (Flehmig, Steinborn, Westhoff, & Langner, 2010; Steinborn & Huestegge, 2016; Steinborn, Langner, Flehmig, et al., 2016). By means of this analysis we were to know whether the observed effects of retesting on average performance speed were because of a global speed-up, or alternatively, because of a selective speed-up of the slow percentiles of the CDF. We expected that the coefficient of variation is sensitive to index such changes, since previous research has shown that this index closely corresponds to the skewness of an empirical response-time distribution (Flehmig et al., 2010; Steinborn, Langner, Flehmig, et al., 2016).

Results

Data Treatment

Although the test handbook provides the recommendation to discard the first and the last of the 14 trials, arguing that the first is biased by a warm-up effect while the last is biased by an end-spurt (final sprint) effect (Brickenkamp, 1962, 1966), we decided not to simply follow this recommendation but to examine this claim experimentally in the present study. Thus, the complete set of 1–14 trials were analyzed with respect to our research questions.

Descriptive Analysis

A formal description of potential performance measures to assess speed, accuracy, and variability is provided in Table 1. Population parameters (*M*, *SD*, skewness, and kurtosis) of standard and alternative performance scores are displayed in Table 2. As expected, the speed measures (Speed, $Speed_C$, including Max and Min) yielded a symmetrical distribution across the sample, moderate performance gains because of retesting, and test–retest stability of the relation between *M* and *SD*. At a descriptive level, therefore, performance speed measures are distinguished by favorable characteristics. Unfortunately, the same cannot be said of all other (performance accuracy and variability) measures, indicating (at a descriptive level) that these performance measures are problematic for reasons of either sample-distributional skewness or test–retest instability.

Correlation Analysis

Reliability coefficients of standard and alternative performance scores are displayed in Table 3. Because some readers might be interested in the intercorrelation between particular performance measures (e.g., Flehmig, Steinborn, Langner, Scholz, et al., 2007; Miles & Proctor, 2012; Schmiedek et al., 2009; van Ravenzwaaij, Brown, & Wagenmakers, 2011), this information is also shown, although the focus here is on test reliability. We would like to remind the reader that we aim to perform basic psychometric analysis of speeded test performance, which stands at the forefront of criteria for evaluating a test's psychometric quality, and which forms the basis for all subsequent research in clinical and neuropsychological subpopulations. This places high demands on sample quality (to ensure preconditions for accurate measurement) and sets particular high standards for evaluating reliability coefficients. Accordingly, reliability coefficients exceeding $r = .90$ are considered to indicate high reliability, coefficients exceeding $r = .80$ are considered to indicate sufficient reliability, while coefficients below these values must be considered insufficiently reliable (cf. Cronbach, 1975, pp. 190–217; Jensen, 2006, pp. 55–74). It is important to note that the fairly rigorous classification standards for reliability that we have to meet here are neither a universal entity nor a generally accepted specification (Lord & Novick, 1968; Miller & Ulrich, 2013; Rasch, 1980). Reliability standards for speeded tests are usually lower in clinical and neuropsychological research contexts, and are also lower for nonspeeded tests and self-report questionnaires assessing related constructs (cf. Bridger, Johnsen, & Brasher, 2013; Broadbent, Cooper, Fitzgerald, & Parkes, 1982; Cheyne, Carriere, & Smilek, 2006; Flehmig, Steinborn, Lang-

Table 2
Population Parameters of Performance Scores in the d2 Sustained-Attention Test

Score	Session 1				Session 2			
	M	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis
1. Speed	36.05	5.39	−.12	−.70	39.46	4.95	−.37	−.54
2. $Speed_C$	34.90	5.25	−.04	−.52	38.77	4.91	−.30	−.48
3. Error rate (%)	2.44	2.55	2.85	11.39	1.48	1.79	3.12	13.14
3. Error $rate_{trans}$	1.44	.68	1.15	2.21	1.06	.59	1.22	2.41
4. Error (omission)	15.4	16.62	2.96	12.22	9.56	11.72	3.18	13.51
4. $Omission_{trans}$	3.51	1.75	1.14	2.34	2.70	1.53	1.24	2.50
5. Error (commission)	.64	1.15	2.82	10.60	.19	.55	4.31	24.57
5. $Commission_{trans}$	.46	.66	1.10	.19	.16	.41	2.37	4.90
6. Variability (*SD*)	3.36	1.07	.41	1.80	2.77	1.15	−.41	.35
7. Variability (CV)	9.52	3.14	.18	.53	7.33	3.37	−.11	.22
8. Max	41.48	5.14	−.62	−.59	43.97	3.88	−1.25	.63
9. Min	30.07	5.62	.28	.05	34.55	6.20	.15	−.38
10. Range (%)	32.55	12.57	.56	.80	25.02	12.64	.11	.15

Note. $N = 100$; statistical measures: M = mean of sample population; SD = standard deviation of sample population; performance scores: Speed = average number of cancelled items per row; $Speed_C$ = error-corrected measure of speed defined as the average number of correctly cancelled items per row; error scores are reported in both original and square-root transformed metric; variability (SD) = reaction time standard deviation of performance speed; variability (CV) = coefficient of variation of performance speed; Min = row with the worst performance (number of cancelled items); Max = row with the best performance (number of cancelled items); range = performance fluctuation defined as relative difference (in %) between Max and Min.

ner, & Westhoff, 2007; Herrmann, 1982; Wilhelm, Witthöft, & Schipolowski, 2010).

Particularly noteworthy is the exceptional test–retest reliability of performance-speed measures ($r = .93^{**}$). This is true for both standard performance speed and error-corrected speed, which are also highly correlated, indicating that both are measuring the same construct and are thus interchangeable. Also, the Max and Min speed measures are relatively reliable ($r = .79^{**}$), which is surprising if one considers that these measures consist of only a small proportion of trials. Error rate ($r = .80^{**}$, $r = .75^{**}$ after square-root transformation) is less reliable than measures of speed, although the observed reliability coefficient is much higher than expected from previous studies (e.g., Brickenkamp, 1962; Hagemeister, 1994). Crucially, the reliability of the overall score of error-rate is solely driven by the omission error ($r = .80^{**}$, $r = .75^{**}$ after square-root transformation) while the commission error is entirely unreliable ($r = .05$, $r = .28^{**}$ after square-root transformation). This at least indicates that error scores should be treated with care. Finally, indices of response-speed variability (*SD*, CV, range) were also insufficiently reliable. In summary, this means that performance speed is clearly the best measure to be used to assess the ability to sustain attention/concentration in practical assessment contexts using the d2 test of sustained-attention.

Experimental Effects on Performance Speed

Within-subject GLM analyses were performed to analyze effects of session (1 vs. 2) and time-on-task (bin 1–14) on performance speed (see Table 4). The results are displayed in Figure 1. Participants became faster with retesting, as indicated by the main effect of session on performance speed ($F(1, 99) = 293.0$, $p < .001$, $\eta^2 = .75$). Further, they slowed down over the testing period, as indicated by a significant main effect of time-on-task on performance speed ($F(13, 1287) = 41.4$, $p < .001$, $\eta^2 = .30$). There was no significant Session × Time-on-Task interaction effect on

Table 3
Reliability Coefficients and Intercorrelations Between Performance Measures in the d2 Test

Session 2	Session 1									
	1	2	3	4	5	6	7	8	9	10
1. Speed	.93**	.98**	.22*	.22*	.07	.00	−.42**	.91**	.88**	−.42**
2. $Speed_C$	.99**	.93**	.00	.00	.03	−.05	−.44**	.88**	.88**	−.44**
3. Error (%)	.13	−.04	.79**	.99**	.17	.21*	.08	.24*	.10	.03
4. Error (omission)	.13	−.04	−.99**	.80**	.11	.21*	.08	.23*	.10	.03
5. Error (comission)	−.03	−.04	.08	.03	.05	.10	.04	.10	.04	.01
6. Variability (*SD*)	−.53	−.55**	.08	.07	.22*	.30**	.90**	.31**	−.41**	.82**
7. Variability (CV)	−.72	−.73**	.01	.01	.18	.96**	.48**	−.09	−.73**	.94**
8. Max	.86**	.84**	.13	.13	.01	−.12	−.34**	.79**	.71**	−.06
9. Min	.93**	.93**	.05	−.05	−.12	−.74**	−.87**	−73**	.79**	−.74
10. Range	−.73**	−.74**	−.01	−.01	.15	.92**	.97**	−.34**	−.88**	.39**

Note. $Speed_C$ = error-corrected measure of speed defined as the average number of correctly cancelled items per row; variability (CV) = coefficient of variation of performance speed; variability (SD) = reaction time standard deviation of performance speed. Test–retest reliability is shown in the main diagonal (denoted grey); correlation coefficients for the first testing session are shown above and for the second testing session below the main diagonal. Significant correlations are denoted ($N = 100$; * for $r \geq .21$, $p < .05$; ** for $r \geq .34$, $p < .01$).

Table 4
Results of the Experimental Effects of Session and Time-on-Task (TOT) on Performance

Source		Speed			Omission			Commission		
	df	*F*	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
1. Session	1,99	292.7	.001	.75	26.3	.001	.21	5.6	.020	.05
2. TOT	1,99	41.4	.001	.30	5.1	.001	.05	1.6	.075	.02
3. Session × TOT	1,99	1.4	.246	.01	1.2	.290	.01	1.0	.477	.01

Note. $N = 100$; effect size: partial η^2; experimental factors: session (1 vs. 2) and time-on-task (TOT: trials 1–14). Performance speed = number of items worked through per trial (statistics for the error-corrected speed measure is identical and therefore not shown); omission error = number of passive errors (targets missed) per trial; commission error = number of active errors (nontargets cancelled) per trial.

performance speed ($F < 1.5$). These results indicate that the d2 test is subject to change as indicated by a clear-cut performance decrement within a session and a performance benefit across sessions.

