TYPE Original Research
PUBLISHED 18 August 2022
DOI 10.3389/fpsyg.2022.946626

OPEN ACCESS

EDITED BY
Miriam Gade,
Medical School Berlin, Germany

REVIEWED BY
Thomas Kleinsorge,
Leibniz Research Centre for Working Environment and Human Factors (IfADo), Germany
Christoph Naefgen,
University of Hagen, Germany

*CORRESPONDENCE
Frank Schumann
schumann@hs-mittweida.de

SPECIALTY SECTION
This article was submitted to Cognition,
a section of the journal Frontiers in Psychology

RECEIVED 17 May 2022
ACCEPTED 29 July 2022
PUBLISHED 18 August 2022

CITATION
Schumann F, Steinborn MB, Flehmig HC, Kürten J, Langner R and Huestegge L (2022) On doing multi-act arithmetic: A multitrait-multimethod approach of performance dimensions in integrated multitasking. *Front. Psychol.* 13:946626. doi: 10.3389/fpsyg.2022.946626

On doing multi-act arithmetic: A multitrait-multimethod approach of performance dimensions in integrated multitasking

Frank Schumann[1]*, Michael B. Steinborn[2], Hagen C. Flehmig[3], Jens Kürten[2], Robert Langner[4,5] and Lynn Huestegge[2]

[1]Mittweida University of Applied Sciences, Mittweida, Germany, [2]Julius-Maximilians-Universität Würzburg, Würzburg, Germany, [3]Technische Universität Dresden, Dresden, Germany, [4]Medical Faculty, Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [5]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Center Jülich, Jülich, Germany

Here we present a systematic plan to the experimental study of test–retest reliability in the multitasking domain, adopting the *multitrait-multimethod (MTMM) approach* to evaluate the psychometric properties of performance in *Düker-type* speeded multiple-act mental arithmetic. These form of tasks capacitate the experimental analysis of *integrated multi-step processing* by combining multiple mental operations in flexible ways in the service of the overarching goal of completing the task. A particular focus was on scoring methodology, particularly measures of response speed variability. To this end, we present data of two experiments with regard to (a) test–retest reliability, (b) between-measures correlational structure, (c) and stability (test–retest practice effects). Finally, we compared participants with high versus low performance variability to assess ability-related differences in measurement precision (typically used as proxy to "simulate" patient populations), which is especially relevant in the applied fields of clinical neuropsychology. The participants performed two classic integrated multi-act arithmetic tasks, combining addition and verification (Exp. 1) and addition and comparison (Exp. 2). The results revealed excellent test–retest reliability for the standard and the variability measures. The analysis of between-measures correlational structure revealed the typical pattern of convergent and discriminant relationships, and also, that absolute response speed variability was highly correlated with average speed ($r > 0.85$), indicating that these measures mainly deliver redundant information. In contrast, speed-adjusted (relativized) variability revealed discriminant validity being correlated to a much lesser degree with average speed, indicating that this measure delivers additional information not already provided by the speed measure. Furthermore, speed-adjusted variability was virtually unaffected by test–retest practice, which makes this measure interesting in situations with repeated testing.

KEYWORDS

sustained attention, concentration, vigilance, reliability, cognitive control

1. Introduction

Sustaining mental focus to continuous activity is perceived as arduous and often hard to keep up over a prolonged time period (Humphreys and Revelle, 1984; Langner and Eickhoff, 2013). This is particularly true when a task requires more than a single mental operation and instead consists of a composite of subordinate actions, or a conglomerate of nested mental acts which form a coordinated ensemble where the whole is more than the integrated sum of its parts (Greenwald, 1970, 1972; Navon and Gopher, 1979; Vandierendonck et al., 2010; Huestegge et al., 2014). Everyday examples can be found in the borrow operation in complex subtractions, or the carry operation in complex multiplications (Ashcraft, 1992). Researchers and practitioners in work-related diagnostic settings often employ speeded (elementary and multitasking) tests for the purpose of assessing individuals' abilities in the context of personnel selection and classification (Cronbach and Meehl, 1955; Campbell and Fiske, 1959). While elementary forms are typically constructed as "*Bourdon-style*" tests, putting excessive demands on perceptual identification, more complex (multitasking) forms typically utilize primary cultural techniques such as mental arithmetic, tapping more into the supervisory scheduling of multiple operations in close temporal succession, which, in their entirety, form an integrated holistic *"Düker-style"* test (Düker, 1949). The present study presents a psychometric analysis of tasks from the latter (multi-operation) category adopting the multitrait-multimethod (MTMM) approach (Campbell and Fiske, 1959; Miller and Ulrich, 2013).

1.1. The psychometrics of multitasking

In their momentous work that formed a milestone in the development of psychometric theory, Düker and Lienert (1949) stated that "*. . .Kapazität entsteht durch das geordnete Zusammenwirken von Einzeltätigkeiten zu einer Handlung durch optimale Koordination. . .". ["*. . .capacity is created by the well-ordered interplay of individual operations to form an action through optimized coordination. . ."]*. The authors held the position that the concept of general cognitive ability is best represented by a test that requires the speeded coordination of elemental mental acts (e.g., recording, calculating, memorizing, rule retrieval, ordering and sequencing, etc.) serving the overarching objective to complete the task. Based on this theorizing, Düker and Lienert (1949) developed the Konzentrations-Leistungs-Test (KLT) ("concentration ability test"), which later on became one of the standard measures in the psychometric testing of elementary cognitive ability. In contrast to *Bourdon-type* cancelation tests (Bates and Lemay, 2004), the KLT requires individuals to work on relatively complex *compound multiple-act* arithmetic tasks over a period of about 20 min. A pivotal point is the inherent multitasking nature of the task (Logan, 1979; Ashcraft, 1992; Pashler, 1994b; Bruning and Manzey, 2018), with each item requiring multistep mental operations that include elemental acts (e.g., addition, subtraction, memorizing) serving to complete the overall task as the primary goal (Oberauer, 2002, 2003; Botvinick and Bylsma, 2005a,b; Janczyk and Kunde, 2010, 2020; Herbort and Rosenbaum, 2014).