Experimental Effects on Omission Errors

A within-subject GLM analysis was performed to analyze effects of session (1 vs. 2) and time-on-task (bin 1–14) on performance accuracy (see Table 4). The results are visually displayed in Figure 2. Participants became less erroneous with retesting, as indicated by the main effect of session on omission errors ($F(1, 99) = 26.3$, $p < .001$, $\eta^2 = .21$). Further, there was a significant main effect of time-on-task on omission errors ($F(13, 1287) = 5.1$, $p < .001$, $\eta^2 = .05$), indicating change over the testing period (although error rate was generally low, as displayed in Figure 1, and the effect size was notably small). There was no significant Session × Time-on-Task interaction effect. It should be mentioned that similar results were obtained with the square-root transformed error score, and thereby is not redundantly reported in more detail.

Experimental (GLM) Effects on Commission Errors

Within-subject GLM analyses were performed to analyze effects of session (1 vs. 2) and time-on-task (bin 1–14) on commission errors (cf. Table 4 and Figure 1). Participants became less erroneous with retesting, as indicated by the main effect of session on commission errors ($F(1, 99) = 5.6$, $p < .05$, $\eta^2 = .05$). There was no significant main effect of time-on-task on commission errors, and also no Session × Time-on-Task interaction effect on commission errors. Again, the results were similar when the square-root transformed error score was used, and therefore, is also not redundantly reported in more detail.

Experimental Effects on Performance Variability

Within-subject GLM analyses were performed to examine effects of the factor session (session 1 vs. 2) on aspects of performance variability. An overview of statistical effects for all performance measures is given in Table 5. Participants' performance became less variable as indicated by a main effect of session on performance variability, as indexed by several alternative scoring methods: the performance range, the performance *SD*, the performance coefficient of variation (CV).

Warm-Up Effects

We performed supplementary GLM analysis that included the within-subject factors session (1 vs. 2) and time-on-task (Trials 1–14) on performance. Remind that it is recommended to discard the first and the last of the 14 trials to increase psychometric quality (Brickenkamp, 1962, 1966). To test the hypothesis that the first trial in the sequence (of 14 trials overall) is demonstrably slow and error prone, we implemented a preplanned (Helmert contrast) single comparison analysis. Remind that Helmert contrasts systematically compare the mean of each factor level with the mean of the succeeding factor levels. In our case, a comparison of the first trial with the mean of its succeeding trials is crucial. This analysis revealed a significant main effect of the factor time-on-task on performance speed ($F(1, 99) = 76.2$, $p < .001$; partial $\eta^2 = .44$). However, the effect was in the opposite direction indicating a decrease not an increase in performance speed. This effect was not different for the first and the second testing sessions ($F < 1$, partial $\eta^2 = .03$). There was also a significant main effect of time-on-task on omission errors ($F(1, 99) = 15.0$, $p < .001$; partial $\eta^2 = .13$), indicating a decrease in omission errors. This effect was indeed larger for the first than for the second testing session, as indicated by an interaction effect of session and time-on-task on omission errors ($F(1, 99) = 6.4$, $p < .05$; partial $\eta^2 = .06$). Regarding commission errors, neither a main effect nor an interaction effect was observed ($F < 0.2$, partial $\eta^2 < .00$). Finally, correlation analysis revealed that the exclusion of the first and/or the last trial did not improve test reliability. Quite the contrary, the inclusion of these trials adds to test reliability in a very normal way through test lengthening.

CDF Analysis

Figure 3 depicts the effect of retesting on the CDF of performance. Visual inspection of Figure 3 (Panel A) indicates that besides the global effect on all CDF percentiles (i.e., the entire distribution shifted rightward), there was a tendency of a selective effect on the slower (vs. faster) CDF percentiles. Statistical relationships were tested by comparing averages across the three lower (vs. upper) CDF percentiles in compound, which leads to a 2 × 2 GLM model (session: test vs. retest; CDF: lower vs. upper percentile). Crucial is the interaction effect on performance ($F(1, 99) = 21.8$, $p < .001$; partial $\eta^2 = .18$), indicating differentially larger retesting benefits for lower (vs. upper) percentiles. This interaction effect was completely abolished for the error-corrected measure of speed, both statistically ($F < 1.2$, partial $\eta^2 = .00$) and visually (Figure 3, Panel B). Figure 4 displays a of the retesting effect, both for standard and the error-corrected performance speed. A is obtained by calculating the difference in mean performance speed as in-

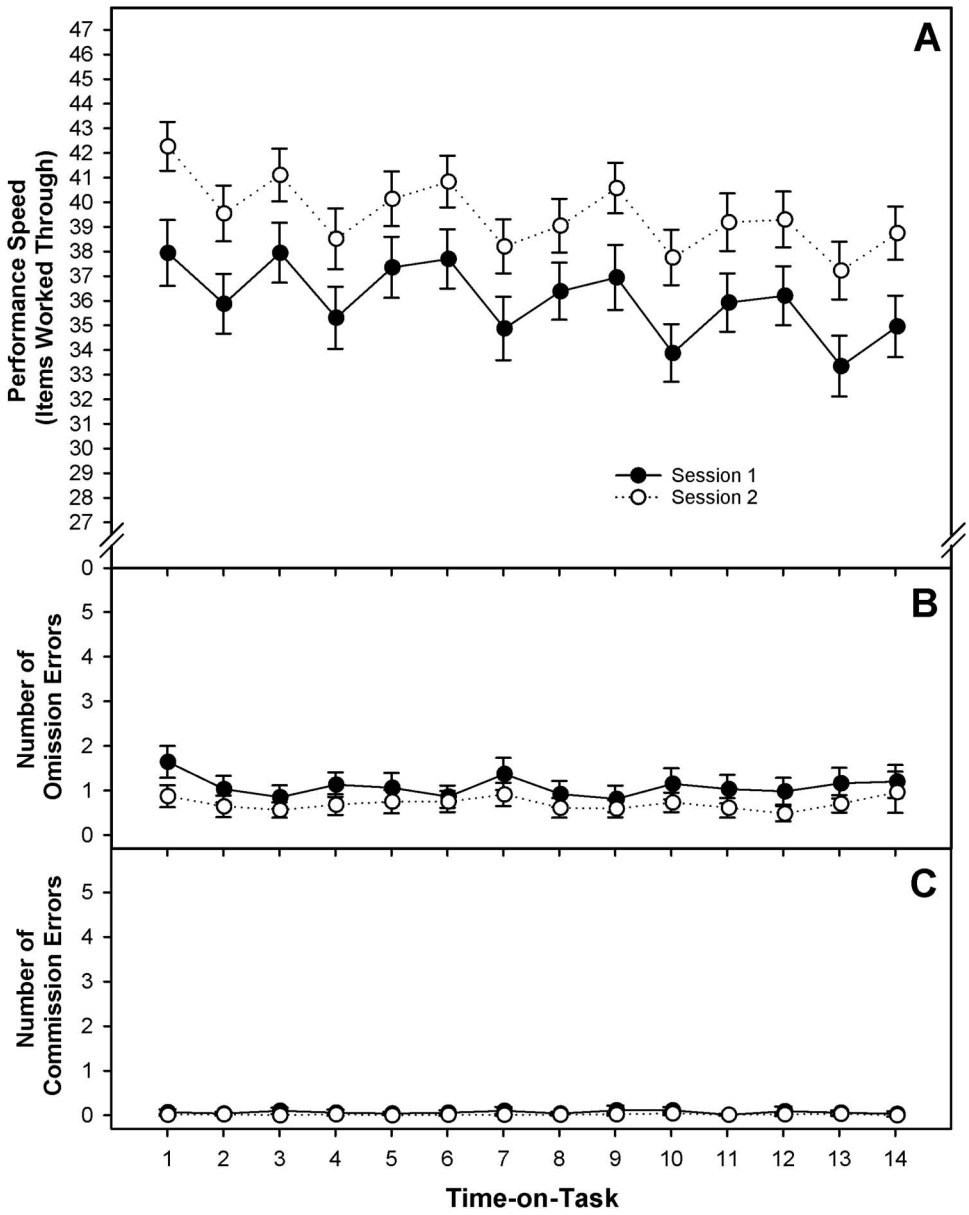

Figure 1. Error-corrected performance speed (number of items worked through minus number of errors committed) as a function of the factors "session" (first vs. second testing) and "time on task" (rows 1–14) in the d2 sustained-attention test. The error bars show the measurement uncertainty of the mean score for a confidence interval range of 95% (Appendix Table A1 and A2).

duced by an experimental manipulation against the mean of the experimental condition for each of the percentiles. By means of this analysis, the effects of retesting can be evaluated relative to the mean of level of performance (cf. De Jong, Liang, & Lauber, 1994; Ridderinkhof, 2002; Steinborn, Langner, & Huestegge, 2016). Note that provide a convenient simplification of the relatively complex information present in the CDFs, and thus serve to improve understanding, though the relationship can also statistically be tested, yielding the same results as argued above (cf. Schwarz & Miller, 2012; Steinborn & Huestegge, 2016).

CRF Analysis

The CRF analysis was performed for both the measures of performance speed and error-corrected performance speed (Figure 5, Panel A and B). By means of a generalized split-half reliability model, we analyzed the within-session reliability as a function of test length, separately for the first testing session and the retesting session. By means of a test–retest reliability model, we analyzed the between-session reliability as a function of test length (cf. Miller & Ulrich, 2013, for details regarding theory and methodology). As becomes evident from

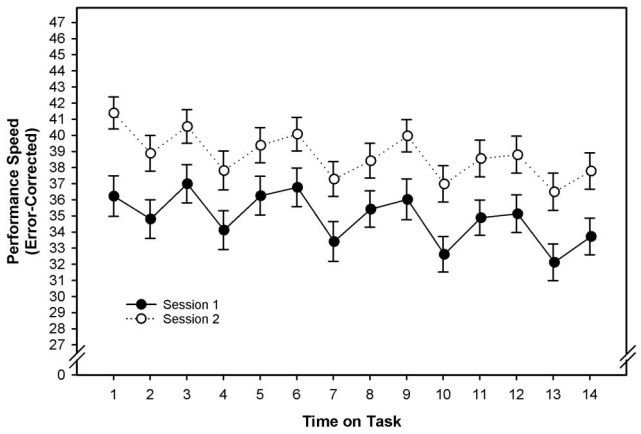

Figure 2. Performance speed (number of items worked through, Panel A) and accuracy (number of omission and commission errors, Panel B and C) as a function of the factors "session" (first vs. second testing) and "time on task" (rows 1–14) in the d2 sustained-attention test. The error bars show the measurement uncertainty of the mean score for a confidence interval range of 95%.