Although Düker and Lienert (1949) were less concerned with the cognitive analysis of tasks but with the utility of cognitive operations for psychometric testing, they are regarded as pioneers in the history of a cognitive–psychometric discipline (Pieters, 1983, 1985; Blotenberg and Schmidt-Atzert, 2019a,b). Unlike most psychometricians at this time, who considered test reliability merely a statistical issue to be resolved by simple score accumulation methods (cf. Cronbach, 1947; Lord and Novick, 1968; Kahneman et al., 2021, pp. 55–69), the authors went even further, asking of what exactly determines test reliability and how these factors can be modeled in an experimental setting. According to their view, a psychologically substantiated perspective of reliability must offer the possibility to theorize on the underlying processes that either promote or hamper measurement precision (e.g., reliability affected by motivation, task complexity, repeated testing, etc.). For example, to understand what is meant by task complexity, researchers typically distinguish between automatic (intuitive) and controlled (reflective) components of arithmetic processing (Ashcraft and Battaglia, 1978; Manstead and Semin, 1980; Strack and Deutsch, 2004), and a standard procedure to measure these components is to experimentally manipulate item difficulty in some way, for example, by varying arithmetic-chain length

(Steinborn and Huestegge, 2016, 2017, 2020), adding a memory load (Logan, 1979; Gopher, 1996; Vallesi et al., 2014), or by dynamically switching ongoing mental operations (Bruning et al., 2020, 2021, 2022).

An important aspect of theorizing concerns the effect of practice on test reliability, which is crucial in situations where repeated testing takes place (Hagemeister, 2007; Hagemeister and Kronmaier, 2017). According to Logan (1988), the most general way of theorizing on practice effects is to assume two distinct processes of solving speeded arithmetic, a calculation-based process and a process based on memory retrieval. Critical is that both processes are running in parallel and in competition to each other. Performance is considered automatic when based on single-act, direct-route retrieval of results from memory, while it is considered controlled when based on algorithmic processing such as counting, adding, memorizing, borrowing (Ashcraft, 1992), or negating a logical term (Deutsch et al., 2006, 2009; Goodwin and Johnson-Laird, 2011, 2013). Key to this idea is that these processes race against each other so that each unique trial is finally cleared up by either the retrieval or the algorithmic operation. Practice gains (due to retesting) occur, according to this conception, because repeated exposure leads to an accumulation of separate episodic traces with experience, which gives them a race advantage over the algorithmic process (Miller and Ulrich, 2003, 2013; Han and Proctor, 2022a,b,c). To say it another way, retesting produces a gradual transition from algorithmic processing to memory-based processing and thus changes the relation (i.e., the mixture parameter) of both as a function of amount of practice (Compton and Logan, 1991; Pashler and Baylis, 1991a,b; Steinborn et al., 2009, 2010b; Los et al., 2014, 2017, 2021; Crowe and Kent, 2019).

Although the concept of coordination, which lies at the core of the mental demands in Düker-style tasks, is considered a trait-like characteristic of normal individuals (Düker, 1949; Bruning et al., 2021), it is also widely resorted to in clinical and neuropsychological contexts to assess the level of cognitive functioning in patients (Bates and Lemay, 2004; Stuss et al., 2005), or wherever individual-case assessment in this cognitive domain is indicated (Willmes, 1985; Stuss et al., 2001). In a broader sense, coordination is a natural ingredient of many everyday activities and of high relevance to research on everyday multitasking (Botvinick and Bylsma, 2005a,b; Salvucci and Taatgen, 2008, 2011). It is crucial to the understanding of individual differences in cognitive performance (Ackerman, 1987; Ackerman and Kanfer, 2009; Steinborn et al., 2016, 2018; Bruning et al., 2020, 2022), though any strict definition naturally depends on the particular (task- and time-based) characteristics (Thomaschke et al., 2012; Thomaschke and Dreisbach, 2015). A study of Bruning et al. (2021) performed an in-depth analysis of individual differences and found a great diversity of how a task is represented by individuals and how this determines response organization (cf. Phillips and Rabbitt, 1995; Watson and Strayer, 2010; Cheyne et al., 2011; Steinborn et al., 2012; Schumann et al., 2022).

1.2. Multitrait-multimethod approach

A unique feature of both elementary and complex (integrated multiple-act) speed tests is that the they are administered in a self-paced mode, which places particular emphasis on the supervisory monitoring of the proper speed–accuracy balance (cf., Rabbitt and Banerji, 1989; Jentzsch and Leuthold, 2006), and that several ways of measuring performance are possible, each with a distinct meaning (Steinborn et al., 2016, 2018). While average performance speed is typically considered the primary measure, error rate serves as a secondary measure, held to indicate rigor, diligence, or punctiliousness, or a lack thereof, respectively (Bates and Lemay, 2004). Speed and accuracy are often not or only modestly correlated which is taken as an argument for the discriminant validity of both measures (Flehmig et al., 2007; Steinborn et al., 2016, 2018; Blotenberg and Schmidt-Atzert, 2019a,b). It is often ignored, however, that error scores (with errors being rare events) exhibit a skewed population distribution, which limits test reliability, which again poses a limit on correlational relationships with other performance indices. Test guidelines often recommend combining measures of speed and accuracy into a single compound dimension, either as penalty-based combination score (Bruyer and Brysbaert, 2011; Gropel et al., 2014; Steinborn et al., 2016, 2018; Wuhr and Ansorge, 2020), or on grounds of model-based reasoning (Vandierendonck, 2017; Liesefeld and Janczyk, 2019), others resort to throughput, defined as the rate of work in a given time frame (Thorne, 2006; Szalma and Teo, 2012). In self-paced tasks, combining speed and accuracy to a compound measure of (inversed) efficiency often yields a slight improvement of reliability (Pieters, 1983, 1985; Van Breukelen et al., 1995; Steinborn et al., 2018, p. 350).

Recent research increasingly focused on measuring the fluctuation of performance, and connected with this point, how this concept could be indexed reliably (Jensen, 1992; Leth-Steensen et al., 2000; Flehmig et al., 2007; Unsworth, 2015; Steinborn et al., 2016, 2018; Fortenbaugh et al., 2017; Unsworth and Robison, 2020). A majority of studies examining intra-individual performance variability resorted to reaction-time standard deviation (RTSD), which seems natural at first sight, given that most statistics textbooks refer to *SD* as the appropriate measure of dispersion of metric-scale data points around their arithmetic mean. However, RTSD is, for pure mathematical reasons, highly correlated with the mean response time (RTM), and due to this redundancy, of only limited diagnostic value. Flehmig et al. (2007) suggested the response time coefficient of variation (RTCV) to index variability, which relates RTSD to the individual's RTM, yielding an index of variability relative to the individual's overall level of performance speed (cf. Steinborn et al., 2016, 2018). RTCV is calculated by dividing the individual RTSD by the individual RTM, multiplied by 100: RTCV = (RTSD/RTM) × 100. As a result, a measure is obtained that allows for comparing intra-individual RT

TABLE 1 A brief guide to understanding the logic underlying the Multitrait-multimethod (MTMM) approach to individual differences in multitasking.

	Step	Metaphor and symbolic assumptions
1	Skew check	- the *molecular* precondition: symmetry of population distribution - skew indicates lack of variance through bias (bottom or ceiling effect) - skew → no information in the data → no expectation of correlation - example: rare events (errors, oddball effects, etc.) & self-ratings often skewed
2	Reliability	- the *molar* precondition: reliability limits correlational relationships between variables - typical retest intervals: 1–2 weeks for performance tests, 4 weeks for questionnaires - factors affecting reliability: number of trials, scaling differences (metric vs. %) - required for claim such as, e.g., "*. . .results show that the constructs are independent. . .*"
3	Convergence	- theoretically similar concepts → expected to be related empirically - is given when two indicators representing the same concept are highly correlated - e.g.: "ability" convergently represented through indicators (efficiency vs. throughput) - a way of judging indicator utility
4	Discriminance	- theoretically different concepts → expected not to be related empirically - is given when two indicators representing different concepts are not correlated - e.g., a high correlation of RTM with RTSD could be seen as lack of discrimination - often produced by method similarity, or natural mathematical relations
5	Stability	- concerns the degree to which (absolute) scores remain constant from test to retest - stability → true score (ability) of person has not changed after repeated testing - main factors biasing stability: test-taker strategies and practice gains - practice gains often unequal across participants → impede reliability scores
6	Reproducibility	- concerns the between-session correlational structure - comparison of correlation structure above vs. below the MTMM reliability diagonal - indicates that the convergent and divergent relationships are stable and reproducible - in essence, a qualitative way of judging the robustness of a nomological network
7	Generalizability	- concerns the replicability of the overall findings with conceptually similar tests - judging whether conclusions are specific or can be made with some scope of validity - generality of findings generates essence → scientific substance → knowledge - on principle, a serial process of corroborating scope and substantiality of concepts

The points 1–2 are basic preconditions that must be fulfilled in order to enable any expectation about relationships of variables with each other. The points 3–4 are the classic evaluation dimensions of the MTMM, as they give an indication of how close (vs. distant) concepts are empirically. The points 5–7 are, in a strict sense, qualitative dimensions of credibility control, achieved through a serial process of replication with slightly varied conceptual variation (cf. Stroebe and Strack, 2014, pp. 61–63; Miller and Ulrich, 2021, 2022).

variability beyond mere–scaling variability (Wagenmakers and Brown, 2007; Steinborn et al., 2017) and is thus suitable for comparing variability even of individuals who differ very much in their average cognitive speed (Jensen and Rohwer, 1966; Neubauer and Fink, 2009).

Miller and Ulrich (2013) have recently developed the individual-differences in response time (IDRT) model which is based on classical test theory and introduces the analysis of standard psychometric criteria (i.e., reliability, convergent vs. discriminant validity, stability) within an RT modeling framework (cf. Campbell and Fiske, 1959; Steinborn et al., 2016, 2018). This approach enables a formal and systematic investigation of the question of which aspects of mental processing time affect the reliabilities and correlations of RT-based measures as assessed with standard psychometric tests. According to the IDRT model, empirically observed RTM is composed of three separate components, individual differences in global processing speed (cf. Bruning et al., 2021; Bruning et al., 2022, for a theoretical view), processing time that is imposed by a certain experimental variable (e.g., effect of increasing workload), a residual term, and an error term. Briefly, IDRT can be described as being composed of (1) person-specific general processing time, (2) task-specific processing time, (3) residual time, and (4) measurement error. In this regard, Miller and Ulrich's theorizing implicates a hierarchical evaluative analysis of empirical correlations within both measures of same test's performance and across measures of different tests' performance (Cronbach and Meehl, 1955; Campbell and Fiske, 1959; Gleser et al., 1965; Rajaratnam et al., 1965; Steinborn et al., 2016, 2018; Hedge et al., 2018).

The *first* precondition is to evaluate whether the population parameter of all performance scores exhibit a sufficient level of symmetry and variance (see **Table 1**). A skewed population distribution would indicate lack of variance either because of a bottom or a ceiling effect (Dunlap et al., 1994; Greer et al., 2006). If there is no information in the data, there will be no expectation of a potential correlation. For example, error scores in speeded tests often reveal a skewed distribution and thus lack test reliability thereof (Hagemeister, 2007; Steinborn et al., 2012; Wessel, 2018). In the *second* step, the reliability diagonal is evaluated which is the precondition to obtain correlations with other measures. This is particularly important in studies where measurable entities are claimed as being *"independent"* concepts. The *third* step is to determine convergent validity, indicating the degree to which a theoretically related concept is actually interrelated empirically. The *fourth* step is to determine discriminant validity, indicating the degree to which a theoretically unrelated concept is not interrelated empirically. In order to determine construct validity of the target measures, one has to demonstrate reliability as a precondition and both convergence and discrimination. Mental arithmetic is especially suitable for constructing tests because of its flexibility to generate items with desirable psychometric characteristics. This concerns two aspects, finding adequate levels of difficulty (e.g., varying problem size) and mitigating practice gains from repeated testing (e.g., increasing item set). Using staggered multi-step processing by combining multiple mental operations (e.g., addition combined with subsequent verification, negation, comparison, memorization) allows for a flexible arrangement of items enabling various options to control for psychometric criteria (Restle, 1970; Ashcraft, 1992; Rickard, 2005; Steinborn et al., 2012).

1.3. Analysis of integrated Multiple–Act arithmetic

Here, we asked whether common performance measures obtained from integrated multi-act arithmetic exhibit sufficient test–retest reliability, discriminant validity, and robustness against practice from repeated testing. We followed the reasoning implied by the IDRT model (Miller and Ulrich, 2013) and the multitrait–multimethod approach (Campbell and Fiske, 1959; Steinborn et al., 2018) as a heuristic means to evaluate convergent and discriminant validity of competing performance measures. To this end, we investigated test–retest reliability, correlational structure, and liability to practice effects, indexing average speed (RTM), error percentage (EP), absolute (RTSD), and relativized (RTCV) reaction-time variability. Data of two experiments are presented, each requiring the integrated coordination of elemental acts. In Experiment 1, we analyzed performance using a mental addition and verification paradigm (Zbrodoff and Logan, 1986, 1990; Steinborn and Huestegge, 2016, 2017, 2020), where participants are presented with an addition term including its result (e.g., 2 + 3 = 5), and are required to verify the correctness of the term by pressing a "yes" or "no" response. In Experiment 2, we used an integrated dual-act mental addition and comparison paradigm (Restle, 1970), where individuals are presented with an addition problem and with a single number, which are spatially separated by a vertical line (e.g., "4 + 5 | 10"); they are instructed to solve the addition problem and then to compare the number value of their calculated result with the number value of the presented digit.