Figure 5, the same results were obtained for both measures. Reliability globally increased with increasing test length. The increase occurred in a monotonous fashion until an asymptotic level is reached. The split-half reliability coefficient revealed a high degree of internal stability, which is preserved with short test length, both at the first and at the second testing session. The test–retest reliability coefficient also revealed good reliability, although this measure is more strongly affected by test length. Notably, excellent test–retest reliability ($r > .90$) is retained even with 50% of the test length. In sum, the results revealed excellent psychometric characteristics of the d2 test's performance primary measures which also prevail under an economic vantage point, similarly for the measures of standard and error-corrected performance speed.

Discussion

What makes a performance speed test psychometrically reliable? According to classical test theory, the reliability of a test is determined by the amount of true score variance relative to error variance, and is known to increase with the number of trials. Reliability places an upper limit on the correlations that can be observed with other relevant measures, usually be taken to examine aspects of construct and criterion validity. Unfortunately, researchers involved in psychometric-test construction issues arguably derive little benefit from such a purely statistical perspective, except for the advice to lengthen the testing time of their instruments to increase reliability. According to Cronbach (1975), a psychologically substantiated view of reliability must offer the possibility to theorize on particular cognitive processes that might either tap into true-score variance or error variance, respectively (cf. Ackerman, 1987; Bornstein, 2011; Kiesel et al., 2010). Cronbach (1975) conjectured that in psychometric testing, true score variance might increase with motivation on part of individuals being tested, and might also increase with item characteristics as far as they are capable to affect motivation. Further it should not be affected by extant factors such as, for example, the way in which the pen is held and used (residual component). Error variance, on the other hand, should increase with any variable that adds coincidental effects to the performance score, originating either from person-specific (e.g., mindwandering) or situation-specific (e.g., disturbances in the test procedure) sources.

Stability of Test Performance With Repeated Testing

According to Miller and Ulrich (2013), crucial to the understanding of what determines test reliability is to know how situational preconditions reveal (ability-related) true-score variance, how demand characteristics evoke (task-related) true-score variance, to determine factors that tap into residual variance, and to reduce the influence of any variable that produces coincidental effects and thus increases error variance. Combining the experimental and the correlational approach to psychometric testing (Pieters, 1983, 1985), we started out by examining experimental effects of repeated testing and time on task on performance. The results revealed sensitivity to change in two different ways, as we found an effect of both time on task (a decrease within a session) and retesting (an increase between sessions) on performance. The gains because of retesting on

Table 5
Retesting Effects of Different Performance Measures in the d2 Sustained-Attention Test

Source	F	p	η_p^2	Cohen d_z	CI (d_z)	Cohen d	CI (d)
1. Speed	292.7	.001	.75	1.71	[1.72–1.74]	.66	[.26–1.06]
2. $Speed_C$	413.1	.001	.81	2.03	[1.99–2.12]	.76	[.36–1.17]
3. Error rate	37.2	.001	.27	.61	[.35–1.09]	.44	[.04–.83]
4. Error (omission)	32.9	.001	.25	.57	[.31–1.05]	.41	[.01–.80]
5. Error (comission)	19.9	.001	.17	.45	[.20–.87]	.50	[.01–.90]
6. Variability (*SD*)	20.0	.001	.16	.45	[.23–.79]	.53	[.13–.93]
7. Variability (CV)	43.1	.001	.30	.66	[.41–1.00]	.67	[.27–1.08]
8. Max	60.0	.001	.39	.78	[.69–.89]	.55	[.15–.95]
9. Min	138.3	.001	.58	1.18	[1.15–1.23]	.76	[.35–1.16]
10. Range (%)	28.8	.001	.23	.54	[.31–.89]	.60	[.20–1.00]

Note. $N = 100$, df (1,99); effect size: partial η^2, Cohen d_z for difference measures (including confidence intervals), and Cohen d (including confidence intervals). CI = 95% confidence interval. $Speed_C$ = error-corrected measure of speed defined as the average number of correctly cancelled items per row; variability (CV) = coefficient of variation of performance speed; variability (SD) = reaction time standard deviation of performance speed. A description of the computation for each of the presented performance measures is provided in Table 1.

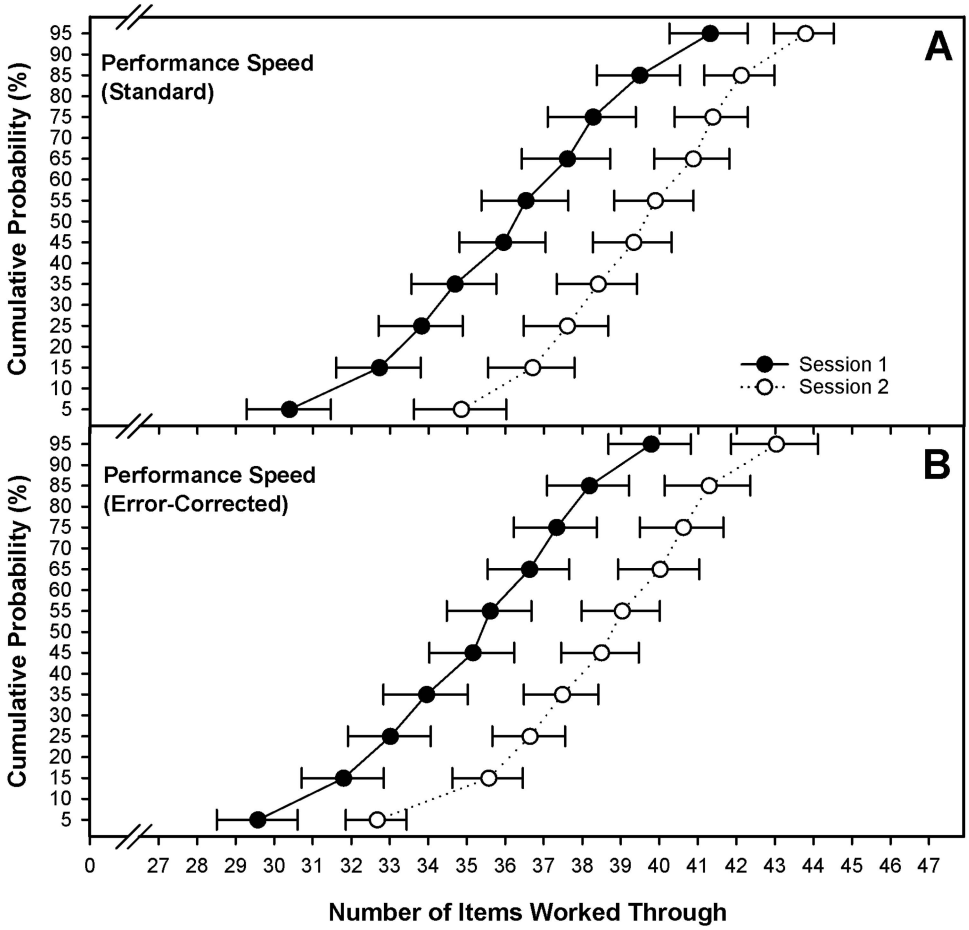

Figure 3. Vincentized interpolated cumulative distributive functions (CDFs) of performance speed (number of items worked through) as a function of the factor "session" (first vs. second testing) in the d2 sustained-attention test, separately displayed for standard (Panel A) and error-corrected performance speed (Panel B). The error bars show the measurement uncertainty of the mean score for a confidence interval range of 95% (Appendix Table A3).

performance speed were within the range of expectations (partial $\eta^2 = .75$; Cohen's $d_z = 1.71$) and seemed not to have hampered measurement accuracy, because the observed retest-reliability coefficient was exceptional ($r = .93$) for a test of such short duration (Cronbach, 1975, pp. 190–217). All these effects were exactly similar for measures of standard performance speed (as recommended by the handbook of the d2 test) and error-corrected performance speed as an alternative measure (Tables 2, 3, and 5). This means that both indices are indistinguishable with respect to test stability and reliability, rendering any debate about preferred use of one or the other measure unnecessary for the use in standard assessment situations (but see Footnote 3, for its use in dissimulation research). Also, the previously hypothesized warm-up effect was not found, since individuals were faster (yet somewhat more error-prone) in the first trial than in the second (Figures 1 and 2), rendering this issue of only minor importance.