2. Experiment 1

Methods

Participants

Twenty-nine young adults (age between 20–30, recruited at the Dresden University of Technology) participated in the study. It is intrinsic to the study of test reliability to have a diversified sample population with a (relatively) balanced gender ratio, including participants not only from the faculty of psychology but also from other faculties (the humanities, natural sciences, etc.). There were two testing sessions with a retest interval of 1 week, which took place under similar conditions (i.e., at the same place and at about the same time). Participants had normal or corrected-to-normal vision and reported to be in normal health condition.

Material and apparatus

The self-paced mental addition and verification was administered twice within a retest interval of 3 days (cf. **Figure 1**). Each item in a trial was presented until response, and was replaced immediately after the response (RSI = 50 ms) by the next item (Steinborn and Huestegge, 2016, 2017, 2020). No feedback was given, neither in case of an erroneous response, nor in case of too slow responses. In a typical trial, they were presented with an addition problem including its result, either correct or incorrect (e.g., "4 + 5 = 9", or "4 + 3 = 6"). Problem size was varied to a maximum of 19 (i.e., 2 + 3; 3 + 2 . . .8 + 7, 7 + 8), and ties (e.g., 2 + 2, 3 + 3, 4 + 4, . . ., etc.) were excluded (cf. Blankenberger, 2001). Participants were required to solve the addition problem and then to decide whether the result is correct or incorrect. In case of a correct result (in half of all trials), they had to press the right button ("yes"), and in case of an incorrect result, they had to press the left button ("no"). In an experimental session, 428 trials were presented. The task required no more than about 30 min of testing time.

Procedure and design

The experiment took place in a noise shielded room and was run on a standard personal computer with color display (19", 100 Hz frequency), controlled *via* the software Experimental Runtime System (ERTS), developed by Behringer (1987). Participants were seated at a distance of about 60 cm in front of the computer screen, and the stimuli were presented at the center of the screen.

Results and discussion

Population parameters and correlations are displayed in **Tables 2**, **3**. The first five responses were regarded warming-up trials and not considered for analysis. Responses faster than 100 ms were regarded outliers and discarded from response

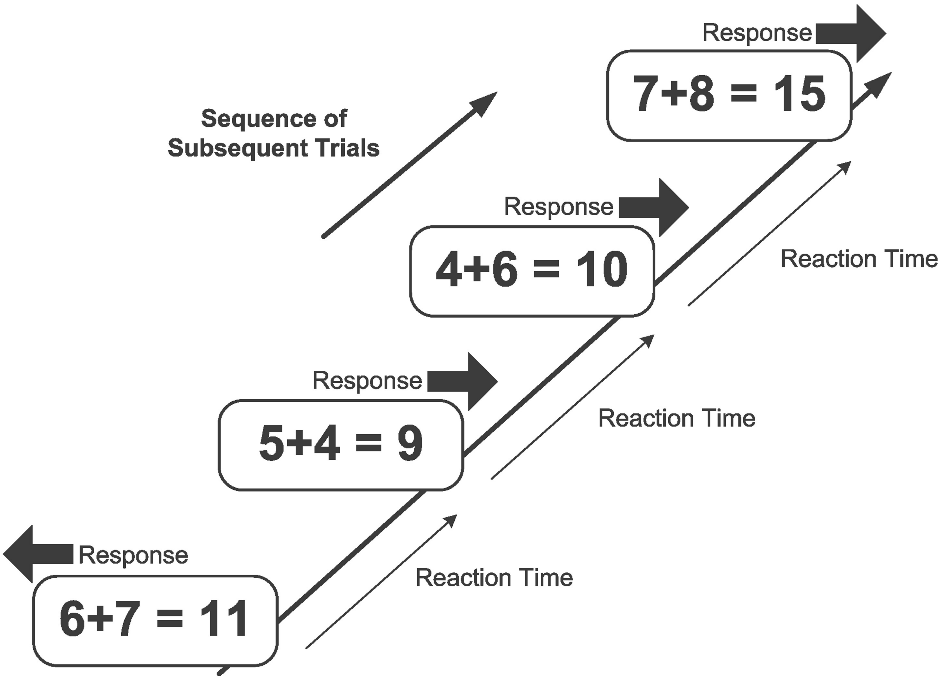

FIGURE 1
Example of a series of trials in the mental addition and verification task. Participants have to perform an addition operation (e.g., “3 + 4 =”) and then to verify (or falsify) the correctness of their mental result with the presented addition term (e.g., “7 = 7”), by pressing the right (or left) response key. The task is self-paced such that each item is presented immediately after the response to the previous item.

time analysis. Correct reactions within this interval were used to compute averaged response time (RTM), standard deviation of response times (RTSD), and coefficient of variation of response times (RTCV). Incorrect responses were computed to index error percentage (EP). MTMM analysis served to evaluate reliability as well as convergent and discriminant correlations of the alternative performance indices in the self-paced speed tests.

Retest reliability

Reliability coefficients are shown along the main diagonal of the correlation matrix (Table 3), presenting the correlations between the first and the second test administration. As expected, RTM was highly reliable (r = 0.93). Surprisingly, good reliability was also obtained for ER (r = 0.84), given that error-score reliability is for the most task low or insufficient (Maloney et al., 2010; Steinborn et al., 2016, 2018; Hedge et al., 2018, for an overview). Note that error scores are often reported as not sufficiently reliable, due to the fact that errors are rare events in chronometric tasks (Jentzsch and Dudschig, 2009; Notebaert et al., 2009; Steinborn et al., 2012; Steinhauser et al., 2017; Wessel, 2018; Dignath et al., 2020; Schaaf et al., 2022), particularly in self-paced psychometric tests (Hagemeister, 2007). Most interestingly, not only RTSD but also RTCV appeared to be highly reliable indices of performance ($r > 0.91$), a finding that deviates a bit from previous studies that demonstrated insufficient reliability for the relativized response speed variability measures.

Correlational structure

There was no relationship between RTM and ER indicating discriminant validity (Table 3). Pronounced (expected) relationships were found between absolute and relativized variability as represented by RTSD and RTM (r = 0.89 and r = 0.87). In fact, the relationship shows that both measures represent a similar aspect of performance. Substantial positive correlations were also found between RTCV and RTM, albeit to a lesser degree (r = 0.69 and r = 0.60). This is in contrast to the previously reported findings where RTCV was observed as less reliable and uncorrelated to RTM.

Practice effects

We performed a repeated-measures analysis of variance including practice (test vs. retest) as factor and performance as dependent measures. A multivariate effect was observed from the first to the second testing session for all performance measures. A more detailed analysis of separate effects revealed that only RTM and RTSD were significantly affected by test–retest practice gains. As expected, individuals became faster on average after practice, indicated by the effects of session on RTM [10% gain; $F(1,28) = 35.1$; $p < 0.000$; $\eta^2 = 0.56$] and on RTSD [10% gain; $F(1,28) = 6.3$; $p < 0.05$; $\eta^2 = 0.19$].

TABLE 2 Descriptive statistics for Experiment 1 and Experiment 2.

Experiment 1 (serial mental addition and verification task)

		Session 1				Session 2			
	Measures	**M**	**SD**	**Skew**	**Range**	**M**	**SD**	**Skew**	**Range**
1	RTM	1402	333	0.23	848–2042	1270	297	0.50	774–2028
2	RTMc	1460	359	0.15	866–2113	1322	304	0.48	794–2112
3	ER	3.86	2.78	0.96	0.23–11.19	3.94	3.68	2.34	0.47–18.18
4	RTSD	679	318	0.25	189–1390	618	280	0.06	185–1141
5	RTCV	0.46	0.14	0.14	0.21–0.78	0.47	0.15	−0.11	0.23–0.76

Experiment 2 (serial mental addition and comparison task)

		Session 1				Session 2			
	Measures	**M**	**SD**	**Skew**	**Range**	**M**	**SD**	**Skew**	**Range**
1	RTM	2063	471	0.78	1191–3035	1744	391	0.51	196–2744
2	RTMc	2128	481	0.78	1237–3229	1785	396	0.52	1133–2776
3	ER	3.05	2.59	3.35	0.33–16.97	2.34	1.97	1.34	0.16–9.14
4	RTSD	1074	494	1.25	383–2743	809	316	0.44	290–1.708
5	RTCV	0.50	0.14	1.59	0.30–1.03	0.45	0.10	0.10	0.26–0.72

Population parameters for all performance measures. N = 29 (Exp. 1); N = 50 (Exp. 2); RTM, response time mean; RTMc, error-corrected RTM (inversed efficiency); ER, error percentage; RTSD, response time standard deviation; RTCV, response time coefficient of variation.

Extended analyses

Miller and Ulrich (2013, p. 824) argued that the average response speed (RTM) should strongly be influenced the intraindividual variability of the same test, as reflected by RTCV, so reliability should decrease as CV increases. The authors argue that "*. . .naturally, researchers should take steps to minimize trial-to-trial fluctuations in arousal, attention, and other factors that might increase RT variability. . .*." Since this assumption remains largely untested for the case of test–retest variability, we here examined it. To this end, we compared groups of individuals who exhibited high versus low performance in the self-paced task (as a proxy to simulate patient populations), partitioning the sample into two groups of individuals according to their RTM, so that each group contained 50% of the sample. Note that though the model theorizes on individual differences in variability, most people would naturally expect to divide them according to speed as critical dimension, with average speed partly depending on variability. For simplicity, therefore, we decided to use speed instead of variability, but we state here the use of variability reveals a very similar pattern. We observed that the reliability of RTM was largest for the high-performance group (r = 0.98**) but was considerably decreased for the low-performance group (r = 0.81**). This finding shows that even with an enormous amount of trials per experimental session (i.e., 424 trials in Exp. 1), RTM measures became less reliable when the analysis was restricted to a subgroup of individuals exhibiting performance deficits, as typically observed in patient groups (Stuss et al., 1996, 2005).

3. Experiment 2

Method

Participants

The sample comprised 50 participants (70% female) who were recruited *via* advertisements and at the Campus of the Dresden University of Technology. All participants had normal or corrected-to-normal vision, and all of them reported to be in good health.

Material and apparatus

The serial mental addition and comparison test (Restle, 1970) was administered twice within a retest interval of 4 days (cf. **Figure 2**). Participants self-paced their responding since each item in a trial was presented until response, replaced "immediately" (i.e., 50 ms RSI) after responding by the next item. No feedback was given, neither in case of an erroneous response, nor in case of too slow responses. In a typical trial, they were presented with an addition problem and with a single number, spatially separated by a vertical line (e.g., "4 + 5 | 10"). They were to solve the addition problem and to compare the number value of their calculated result with the number value of the presented digit. In all trials, the value of the digit was either one point smaller or one point larger than the value of the addition term but never of equal value. Participants were required to "choose" the larger number value by pressing

TABLE 3 Multitrait-multimethod-matrix for Experiment 1 and Experiment 2.

Experiment 1 (serial mental addition and verification task)

		Session 1				
Session 2		**1**	**2**	**3**	**4**	**5**
RTM	*1*	0.93**	0.99**	0.05	0.89**	0.69**
RTMc	*2*	0.99**	0.94**	0.18	0.89**	0.70**
ER	*3*	−0.17	−0.01	0.84**	0.13	0.22
RTSD	*4*	0.88**	0.88**	−0.04	0.91**	0.93**
RTCV	*5*	0.59**	0.62**	−0.07	0.90**	0.91**

Experiment 2 (serial mental addition and comparison task)

		Session 1				
Session 2		**1**	**2**	**3**	**4**	**5**
RTM	1	0.94**	0.99**	−0.13	0.85**	0.63**
RTMc	2	0.99**	0.94**	−0.01	0.84**	0.62**
ER	3	−0.17	−0.08	0.78**	−0.11	−0.05
RTSD	4	0.88**	0.89**	−0.08	0.83**	0.93**
RTCV	5	0.59**	0.60**	0.05	0.89**	0.78**

Test–retest reliability and intercorrelation structure (convergent vs. divergent) of all performance measures, separately for session 1 and session 2. $N = 29$ (Exp. 1); $N = 50$ (Exp. 2); RTM, response time mean; ER, error percentage; RTSD, response time standard deviation; RTCV, response time coefficient of variation. Test–retest reliability is shown in the main diagonal (denoted with gray); correlations for the first session are shown above, for the second session below the main diagonal. $^{**}p < 0.01$.

either the left or the right key. That is, when the number value on the left side was larger (e.g., "2 + 3 | 4"), they had to respond with the left key, and when the number value on the right side was larger (e.g., "5 | 2 + 4"), they had to respond with the right key. Thus, the task required coordinated addition and comparison demands. The item set contained 148 items, with ties being excluded (cf. Blankenberger, 2001), and with a problem size ranging from 4 to 19. Both the large number of items (set-size effect) and the diversification of item difficulty (*"peek-a-boo"* uncertainty effect) are effective means to preventing mindless (rhythmic) responding (Lupker et al., 1997; Schmidt et al., 2016; Braem et al., 2019; Schumann et al., 2022), similar to the shuffling of preparatory intervals (Grosjean et al., 2001; Steinborn et al., 2009, 2010b; Langner et al., 2010, 2011, 2018; Wehrman and Sowman, 2019, 2021). In a session, each item was presented four times, amounting to a total of 592 randomly presented trials, requiring no longer than 30 min of testing.