We further examined effects of repeated testing on performance variability by considering commonly accepted indices of performance variability (*SD*, CV, range, etc.) as well as the CDF of trials (cf. De Jong et al., 1994; Schweickert, Giorgini, & Dzhafarov, 2000; Ulrich et al., 2007). By means of distributional analysis, we were able to more precisely characterize the changes that might occur with retesting. Essentially, we examined whether the observed effects on average performance speed originate from a global speed-up of all CDF percentiles or from a selective speed-up of only the slowest CDF percentiles, which would indicate a stabilization of performance (Langner & Eickhoff, 2013; Steinborn & Huestegge, 2016; Steinborn, Langner, & Huestegge, 2016). Although retesting essentially affected all CDF percentiles to some degree, the effect was more pronounced in the slower than in the faster percentiles of the CDF. This becomes visually evident in Figure 3 (Panel A). Yet, the selective effect was absent in the error-corrected CDF, a pattern that is typically found in self-paced tests (Bertelson & Joffe, 1963; Bills, 1931, 1935; Sanders & Hoogenboom, 1970; Steinborn & Huestegge, 2016). It indicates that the error-penalty effect is more pronounced in the slower (vs. faster) CDF percentiles (Figures 4; Appendix Table A3), even though

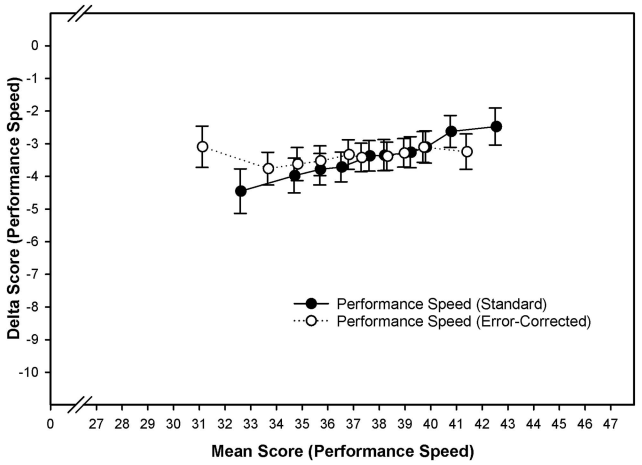

Figure 4. Delta plots of the test–retest effect, separately displayed for the standard (Panel A) and error-corrected performance speed (Panel B). For each percentile, the performance speed difference between the experimental conditions (test vs. retest) is plotted against the mean of the conditions in that percentile. The error bars show the measurement uncertainty of the mean score for a confidence interval range of 95%.

it did not affect reliability (see Figure 5).[1] One should always bear in mind, however, that these findings hold only for the case of repeated testing. Extensive training to the asymptote of performance might produce an entirely different pattern, as very extensive practice causes a transition from controlled information processing to pure memory retrieval (Logan, 1988, 1992; Rickard, Lau, & Pashler, 2008).

Given that extensive practice also changes the underlying construct that is intended to be measured (e.g., Ackerman, 1987), we also evaluated a list of commonly used alternative performance measures (that mostly tap into aspects of performance variability) with respect to test–retest effects (cf. Table 5). Remind that most of these indices can already be refuted on grounds of insufficient test reliability (see Table 3 and Appendix Table A4), but for the sake of completeness, it still might be useful to further characterize them with respect to practice effects. The robustness of performance indices because of repeated testing and practice is an important requirement concerning test validity, especially in applied contexts in which the amount of prior test experience typically cannot be established. With regard to performance speed, our results are about similar to and in line with previous observations both under normal (Flehmig, Steinborn, Langner, Scholz, et al., 2007; Hagemeister, 2007; Steinborn et al., 2008; Steinborn, Flehmig, et al., 2009) and detrimental testing conditions produced by time on task (Langner, Eickhoff, & Steinborn, 2011; Langner, Steinborn, Chatterjee, Sturm, & Willmes, 2010) and prolonged wakefulness (Bratzke, Rolke, Steinborn, & Ulrich, 2009; Bratzke, Steinborn, Rolke, & Ulrich, 2012; Steinborn, Bratzke, et al., 2010). With regard to types of errors, gains were observed for both omission and commission errors, although one should always consider the problems with leveraging effects of low-probability events in psychometric testing of speeded performance.[2]

Analysis of Test Reliability

Reliability is the top-priority issue in test-evaluation research and its accurate determination places high demands on sample quality and sets particularly high standards for evaluating reliability coefficients (cf. Cronbach, 1975, pp. 190–217; Jensen, 2006, pp. 55–74). A rigorous classification scheme (i.e., $r >$ .90 = high, $r >$ .80 = sufficient, $r <$ .80 = problematic) is therefore needed. Our results revealed excellent reliability coefficients ($r >$ .90) for average performance speed while error rate was reliable only under some constraints (see Table 3 and Appendix Table A4). While omission errors were still sufficiently reliable (r = .80), commission errors were absolutely unreliable (r = .05). And yet, despite the rigorous classification scheme, the reliability of the omission-error score is still unusually high as compared with related findings (Brickenkamp, 1962, 1966; Hagemeister, 1994; Westhoff, 1985; Westhoff & Lemme, 1988). Reliability problems related to error scores are a well-known psychometric test-construction issue albeit woefully neglected in applied-research contexts, resulting in an increased risk of error scores being inadequately used and interpreted, as is the case in many applied studies (e.g., Gropel et al., 2014; Wetter, Wegge, Jonas, & Schmidt, 2012). It should be noted once again that the lower reliability of error scores naturally arises from its low occurrence probability, to be overcome by lengthening a test (Miller & Ulrich, 2013; Stuss et al., 1996; Ulrich & Miller, 1994). Hagemeister (1994) has conducted a large number of studies on this subject, wondering as to whether reliability might be increased to a tolerable degree by lengthening the test. In one of her studies, participants were administered four times with the d2 test, which yielded a decrease in population-parameter skewness and consequently, an increase of test reliability.

[1] One reviewer wondered about the nature of errors in fast (vs. slow) trials and about impulsivity (vs. overload) as potential underlying mechanisms. We would like to give a short summary of findings from RT-based (mental-chronometry) research: Whether errors in computerized serial choice-RT experiments occur at fast or slow percentiles critically depends on the response–stimulus interval. Given the RSI is constant, then the following rule applies: First, with short RSIs (0–50 ms), or in self-paced tasks (as is the d2 test), errors usually occur at the slower CDF percentiles (thus, referred to as overload errors), whereas with somewhat longer RSIs (500–750 ms), errors occur at the faster percentiles (thus, referred to as impulsive errors). Second, overall error rate also increases toward longer and thus more optimal (500–750 ms) RSIs but decreases again with further lengthening of this interval. These issues are mostly addressed in the literature on time-preparatory (foreperiod) effects and further reading can be found there (Niemi & Näätänen, 1981; Ollman & Billington, 1972; Posner, Klein, Summers, & Buggie, 1973; Steinborn & Langner, 2011, 2012; Steinborn, Langner, & Huestegge, 2016).

[2] Response to reviewer comment: The trouble with error scores (and any other low-probability events) in self-paced speed tests is that they cannot accurately be predicted, because they entail a skewed sample distribution. This directly results in low test–retest correlations. The reason for this is that errors in speeded tests are (both theoretically conceptualized and empirically found as) rare events and that substantial difficulties arise connected with the unpredictability of low-probability events (Jentzsch & Dudschig, 2009; Notebaert et al., 2009; Steinborn, Flehmig, Bratzke, & Schröter, 2012). Transformation procedures to reduce skew might be helpful (cf. Dunlap, Chen, & Greer, 1994), but one should always bear in mind that the lack of reliability of error scores arises from their nature as being rare events; thus, low predictability of error scores is a theoretical prediction in self-paced speeded tests.

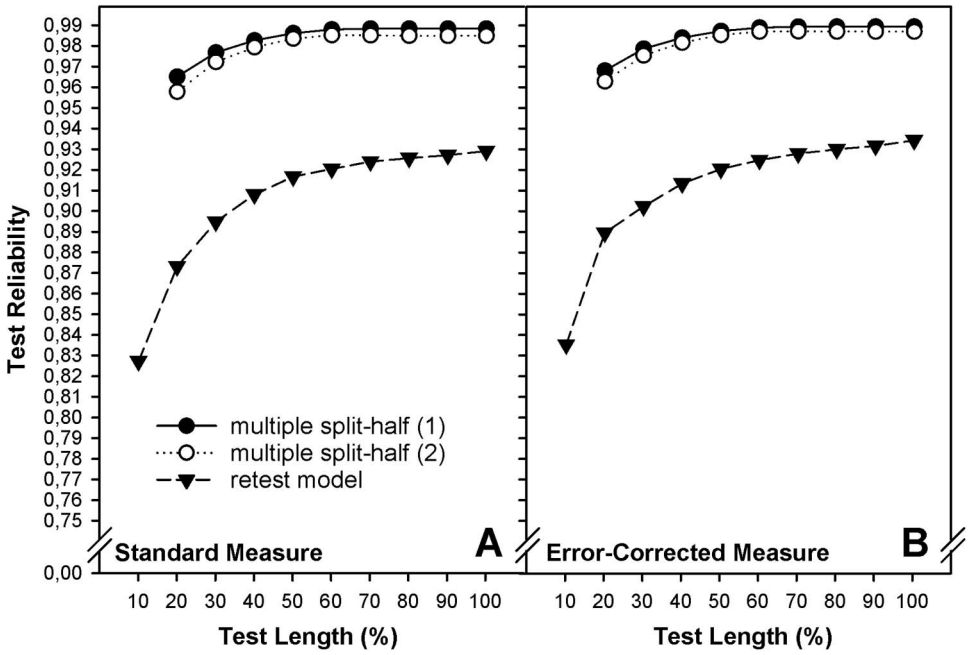

Figure 5. Reliability of performance speed in the d2 test for the first and the second test administration (multiple split-half reliability model) and for repeated test administration (retest reliability model). Data are separately displayed for standard (Panel A) and error-corrected measures (Panel B).

As a further novel aspect, at least not found very often in empirical research (Miller & Ulrich, 2013), we performed CRF analysis to evaluate reliability as a function of test length (Lord & Novick, 1968; Miller & Ulrich, 2013). As becomes evident from Figure 5, the multiple split-half coefficient yielded exceptional values, indicating that internal stability of performance is preserved across the test from the first to the last trial. Further, the test–retest reliability coefficient increased empirically with the length, exhibiting relatively low reliability ($r = .83$) when only 10% of the test are considered but extraordinarily high reliability ($r = .93$) when 100% of test length is considered. The same results were notably obtained for both standard and error-corrected performance measures. Crucially, it becomes also evident from Figure 5 that sufficiently high reliability is already obtained with 50% of the original test length. This might be important news for researchers that usually use the d2 test for purposes such as to establishing a workload (interference) condition in memory-recall experiments (e.g., Engelkamp, Jahn, & Seiler, 2003; Golly-Häring & Engelkamp, 2003; Jahn & Engelkamp, 2003), or for standard purposes such as population screening or to globally determine participants' cognitive ability (Strobach, Frensch, Muller, & Schubert, 2015; Strobach, Frensch, & Schubert, 2012). Because the d2 test is one of the most widely used instruments to assessing elementary cognitive ability (Krumm et al., 2009, 2008), it might be particularly appropriate whenever economic considerations are important.