Results and discussion

Population parameters and correlations are displayed in **Tables 2, 3**. Responses faster than 100 ms were discarded, correct reactions within this interval were used to compute RTM, RTSD, and RTCV. Incorrect reactions were regarded as error.

Retest reliability

Response time was again highly reliable ($r = 0.94$), and also good reliability was obtained for ER ($r = 0.79$) and RTSD ($r = 0.83$). Most important, both indices of performance variability appeared to have good reliability ($r = 0.83$ for RTSD; $r = 0.78$ for RTCV).

Correlational structure

Response time and ER were uncorrelated. Strong correlations were found between RTSD and RTM ($r > 0.85$ and $r = 0.88$), corroborating the redundancy of both measures. On the other hand, substantial positive correlations were also found between RTCV and RTM ($r > 0.63$ and $r = 0.59$), again indicating that even mean-corrected variability has some conceptual overlap with speed.

Practice effects

A multivariate effect was revealed overall and separately for all measures. A more detailed analysis revealed substantial practice gains for RTM [18% gain; $F(1,49) = 169.7$; $p < 0.000$; partial $\eta^2 = 0.78$], RTSD [32% gain; $F(1,49) = 41.6$; $p < 0.000$; partial $\eta^2 = 0.46$], and ER [27% gain; $F(1,49) = 9.9$; $p < 0.001$; partial $\eta^2 = 0.17$]; less pronounced gains were observed for RTCV [11% gain; $F(1,49) = 16.5$; $p < 0.000$; partial $\eta^2 = 0.25$]. Thus, participants became faster, more accurate, and more constant after retesting.

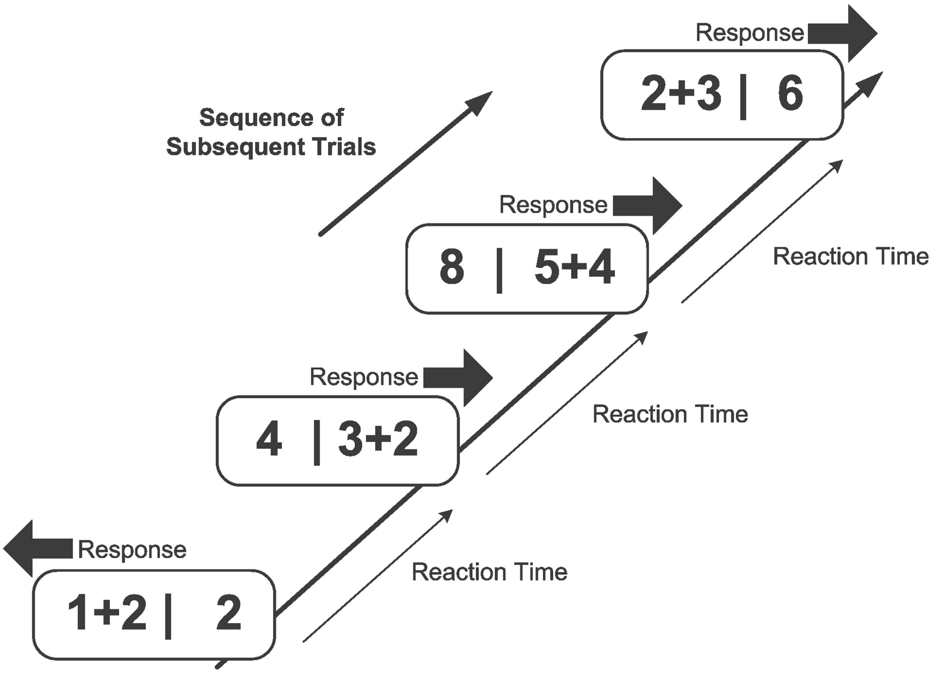

FIGURE 2
Example of a series of trials in the mental addition and comparison task. The task is to solve the addition term (e.g., "2 + 3 = . . . ?") and to decide whether the result is larger or smaller than the presented number value (e.g., "5 > 6 . . . ?"). The participants are to respond by "choosing" the larger of the number values. Each item in the task is presented immediately after the response to the previous item.

Extended analyses

We again and in the same way compared low- vs. high-performance participants. Reliability of RTM was largest for the low-performance group (r = 0.92) but was decreased for the high-performance group (r = 0.79), indicating again that even with many trials per experimental session (i.e., 529 trials in Exp. 1), RTM measures became prone to unreliability in slightly deficient subgroups (cf. **Figure 3**).

4. General discussion

The main results can be summarized as follows: (1) *Molecular precondition:* There was an approximately symmetric (sample-population) distribution (except skew in error rate), permitting to expect correlations between and across sessions. (2) *Molar precondition* (*reliability*): The relevant performance indices exhibited high test–retest reliability as obtained from the correlation of two testing sessions. This utility effect arguably stems from two sources, the time-compression property (Miller and Ulrich, 2003, 2013) and multitasking property (Vandierendonck et al., 2010; Vandierendonck, 2017). (3–4) *Convergence and discriminance*: Speed (RTM) and error rate (ER) were uncorrelated, both within and across testing sessions, indicating discriminant relationships. Response speed variability as indexed by RTCV was both reliable and relatively uncorrelated to other performance dimensions such as speed and accuracy, qualifying RTCV as discriminant from measures of speed and accuracy. By contrast, absolute RT variability (RTSD) was highly correlated with RTM, indicating redundancy. (5) *Stability:* Further, there were substantial practice gains for RTM and RTSD, and partly for ER; however, RTCV remained relatively stable after retesting – a finding that is consistent with the previous result of Flehmig et al. (2007). (6) *Reproducibility*. The correlational structure was relatively similar at the first relative to the second testing session, indicating stable relationships. (7) *Generalizability*. The overall picture of reliability, correlational structure and stability was similar in both tasks, enabling similar overall conclusions, and supporting the claim that speed tests based on multi-act mental arithmetic are promising in the experimental study of mental testing.