The findings presented here might be important for practitioners who use available psychometric tests such as the d2 test, because of its convenience of providing recommendations for scoring and interpreting performance. While indices of performance speed are perfectly reliable (that is retained with even half of the test length), the situation is less promising regarding performance accuracy (though omission-error reliability was unexpectedly high). Given the common opinion to treat indices of performance speed and accuracy as equally important yet distinct aspects of cognitive ability (e.g., Becker, Roth, Andrich, & Margraf, 1999; Mayer et al., 2003; Mussweiler & Strack, 2000), our results provide important implications regarding their proper use and interpretation. We argue that one should always keep a close eye on matters of test reliability connected with error scores when interpreting correlative relationships to respective validity criteria (cf. Hughes, Linck, Bowles, Koeth, & Bunting, 2014). We further regard it advisable to consider errors in the d2 test a heterogeneous (not a homogeneous) construct with different underlying sources. In some sense, this holds for any Bourdon-like psychometric test, that is, this suggestion applies to any cancellation test. Finally, it is important to note that none of the widely used variability measures reached sufficient test reliability. It holds for the case of the d2 test of sustained-attention that these measures should not be taken at all.

Practical Recommendations for Scoring Performance

How should performance measures of the d2 test of sustained-attention be used by practitioners and researchers in applied research fields? Given the widespread reputation of the d2 test in Europe (cf. Schorr, 1995), this question is of particular importance. Handbook guidelines and common practice suggest to exclude the first and last trials to improve test reliability (Brickenkamp, 1962, 1966; Hagemeister, 1994; Westhoff, 1985; Westhoff & Lemme, 1988). Further it consid-

ers the work done per time unit to index information-processing speed, the rate of errors committed to index diligence, and the performance range to index persistence (i.e., the ability to keep the system focused for a certain period of time). Moreover, a variety of alternative measures are also in widespread use in both clinical/neuropsychological and social-psychological research (Bates & Lemay, 2004). Despite a widespread use, however, a thorough investigation of basic psychometric properties has *not* yet taken place, whereby it remains uncertain whether these measures might be used at all. Our study is aimed to serve as a guideline for the selection of a specific performance index that is of particular interest. Contrary to common opinion, our results imply that excluding the first and the last trial from the data does not improve (but decrease) test reliability, whereby we recommend inclusion of all trials for performance analysis. Further, while we certify excellent psychometric quality for performance speed, we cannot thoroughly recommend the use of error rate (%) to index diligence. Instead, we suggest to treat error rate a tradeoff variable which might be integrated into a compound measure of error-corrected speed.[3] Finally, variability measures should not be used at all.

For practitioners working in clinical and neuropsychological contexts, we recommend that performance speed (or error-corrected speed, respectively) is the only measure eligible for highly reliable assessment, and thus should be regarded as the primary performance index. Whether to use the standard score or the error-corrected score seems of minor importance, from the perspective of reliability, because both indices revealed similar psychometric quality and are therefore interchangeable as evidenced from the experimental and the correlational analyses. Errors might also be used as secondary (not primary) performance dimension. For researchers in ergonomic contexts, where transient states are induced via experimental conditions (Hancock, Ross, & Szalma, 2007; Szalma & Hancock, 2011), aspects of performance accuracy (i.e., errors of omission and commission) and variability might serve as secondary aspects besides performance speed, although caution is advised regarding interpretations. Given the wide use of the d2 test as a standard instrument to assess attentional efficiency (e.g., Kramer, Weger, & Sharma, 2013; Moore & Malinowski, 2009), to globally characterize participants in a particular research context such as cognitive aging (e.g., Getzmann, Wascher, & Falkenstein, 2015; Strobach et al., 2015), as a distractor condition in memory recall experiments (e.g., Engelkamp et al., 2003; Golly-Häring & Engelkamp, 2003; Jahn & Engelkamp, 2003), or to indicate self-control depletion (e.g., Friese, Messner, & Schaffner, 2012; Gropel et al., 2014), these recommendations should not be taken lightly in future studies.

While we provide guidance for the d2 test, this does not mean we consider scores of performance accuracy or variability as generally unreliable. For mathematical reasons, measures of intraindividual variability can never reach the same degree of reliability as compared with measures of central tendency, but their reliability will certainly increase with increasing length of a test. Although Stuss et al. (1996) could not identify any advantage of long psychometric tests over short ones, as a result of a comparative study of tests for detecting cognitive decline in older adults, the actual advantage of long (over short) tests might consist in the potential to bring forth a measure of performance variability of sufficiently high reliability. Thus, an evaluation of short versus longer test-versions of the d2 test might be a promising field of subsequent research (cf. Hagemeister, 1994, 2007; Westhoff, 1985; Westhoff & Lemme, 1988), and might be useful in practical and clinical contexts whenever economic considerations are of only minor importance (Green, Chen, Helms, & Henze, 2011; Sturm, Willmes, Orgass, & Hartje, 1997; Stuss et al., 2001). As already mentioned, error-score reliability is expected to increase with increasing opportunity to commit errors (i.e., with test lengthening) and also to increase with time pressure (e.g., by natural deadlines Los, Hoorn, Grin, & Van der Burg, 2013; Ruthruff, Johnston, & Remington, 2009). Crucially however, one should bear in mind that the construct underlying error scores will also change accordingly.

Final Conclusion

The key contribution that this psychometric analysis delivers covers three aspects, (a) knowledge in terms of novel, theoretically important insights into the temporal dynamics of sustained-attention performance, (b) methodology of design and research logic within the framework of mental chronometry, (c) and advanced measurement technology to index test reliability. *First*, despite prior studies, our results provide new knowledge since we are the first to exactly determine the size of both within-session (time-on-task effects) and between-session (test–retest effects) temporal performance dynamics. *Second*, we provide a methodical advancement to psychometric evaluation of speeded test performance within the realm of mental chronometry. This includes the goals of manipulating critical experimental variables and measuring its effects by analyzing performance distributions instead of only analyzing mean performance, which is a major advancement to previous research in this domain. Third, we provide an advanced approach to evaluate a test's measurement accuracy by computing CRF. As argued earlier, test–retest correlations of overall test performance are not adequate to be used for comparison purposes (e.g., Habekost, Petersen, & Vangkilde, 2014), if the to-be-compared tests differ in length (Miller & Ulrich, 2013). The

[3] One reviewer had questions regarding effects of instruction on omission and commission errors in the d2 test, which we would like to answer here: Standard instructions (to either emphasize speed or accuracy) usually affect omission errors (more misses with greater speed) but not commission errors. In fact, individuals do not tend to actively commit errors in the d2 test (cf. Table 2). At the first sight, that may well seem surprising but it actually lies in the nature of the Bourdon-test (scan, check, and cancellation) principle. A strategy to emphasize speed (at the cost of accuracy) implies faster scanning, cursory checking, and less frequent responding (so costs incurred by additional, false-positive, and responses are avoided). A strategy to emphasize accuracy, on the other hand, implies slower scanning, and intense checking (so the increased discrimination time available counteracts false-positive responding). Notably, the relation of the error types are considered an indication of dissimulation. It is regarded useful in applied clinical contexts, where individuals could be expected either to feign (at least exaggerate) symptoms to receive therapeutic intervention, or because rigidified expectations turned into a self-fulfilling prophecy (Bates & Lemay, 2004; Brickenkamp, 1962; Hagemeister, 1994; Suhr & Wei, 2013; Wei & Suhr, 2015, for further reading). When student participants are instructed (being a patient in a role playing game) to dissimulate poor performance, then they usually exhibit unusually high numbers of commission (not omission) errors.

central message of our report to be delivered to the community therefore is that test–retest correlations are not interpretable by itself but must be evaluated as a function of test length.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information-processing perspectives. *Psychological Bulletin, 102,* 3–27. http://dx.doi.org/10.1037/0033-2909.102.1.3

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15,* 163–181. http://dx.doi.org/10.1037/a0015719

Bates, M. E., & Lemay, E. P., Jr. (2004). The d2 Test of attention: Construct validity and extensions in scoring techniques. *Journal of the International Neuropsychological Society, 10,* 392–400. http://dx.doi.org/10.1017/S135561770410307X

Becker, E. S., Roth, W. T., Andrich, M., & Margraf, J. (1999). Explicit memory in anxiety disorders. *Journal of Abnormal Psychology, 108,* 153–163. http://dx.doi.org/10.1037/0021-843X.108.1.153

Bertelson, P., & Joffe, R. (1963). Blockings in prolonged serial responding. *Ergonomics, 6,* 109–116. http://dx.doi.org/10.1080/00140136308930682

Bills, A. G. (1931). Blocking: A new principle of mental fatigue. *The American Journal of Psychology, 43,* 230–245. http://dx.doi.org/10.2307/1414771

Bills, A. G. (1935). Some causal factors in mental blocking. *Journal of Experimental Psychology, 18,* 172–185. http://dx.doi.org/10.1037/h0059285

Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment, 23,* 532–544. http://dx.doi.org/10.1037/a0022402

Bratzke, D., Rolke, B., Steinborn, M. B., & Ulrich, R. (2009). The effect of 40 h constant wakefulness on task-switching efficiency. *Journal of Sleep Research, 18,* 167–172. http://dx.doi.org/10.1111/j.1365-2869.2008.00729.x

Bratzke, D., Steinborn, M. B., Rolke, B., & Ulrich, R. (2012). Effects of sleep loss and circadian rhythm on executive inhibitory control in the Stroop and Simon tasks. *Chronobiology International, 29,* 55–61. http://dx.doi.org/10.3109/07420528.2011.635235

Brickenkamp, R. (1962). *Aufmerksamkeits-Belastungs-Test (Test d2)* (1st ed.). Göttingen, Germany: Hogrefe.