4.1. Reliability, inter-correlation, practice effects

Reliability

In order to construct a test with excellent psychometric characteristics, one has to consider three aspects, the *principles of measurement theory* (Lord and Novick, 1968; Lienert, 1969) the principles of *chronometric-design theory*

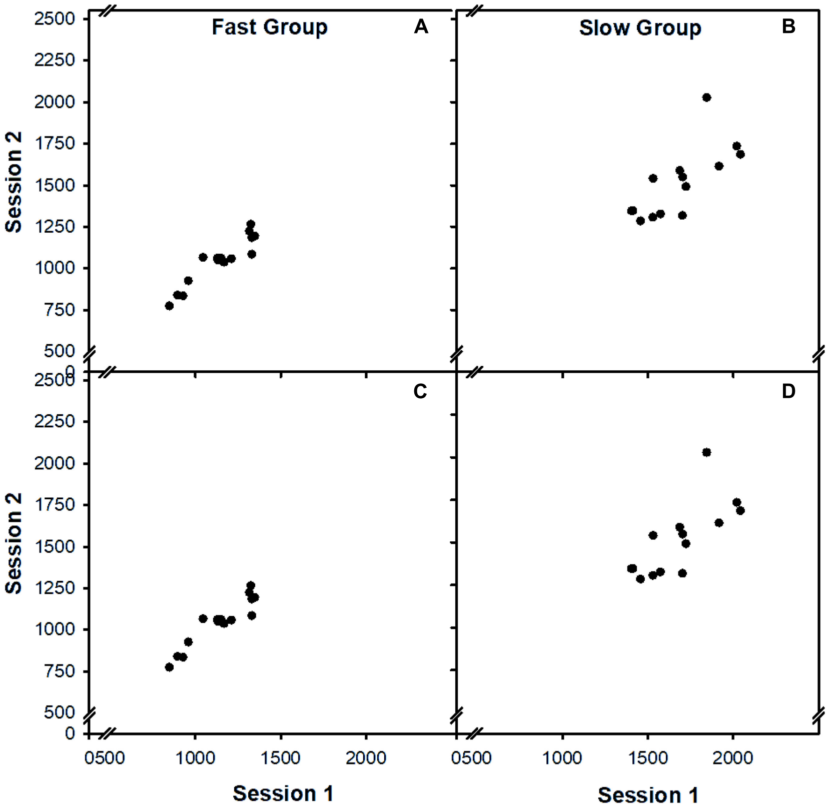

FIGURE 3
Results of Experiment 1 and Experiment 2. Scatterplot of the relationship between test and retest performance for both Experiment 1 (Panels **A,B**) and 2 (Panels **C,D**), separately for a group of fast individual and a group of slow individuals.

(Miller and Ulrich, 2013, 2021, 2022), and the principles of *concept-generalizability theory* (Cronbach and Meehl, 1955; Stroebe and Strack, 2014, pp. 61–63). On the side of measurement theory, the reliability of a test is determined by the amount of true score variance relative to error variance, and this ratio increases with number of trials. On the side of design theory, the amount of processing time should be maximal relative to testing time (Schumann et al., 2022, pp. 14–15). To this end, we employed a sufficient trial number in a self-paced presentation mode. As a result, exceptional test–retest reliability was obtained for the speed measure (r = 0.90). With respect to error rate (ER), retest reliability was surprisingly good (r = 0.79), given that error rate is typically low in self-paced tasks (∼5%), which limits reliability of error scores for pure mathematical reasons. However, low-event rate unreliability can partially be compensated by aggregating the absolute number of errors through lengthening a test (Hagemeister, 2007), that is, by trading off test economy with test reliability (Steinborn et al., 2016, 2018). Remarkable is that RTCV as relativized measure of variability exhibited good reliability in both tasks, given the many reports of a lack of reliability. How can these findings be accounted for? According to Miller and Ulrich (2013), reliability is predicted to increase with individual differences in person-specific processing time and individual differences evoked by task demands, but to decrease with increasing residual time and measurement error. Likely, exceptional reliability was obtained because mental arithmetic enables diversification of individual items: it is possible to use a great number of different individual items, preventing practice effects, and it allows for a finely graduated variegation of item difficulty (e.g., problem size ranging from 3–19), which mitigates mindlessness (Sanabria et al., 2011; Bedi et al., 2022) or other kinds of unfocused rhythmic responding (Steinborn and Langner, 2011, 2012; Schmidt, 2017; Braem et al., 2019).

Convergence and discriminance

Campbell and Fiske (1959) provided a practical methodology for purposes of construct validation and test construction. At its heart are two types of validity termed convergent and discriminant, defined as subcategories of construct validity. Convergent validity refers to the degree to which concepts that are related theoretically are actually interrelated empirically (Gleser et al., 1965; Rajaratnam et al., 1965). Discriminant validity refers to the degree to which theoretically distinct concepts are not interrelated empirically. To determine construct validity of a test's target measure is to demonstrate convergence and discrimination. Across experiments, speed and error rate were uncorrelated, indicating discriminant validity. RTCV was uncorrelated to error rate but somewhat related to speed, yet to a lesser degree than RTSD. Given the high reliability, this would clearly qualify RTCV over RTSD as the measure of choice to index performance variability. *One could ask, at this point, why a correlation of RTCV with RTM is still evident in a relativized (mean-corrected) measure like RTCV*? According to Flehmig et al. (2007), the answer is that while RTSD shares natural commonality with RTM for pure mathematical reasons, a relativized (percentage-based) measure captures performance variability beyond mere scaling–variability (Wagenmakers and Brown, 2007; Steinborn et al., 2016, 2018). A relationship between RTM and RTCV is still possible when an experimental factor evokes variability through propagation, or when variability is evoked along the trait (i.e., ability) dimension, for example, in a situation where lower ability is *in a specific way* related to more distraction during a task. A final note: we strongly advise against the method of removing shared variance of RTM with RTSD by means of partial correlation, as is sometimes seen in the literature, for two reasons. First, the obtained "mean-corrected" residual of a variability is "sample-dependent," that is, it cannot be interpreted individually (e.g., each individual added to a sample would change the score values for all other individuals of the sample). In general, we would not recommend indices computed from z-standardized scores, either as sum, subtractive, or residual score, or based on covariance-analytical techniques, as found in the literature (e.g., Seli et al., 2013; Klein and Robinson, 2019; Liesefeld and Janczyk, 2019), because they are unsuitable in practical assessment situation, or in clinical diagnostics (Willmes, 1985; Stuss, 2006; Vallesi and Shallice, 2007).