Brickenkamp, R. (1966). Anmerkungen zur Interpretierbarkeit von Leistungsschwankungen im Test d2. *Diagnostica, 19,* 125–131.

Bridger, R. S., Johnsen, S. A. K., & Brasher, K. (2013). Psychometric properties of the Cognitive Failures Questionnaire. *Ergonomics, 56,* 1515–1524. http://dx.doi.org/10.1080/00140139.2013.821172

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology, 21,* 1–16. http://dx.doi.org/10.1111/j.2044-8260.1982.tb01421.x

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cheyne, J. A., Carriere, J. S. A., & Smilek, D. (2006). Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition, 15,* 578–592. http://dx.doi.org/10.1016/j.concog.2005.11.009

Corcoran, D. W. (1966). An acoustic factor in letter cancellation. *Nature, 210,* 658. http://dx.doi.org/10.1038/210658a0

Cronbach, L. J. (1947). Test reliability; its meaning and determination. *Psychometrika, 12,* 1–16. http://dx.doi.org/10.1007/BF02289289

Cronbach, L. J. (1975). *Essentials of psychological testing*. New York, NY: Harper & Row.

De Jong, R., Liang, C. C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 731–750. http://dx.doi.org/10.1037/0096-1523.20.4.731

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193–222. http://dx.doi.org/10.1146/annurev.ne.18.030195.001205

Doebler, P., & Scheffler, B. (2016). The relationship of choice reaction time variability and intelligence: A meta-analysis. *Learning and Individual Differences, 52,* 157–166. http://dx.doi.org/10.1016/j.lindif.2015.02.009

Dunlap, W. P., Chen, R. S., & Greer, T. (1994). Skew reduces test-retest reliability. *Journal of Applied Psychology, 79,* 310–313. http://dx.doi.org/10.1037/0021-9010.79.2.310

Engelkamp, J., Jahn, P., & Seiler, K. H. (2003). The item-order hypothesis reconsidered: The role of order information in free recall. *Psychological Research, 67,* 280–290. http://dx.doi.org/10.1007/s00426-002-0118-1

Flehmig, H. C., Steinborn, M. B., Langner, R., Scholz, A., & Westhoff, K. (2007). Assessing intraindividual variability in sustained attention: Reliability, relation to speed and accuracy, and practice effects. *Psychology Science, 49,* 132–149.

Flehmig, H. C., Steinborn, M. B., Langner, R., & Westhoff, K. (2007). Neuroticism and the mental noise hypothesis: Relation to lapses of attention and slips of action in everyday life. *Psychological Science, 49,* 343–360.

Flehmig, H. C., Steinborn, M. B., Westhoff, K., & Langner, R. (2010). Neuroticism and speed-accuracy tradeoff in self-paced speeded mental addition and comparison. *Journal of Individual Differences, 31,* 130–137. http://dx.doi.org/10.1027/1614-0001/a000021

Friese, M., Messner, C., & Schaffner, Y. (2012). Mindfulness meditation counteracts self-control depletion. *Consciousness and Cognition, 21,* 1016–1022. http://dx.doi.org/10.1016/j.concog.2012.01.008

Frings, C., Rothermund, K., & Wentura, D. (2007). Distractor repetitions retrieve previous responses to targets. *The Quarterly Journal of Experimental Psychology, 60,* 1367–1377. http://dx.doi.org/10.1080/17470210600955645

Getzmann, S., Wascher, E., & Falkenstein, M. (2015). What does successful speech-in-noise perception in aging depend on? Electrophysiological correlates of high and low performance in older adults. *Neuropsychologia, 70,* 43–57. http://dx.doi.org/10.1016/j.neuropsychologia.2015.02.009

Golly-Häring, C., & Engelkamp, J. (2003). Categorical-relational and order-relational information in memory for subject-performed and experimenter-performed actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 965–975. http://dx.doi.org/10.1037/0278-7393.29.5.965

Graveson, J., Bauermeister, S., McKeown, D., & Bunce, D. (2016). Intra-individual reaction time variability, falls, and gait in old age: A systematic review. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 71,* 857–864. http://dx.doi.org/10.1093/geronb/gbv027

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in Psychological Assessment: Recognizing the people behind the data. *Psychological Assessment, 23,* 656–669. http://dx.doi.org/10.1037/a0023089

Gröpel, P., Baumeister, R. F., & Beckmann, J. (2014). Action versus state orientation and self-control performance after depletion. *Personality and Social Psychology Bulletin, 40,* 476–487. http://dx.doi.org/10.1177/0146167213516636

Habekost, T., Petersen, A., & Vangkilde, S. (2014). Testing attention: Comparing the ANT with TVA-based assessment. *Behavior Research Methods, 46,* 81–94. http://dx.doi.org/10.3758/s13428-013-0341-2

Hagemeister, C. (1994). *Fehlerneigung beim konzentrierten Arbeiten* [Error liability during continuous mental work]. Berlin, Germany: Köster.

Hagemeister, C. (2007). How useful is the power law of practice for recognizing practice in concentration tests? *European Journal of Psychological Assessment, 23,* 157–165. http://dx.doi.org/10.1027/1015-5759.23.3.157

Hancock, P. A., Ross, J. M., & Szalma, J. L. (2007). A meta-analysis of performance response under thermal stressors. *Human Factors, 49,* 851–877. http://dx.doi.org/10.1518/001872007X230226

Helton, W. S., & Russell, P. N. (2011a). Feature absence-presence and two theories of lapses of sustained attention. *Psychological Research, 75,* 384–392. http://dx.doi.org/10.1007/s00426-010-0316-1

Helton, W. S., & Russell, P. N. (2011b). Working memory load and the vigilance decrement. *Experimental Brain Research, 212,* 429–437. http://dx.doi.org/10.1007/s00221-011-2749-1

Herrmann, D. J. (1982). Know thy memory: The use of questionnaire to assess and study memory. *Psychological Bulletin, 92,* 434–452. http://dx.doi.org/10.1037/0033-2909.92.2.434

Hommel, B. (1998). Automatic stimulus-response translation in dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1368–1384. http://dx.doi.org/10.1037/0096-1523.24.5.1368

Huestegge, L., & Koch, I. (2013). Constraints in task-set control: Modality dominance patterns among effector systems. *Journal of Experimental Psychology: General, 142,* 633–637. http://dx.doi.org/10.1037/a0030156

Huestegge, L., Pieczykolan, A., & Koch, I. (2014). Talking while looking: On the encapsulation of output system representations. *Cognitive Psychology, 73,* 72–91. http://dx.doi.org/10.1016/j.cogpsych.2014.06.001

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods, 46,* 702–721. http://dx.doi.org/10.3758/s13428-013-0411-5

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review, 91,* 153–184. http://dx.doi.org/10.1037/0033-295X.91.2.153

Jahn, P., & Engelkamp, J. (2003). Design-effects in prospective and retrospective memory for actions. *Experimental Psychology, 50,* 4–15. http://dx.doi.org/10.1027//1618-3169.50.1.4

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and Individual Differences, 13,* 869–881. http://dx.doi.org/10.1016/0191-8869(92)90004-9

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam, the Netherlands: Elsevier.

Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *The Quarterly Journal of Experimental Psychology, 62,* 209–218. http://dx.doi.org/10.1080/17470210802240655

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin, 136,* 849–874. http://dx.doi.org/10.1037/a0019842

Koriat, A., & Greenberg, S. N. (1996). The enhancement effect in letter detection: Further evidence for the structural model of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1184–1195. http://dx.doi.org/10.1037/0278-7393.22.5.1184

Kraepelin, E. (1902). Die Arbeitskurve [the work curve]. *Philosophische Studien, 19,* 459–507.

Kramer, R. S. S., Weger, U. W., & Sharma, D. (2013). The effect of mindfulness meditation on time perception. *Consciousness and Cognition, 22,* 846–852. http://dx.doi.org/10.1016/j.concog.2013.05.008

Krumm, S., Schmidt-Atzert, L., Buehner, M., Ziegler, M., Michalczyk, K., & Arrow, K. (2009). Storage and non-storage components of working memory predicting reasoning: A simultaneous examination of a wide range of ability factors. *Intelligence, 37,* 347–364. http://dx.doi.org/10.1016/j.intell.2009.02.003

Krumm, S., Schmidt-Atzert, L., & Eschert, S. (2008). Investigating the structure of attention: How do test characteristics of paper-pencil sustained attention tests influence their relationship with other attention tests? *European Journal of Psychological Assessment, 24,* 108–116. http://dx.doi.org/10.1027/1015-5759.24.2.108

Langner, R., & Eickhoff, S. B. (2013). Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychological Bulletin, 139,* 870–900. http://dx.doi.org/10.1037/a0030694

Langner, R., Eickhoff, S. B., & Steinborn, M. B. (2011). Mental fatigue modulates dynamic adaptation to perceptual demand in speeded detection. *PLoS ONE, 6,* e28399. http://dx.doi.org/10.1371/journal.pone.0028399

Langner, R., Steinborn, M. B., Chatterjee, A., Sturm, W., & Willmes, K. (2010). Mental fatigue and temporal preparation in simple reaction-time performance. *Acta Psychologica, 133,* 64–72. http://dx.doi.org/10.1016/j.actpsy.2009.10.001

Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychologica, 104,* 167–190. http://dx.doi.org/10.1016/S0001-6918(00)00019-6

Lim, J., & Dinges, D. F. (2010). A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychological Bulletin, 136,* 375–389. http://dx.doi.org/10.1037/a0018883

Logan, G. D. (1988). Toward and instance theory of automatization. *Psychological Review, 95,* 492–527. http://dx.doi.org/10.1037/0033-295X.95.4.492

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 883–914. http://dx.doi.org/10.1037/0278-7393.18.5.883

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addision Wesley.

Los, S. A., Hoorn, J. F., Grin, M., & Van der Burg, E. (2013). The time course of temporal preparation in an applied setting: A study of gaming behavior. *Acta Psychologica, 144,* 499–505. http://dx.doi.org/10.1016/j.actpsy.2013.09.003

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Maloney, E. A., Risko, E. F., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Psychologica, 134,* 154–161. http://dx.doi.org/10.1016/j.actpsy.2010.01.006

Mayer, E., Reicherts, M., Deloche, G., Willadino-Braga, L., Taussik, I., Dordain, M., . . . Annoni, J. M. (2003). Number processing after stroke: Anatomoclinical correlations in oral and written codes. *Journal of the International Neuropsychological Society, 9,* 899–912. http://dx.doi.org/10.1017/S1355617703960103

Miles, J. D., & Proctor, R. W. (2012). Correlations between spatial compatibility effects: Are arrows more like locations or words? *Psychological Research, 76,* 777–791. http://dx.doi.org/10.1007/s00426-011-0378-8

Miller, J. (2006). A likelihood ratio test for mixture effects. *Behavior Research Methods, 38,* 92–106. http://dx.doi.org/10.3758/BF03192754

Miller, J., & Schröter, H. (2002). Online response preparation in a rapid serial visual search task. *Journal of Experimental Psychology: Human*

Perception and Performance, 28, 1364–1390. http://dx.doi.org/10.1037/0096-1523.28.6.1364

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review, 20,* 819–858. http://dx.doi.org/10.3758/s13423-013-0404-5

Moore, A., & Malinowski, P. (2009). Meditation, mindfulness and cognitive flexibility. *Consciousness and Cognition, 18,* 176–186. http://dx.doi.org/10.1016/j.concog.2008.12.008

Mussweiler, T., & Strack, F. (2000). The "relative self": Informational and judgmental consequences of comparative self-evaluation. *Journal of Personality and Social Psychology, 79,* 23–38. http://dx.doi.org/10.1037/0022-3514.79.1.23

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction-time. *Psychological Bulletin, 89,* 133–162. http://dx.doi.org/10.1037/0033-2909.89.1.133

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition, 111,* 275–279. http://dx.doi.org/10.1016/j.cognition.2009.02.002

Ollman, R. T., & Billington, M. J. (1972). The deadline model for simple reaction times. *Cognitive Psychology, 3,* 311–336. http://dx.doi.org/10.1016/0010-0285(72)90010-2

Pieters, J. P. M. (1983). Sternberg's additive factor method and underlying psychological processes: Some theoretical considerations. *Psychological Bulletin, 93,* 411–426. http://dx.doi.org/10.1037/0033-2909.93.3.411

Pieters, J. P. M. (1985). Reaction time analysis of simple mental tasks: A general approach. *Acta Psychologica, 59,* 227–269. http://dx.doi.org/10.1016/0001-6918(85)90046-0

Posner, M. I., Klein, R., Summers, J., & Buggie, S. (1973). On the selection of signals. *Memory & Cognition, 1,* 2–12. http://dx.doi.org/10.3758/BF03198062

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Rickard, T. C., Lau, J. S-H., & Pashler, H. (2008). Spacing and the transition from calculation to retrieval. *Psychonomic Bulletin & Review, 15,* 656–661. http://dx.doi.org/10.3758/PBR.15.3.656

Ridderinkhof, K. R. (2002). Micro- and macro-adjustments of task set: Activation and suppression in conflict tasks. *Psychological Research, 66,* 312–323. http://dx.doi.org/10.1007/s00426-002-0104-7

Ruthruff, E., Johnston, J. C., & Remington, R. W. (2009). How strategic is the central bottleneck: Can it be overcome by trying harder? *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1368–1384. http://dx.doi.org/10.1037/a0015784

Sanders, A. F., & Hoogenboom, W. (1970). On effects of continuous active work on performance. *Acta Psychologica, 33,* 414–431. http://dx.doi.org/10.1016/0001-6918(70)90151-4

Saville, C. W. N., Pawling, R., Trullinger, M., Daley, D., Intriligator, J., & Klein, C. (2011). On the stability of instability: Optimising the reliability of intra-subject variability of reaction times. *Personality and Individual Differences, 51,* 148–153. http://dx.doi.org/10.1016/j.paid.2011.03.034

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2009). On the relation of mean reaction time and intraindividual reaction time variability. *Psychology and Aging, 24,* 841–857. http://dx.doi.org/10.1037/a0017799

Schorr, A. (1995). Stand und Perspektiven der angewandten psychologischen Diagnostik: Ergebnis einer repräsentativen Befragung von praktisch arbeitenden westdeutschen Psychologen [Status and perspectives of applied psychometric-testing: Results of a representative survey of western German psychologists]. *Diagnostica, 41,* 3–20.

Schwarz, W., & Miller, J. (2012). Response time models of delta plots with negative-going slopes. *Psychonomic Bulletin & Review, 19,* 555–574. http://dx.doi.org/10.3758/s13423-012-0254-6

Schweickert, R., Giorgini, M., & Dzhafarov, E. (2000). Selective influence and response time cumulative distribution functions in serial-parallel task networks. *Journal of Mathematical Psychology, 44,* 504–535. http://dx.doi.org/10.1006/jmps.1999.1268

Schweizer, K. (1996). The speed-accuracy transition due to task complexity. *Intelligence, 22,* 115–128. http://dx.doi.org/10.1016/S0160-2896(96)90012-4

Schweizer, K. (2001). Preattentive processing and cognitive ability. *Intelligence, 29,* 169–186. http://dx.doi.org/10.1016/S0160-2896(00)00049-0

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition, 33,* 644–653. http://dx.doi.org/10.3758/BF03195331

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steinborn, M. B., Bratzke, D., Rolke, B., Gordijn, M. C. M., Beersma, D. G. M., & Ulrich, R. (2010). The effect of 40 hours of constant wakefulness on number comparison performance. *Chronobiology International, 27,* 807–825. http://dx.doi.org/10.3109/07420521003778765

Steinborn, M. B., Flehmig, H. C., Bratzke, D., & Schröter, H. (2012). Error reactivity in self-paced performance: Highly-accurate individuals exhibit largest post-error slowing. *The Quarterly Journal of Experimental Psychology, 65,* 624–631. http://dx.doi.org/10.1080/17470218.2012.660962

Steinborn, M. B., Flehmig, H. C., Westhoff, K., & Langner, R. (2008). Predicting school achievement from self-paced continuous performance: Examining the contributions of response speed, accuracy, and response speed variability. *Psychology Science Quarterly, 50,* 613–634.

Steinborn, M. B., Flehmig, H. C., Westhoff, K., & Langner, R. (2009). Differential effects of prolonged work on performance measures in self-paced speed tests. *Advances in Cognitive Psychology, 5,* 105–113. http://dx.doi.org/10.2478/v10053-008-0070-8

Steinborn, M. B., & Huestegge, L. (2016). A walk down the lane gives wings to your brain: Restorative benefits of rest breaks on cognition and self-control. *Applied Cognitive Psychology, 30,* 795–805. http://dx.doi.org/10.1002/acp.3255

Steinborn, M. B., & Langner, R. (2011). Distraction by irrelevant sound during foreperiods selectively impairs temporal preparation. *Acta Psychologica, 136,* 405–418. http://dx.doi.org/10.1016/j.actpsy.2011.01.008

Steinborn, M. B., & Langner, R. (2012). Arousal modulates temporal preparation under increased time uncertainty: Evidence from higher-order sequential foreperiod effects. *Acta Psychologica, 139,* 65–76. http://dx.doi.org/10.1016/j.actpsy.2011.10.010

Steinborn, M. B., Langner, R., Flehmig, H. C., & Huestegge, L. (2016). Everyday life cognitive instability predicts simple reaction-time variability: Analysis of reaction time distributions and delta plots. *Applied Cognitive Psychology, 30,* 92–102. http://dx.doi.org/10.1002/acp.3172

Steinborn, M. B., Langner, R., & Huestegge, L. (2016). Mobilizing cognition for speeded action: Try-harder instructions promote motivated readiness in the constant-foreperiod paradigm. *Psychological Research*. Advance online publication. http://dx.doi.org/10.1007/s00426-016-0810-1

Steinhauser, M., & Hübner, R. (2006). Response-based strengthening in task shifting: Evidence from shift effects produced by errors. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 517–534. http://dx.doi.org/10.1037/0096-1523.32.3.517

Stolz, J. A., Besner, D., & Carr, T. H. (2005). Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition, 12,* 284–336.

Strobach, T., Frensch, P., Müller, H., & Schubert, T. (2015). Evidence for the acquisition of dual-task coordination skills in older adults. *Acta Psychologica, 160,* 104–116. http://dx.doi.org/10.1016/j.actpsy.2015.07.006

Strobach, T., Frensch, P. A., & Schubert, T. (2012). Video game practice optimizes executive control skills in dual-task and task switching situations. *Acta Psychologica, 140,* 13–24. http://dx.doi.org/10.1016/j.actpsy.2012.02.001

Sturm, W., Willmes, K., Orgass, B., & Hartje, W. (1997). Do specific attention deficits need specific training? *Neuropsychological Rehabilitation, 7,* 81–103. http://dx.doi.org/10.1080/713755526

Stuss, D. T., Bisschop, S. M., Alexander, M. P., Levine, B., Katz, D., & Izukawa, D. (2001). The Trail Making Test: A study in focal lesion patients. *Psychological Assessment, 13,* 230–239. http://dx.doi.org/10.1037/1040-3590.13.2.230

Stuss, D. T., Meiran, N., Guzman, D. A., Lafleche, G., & Willmer, J. (1996). Do long tests yield a more accurate diagnosis of dementia than short tests? A comparison of 5 neuropsychological tests. *Archives of Neurology, 53,* 1033–1039. http://dx.doi.org/10.1001/archneur.1996.00550100119021

Stuss, D. T., Murphy, K. J., Binns, M. A., & Alexander, M. P. (2003). Staying on the job: The frontal lobes control individual performance variability. *Brain: A Journal of Neurology, 126,* 2363–2380. http://dx.doi.org/10.1093/brain/awg237

Suhr, J. A., & Wei, C. (2013). Symptoms as an excuse: Attention deficit/hyperactivity disorder symptom reporting as an excuse for cognitive test performance in the context of evaluative threat. *Journal of Social and Clinical Psychology, 32,* 753–769. http://dx.doi.org/10.1521/jscp.2013.32.7.753

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: A meta-analytic synthesis. *Psychological Bulletin, 137,* 682–707. http://dx.doi.org/10.1037/a0023987

Szalma, J. L., & Teo, G. W. L. (2012). Spatial and temporal task characteristics as stress: A test of the dynamic adaptability theory of stress, workload, and performance. *Acta Psychologica, 139,* 471–485. http://dx.doi.org/10.1016/j.actpsy.2011.12.009

Thomaschke, R., & Dreisbach, G. (2015). The time-event correlation effect is due to temporal expectancy, not to partial transition costs. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 196–218. http://dx.doi.org/10.1037/a0038328

Thomaschke, R., Hoffmann, J., Haering, C., & Kiesel, A. (2016). Time-based expectancy for task-relevant stimulus features. *Time and Time Perception, 4,* 248–270. http://dx.doi.org/10.1163/22134468-00002069

Thorne, D. R. (2006). Throughput: A simple performance index with desirable characteristics. *Behavior Research Methods, 38,* 569–573. http://dx.doi.org/10.3758/BF03193886

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123,* 34–80. http://dx.doi.org/10.1037/0096-3445.123.1.34

Ulrich, R., Miller, J., & Schröter, H. (2007). Testing the race model inequality: An algorithm and computer programs. *Behavior Research Methods, 39,* 291–302. http://dx.doi.org/10.3758/BF03193160

Vanbreukelen, G. J. P., Roskam, E. E. C. I., Eling, P. A. T. M., Jansen, R. W. T. L., Souren, D. A. P. B., & Ickenroth, J. G. M. (1995). A model and diagnostic measures for response time series on tests of concentration: Historical background, conceptual framework, and some applications. *Brain and Cognition, 27,* 147–179. http://dx.doi.org/10.1006/brcg.1995.1015

van Ravenzwaaij, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition, 119,* 381–393. http://dx.doi.org/10.1016/j.cognition.2011.02.002

Waechter, S., Stolz, J. A., & Besner, D. (2010). Visual word recognition: On the reliability of repetition priming. *Visual Cognition, 18,* 537–558. http://dx.doi.org/10.1080/13506280902868603

Wei, C., & Suhr, J. A. (2015). Examination of the role of expectancies on task performance in college students concerned about ADHD. *Applied Neuropsychology Adult, 22,* 204–208. http://dx.doi.org/10.1080/23279095.2014.902836

West, R., Murphy, K. J., Armilio, M. L., Craik, F. I. M., & Stuss, D. T. (2002). Lapses of intention and performance variability reveal age-related increases in fluctuations of executive control. *Brain and Cognition, 49,* 402–419. http://dx.doi.org/10.1006/brcg.2001.1507

Westhoff, K. (1985). Eine erste Prüfung einer Konzentrationstheorie [A first test of an assessment model of sustained attention in active tasks]. *Diagnostica, 31,* 265–278.

Westhoff, K., & Lemme, M. (1988). Eine erweiterte Prüfung einer Konzentrationstheorie [Extended testing of an assessment model of sustained attention in active tasks]. *Diagnostica, 34,* 244–255.

Wetter, O. E., Wegge, J., Jonas, K., & Schmidt, K.-H. (2012). Dual goals for speed and accuracy on the same performance task: Can they prevent speed-accuracy trade-offs? *Journal of Personnel Psychology, 11,* 118–126. http://dx.doi.org/10.1027/1866-5888/a000063

Wilhelm, O., Witthöft, M., & Schipolowski, S. (2010). Self-reported cognitive failures: Competing measurement models and self-report correlates. *Journal of Individual Differences, 31,* 1–14. http://dx.doi.org/10.1027/1614-0001/a000001

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General, 136,* 623–638. http://dx.doi.org/10.1037/0096-3445.136.4.623

(*Appendix follows*)

Appendix

Table A1
Descriptive Performance Data for the Time-on-Task (TOT) Effect in the d2 Sustained-Attention Test

	Session 1				Session 2			
	Speed		$Speed_C$		Speed		$Speed_C$	
TOT-level	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	37.94	6.74	36.23	6.36	42.27	5.00	41.39	5.01
2	35.88	6.13	34.81	6.06	39.55	5.70	38.89	5.58
3	37.95	6.13	37.00	5.97	41.11	5.40	40.55	5.28
4	35.31	6.32	34.12	6.04	38.52	6.19	37.82	6.09
5	37.36	6.25	36.26	6.06	40.14	5.61	39.39	5.52
6	37.70	6.07	36.78	6.03	40.84	5.28	40.08	5.26
7	34.88	6.50	33.41	6.23	38.21	5.51	37.29	5.45
8	36.39	5.85	35.43	5.67	39.05	5.50	38.43	5.41
9	36.95	6.66	36.03	6.40	40.58	5.15	39.97	5.04
10	30.88	5.90	32.62	5.54	37.76	5.67	36.99	5.66
11	35.93	5.96	34.89	5.54	39.19	5.90	38.57	5.77
12	36.21	6.03	35.14	5.91	39.30	5.71	38.80	5.78
13	33.35	6.23	32.13	5.74	37.23	5.90	36.50	5.85
14	34.96	6.28	33.73	5.75	38.75	5.44	37.79	5.70

Note. $Speed_C$ = error-corrected measure of speed defined as the average number of correctly cancelled items per row. A description of the computation for the presented performance measures is provided in Table 1.

Table A2
Descriptive Performance Data for the Time-on-Task (TOT) Effect in the d2 Sustained-Attention Test

	Session 1				Session 2			
	Omission		Commission		Omission		Commission	
TOT-level	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	1.64	1.80	.07	.33	.87	1.24	.01	.19
2	1.03	1.51	.04	.20	.64	1.23	.02	.14
3	.85	1.32	.10	.36	.56	.87	.00	.00
4	1.13	1.38	.06	.34	.86	1.17	.02	.14
5	1.06	1.67	.04	.19	.75	1.33	.00	.00
6	.86	1.26	.06	.28	.75	1.20	.01	.10
7	1.37	1.81	.10	.44	.91	1.31	.01	.10
8	.92	1.47	.04	.20	.61	1.11	.01	.10
9	.81	1.50	.11	.55	.59	.99	.02	.14
10	1.15	1.75	.11	.37	.73	1.09	.04	.20
11	1.03	1.60	.01	.10	.61	1.11	.01	.10
12	.98	1.52	.09	.53	.48	.86	.02	.14
13	1.16	1.74	.06	.28	.70	.99	.03	.22
14	1.20	1.86	.03	.30	.96	2.34	.00	.00

Note. A description of the computation for the presented performance measures is provided in Table 1.

(*Appendix continues*)

Table A3
Descriptive Data for the Cumulative Distributive Function in the d2 Sustained-Attention Test

	Session 1				Session 2			
	CDF		CDF_C		CDF		CDF_C	
Percentile	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
.05	30.37	5.51	29.56	5.39	34.82	6.04	32.65	5.67
.15	32.70	5.52	31.78	5.37	36.67	5.65	35.54	5.61
.25	33.79	5.50	32.99	5.43	37.57	5.51	36.61	5.46
.35	34.66	5.56	33.93	5.33	38.37	5.24	37.45	5.31
.45	35.92	5.64	35.13	5.55	39.29	5.14	38.46	5.11
.55	36.50	5.70	35.58	5.56	39.85	5.18	39.00	5.08
.65	37.57	5.78	36.60	5.52	40.83	4.92	39.98	4.87
.75	38.24	5.75	37.30	5.42	41.34	4.79	40.58	4.76
.85	39.45	5.44	38.15	5.38	42.07	4.57	41.25	4.62
.95	41.27	5.12	39.75	5.28	43.74	3.93	42.99	3.98

Note. Values are taken from a vincentized interpolated cumulative distributive function (CDF) of performance speed. A detailed description is provided in the Method section.

Table A4
Reliability Coefficients of Square-Root Transformed Error Scores and Intercorrelations with the Performance Measures in the d2 Sustained-Attention Test (Supplement in Addition to Table 3)

	Session 1									
Session 2	1	2	3	4	5	6	7	8	9	10
1. Speed	—	—	.18	.18	.04	—	—	—	—	—
2. $Speed_C$	—	—	−.04	−.04	.01	—	—	—	—	—
3. Error (%)	.05	−.11	.75**	.99**	.19	.23*	.12	.22*	.07	.07
4. Error (omission)	.06	−.10	−.99**	.75**	.11	.23*	.12	.22*	.07	.07
5. Error (comission)	−.09	−.11	.14	.08	.28**	.08	.03	.07	.02	.00
6. Variability (*SD*)	—	—	.14	.13	.25*	—	—	—	—	—
7. Variability (CV)	—	—	.09	.07	.22*	—	—	—	—	—
8. Max	—	—	.07	.08	−.03	—	—	—	—	—
9. Min	—	—	−.03	−.01	−.17	—	—	—	—	—
10. Range	—	—	.05	.04	.19	—	—	—	—	—

Note. Test–retest reliability is shown in the main diagonal (denoted grey); Correlation coefficients for the first testing session are shown above and for the second testing session below the main diagonal. $Speed_C$ = error-corrected measure of speed defined as the average number of correctly cancelled items per row; variability (CV) = coefficient of variation of performance speed; variability (SD) = reaction time standard deviation of performance speed. Significant correlations are denoted ($N = 100$. * for $r \geq .21$. $p < .05$. ** for $r \geq .34$. $p < .01$).

Received June 23, 2016
Revision received January 30, 2017
Accepted March 16, 2017 ■