Practice effects

In theory, a test should be robust against practice effects. In reality, however, this requirement is unrealizable, since virtually every thinkable psychometric test will be affected by retesting, at least to some extent (Lemay et al., 2004; Calamia et al., 2013; Scharfen et al., 2018a,b,c; Soveri et al., 2018; Blotenberg and Schmidt-Atzert, 2019a,b; Williams et al., 2022). Crucial to the evaluation of a test is whether retesting changes the psychometric properties, which is relevant in situations where retesting is unavoidable, such as in the study of sleep deprivation, memory, or some sort of patient studies (Bratzke et al., 2009, 2012; Strobach et al., 2012; Mattiesing et al., 2017; Kunchulia et al., 2022). The results of the present two experiments demonstrate that despite retesting yielding substantial practice gains, the interrelation of the performance measures remained stable throughout. Overall, we can consider convergent and discriminant validity of the performance measures as being similar at both testing sessions, and across both experiments. Most interesting, the relativized measure of performance variability (RTCV) was to a much lesser degree affected by practice as compared to the measure of central tendency (RTM), which demonstrates that the relative performance fluctuations that occur during the test are stable with repeated testing. From a psychometric perspective, this is a remarkable feature of RTCV of assessing performance effects in psychometric speed tests (Pieters, 1983, 1985; Steinborn and Huestegge, 2016, 2017, 2020).

4.2. Multitasking measures in cognitive psychometrics

The study of multitasking is both multifarious and multitudinous given the multiple definitions and paradigms proposed in the literature. Depending on the particular objective, research can be classified into three categories, one concerned with the cognitive analysis of tasks (Pashler, 1994a,b; Meyer and Kieras, 1997a,b; Hommel, 1998a,b; Salvucci and Taatgen, 2008, 2011), another with the "human" factor at work and leisure (Wickens, 2008; Parasuraman and Manzey, 2010; Schumann et al., 2022) as well as in competitive sports activities (e.g., Kunde et al., 2011; Wehrman and Sowman, 2019, 2021; Pedraza-Ramirez et al., 2020; Polzien et al., 2022), and the third one with the utility of cognitive operations for psychometric test construction (Miller and Ulrich, 2003, 2013; Steinborn et al., 2016, 2018). By definition, multitasking can be conceived of as a special form of performance behavior that requires more than one mental act at a time or in close succession with the unit of observation being either a manifest or latent variable. For example, in a dual-response paradigm (Huestegge and Koch, 2013; Raettig and Huestegge, 2021), the performance registration is directly observable, while in a multi-act (chained) arithmetic paradigm, the units of interest lay hidden and can only indirectly be inferred (Pieters, 1983, 1985). Since even the simplest decision imposes a considerable demand requiring multiple inhibitory control of alternative response options (Moeller and Frings, 2019; Raettig and Huestegge, 2021), it is utterly impossible to reach a definition of multitasking that is minimalistic and universal at the same time (Ackerman, 1987; Logan, 2002, 2004; Altmann and Gray, 2008; Kiesel et al., 2010; Paas Oliveros et al., 2022; Schumann et al., 2022).

Düker-type vs. Bourdon-type tests

In the psychometric discipline (Jensen and Rohwer, 1966; Jensen, 1992; Miller and Ulrich, 2003, 2013), cognitive theory serves the practical purpose of constructing tests that meet standard psychometric criteria, and this division is thus more open to the general definition with respect to mental operations. According to Düker, capacity is *created through coordination* *of elemental acts*, which can be conceived of as the well-ordered orchestration of internal operations in close temporal proximity toward completing the ongoing task (Groen and Parkman, 1972; Ashcraft, 1992; Huber et al., 2016). The speeded arithmetic of the present study fall into this category of tasks. In the mental addition and comparison test (Exp. 1), for example, an addition problem is presented together with a single number, both spatially separated (e.g., "4 + 5 | 10"). Completing the task requires two elemental acts, solving the addition problem, and comparing the result with the number value of the presented digit (Restle, 1970; Steinborn et al., 2012). In this way, Düker-type tasks represent a natural form of an integrated-dual task, and by this means, provide opportunities of studying (sub-)task integration processes in goal-oriented multitasking (Restle, 1970; Treisman and Gelade, 1980; Huber et al., 2016; Moeller and Frings, 2019; Pedraza-Ramirez et al., 2020). With this regard, the use of culture techniques such as adding or subtracting numbers or multitasking combinations in gamified environments (Bilalic et al., 2009; Meyerhoff et al., 2017; Strobach and Huestegge, 2017; Pedraza-Ramirez et al., 2020) are superior as task medium for psychometric testing (Restle, 1970; Ashcraft, 1992; Huber et al., 2016), both in paper-pencil and computerized testing (Steinborn et al., 2016, 2018; Wuhr and Ansorge, 2020), in laboratory and real-assessment situations (Schumann et al., 2022).

Guidelines for constructing speed tests

According to Cronbach (1975), there are five reasons why the use of more complex speed tests based on mental addition are preferable: (a) test reliability, (b) test economy, (c) culture fairness, (d) flexibility, (e) and broadband validity (Thorndike, 1971; Jensen, 1980; Wuhr and Ansorge, 2020). In general terms, speed tests deliver a measure to assess ability as broadband concept, typically indicated by a wide spectrum of criterion validity. The basic principle is to employ a large number of trials and to administering the test in a self-paced mode so that the processing of items are compressed per unit of time, capturing maximal processing time relative to overall testing time (Miller and Ulrich, 2003, 2013). While 5–10 min of testing time seems optimal, it should not exceed 20–30 min. Although it is not exactly clear how complex a task should be to reach optimal reliabilities, medium task complexity (RT's approximately 1–2 s, 5–10% errors) seems to be a good option, as indicated by the present results and previous findings (cf. Schumann et al., 2022, for a review). Increasing complexity beyond some point (RTs > 3 s) is "increasingly" problematic as error rate will dramatically escalate, which produces ambiguity on the speed–error relation, for both (easy vs. hard) task condition and (slow vs. fast) individuals. Setting optimal levels and variegating (shuffling) difficulty are the most important control options for test development, typically achieved by varying arithmetic problem size (Ashcraft, 1992; Imbo and Vandierendonck, 2008a,b), arithmetic-chain length (Steinborn and Huestegge, 2016, 2017, 2020), or by using forms of multi-act mental arithmetic.

Indexing performance variability

Across two experiments, we observed both RTSD and RTCV as being sufficiently reliable with respect to retesting, indicating that this measure does not reflect random fluctuations but systematic variance that is replicable at a second testing session. Therefore, the correlations of measures of central tendency and variability may be interpreted without being concerned about insufficient reliability (Miller and Ulrich, 2013), as sometimes argued by a "reliability paradox" (Hedge et al., 2018). In fact, the correlations were around $r = 0.60$ at both first and second testing session and in both experiments (**Table 3**), indicating discriminant validity. Since the correlation of RT and RTSD were much higher, being between values of $r = 0.85–0.89$, the present results show that RTCV is the measures that should be preferred when one intends to measure performance variability in practical assessment contexts. This feature makes RTCV quite interesting for practical assessment purposes where test validity is at danger of being compromised by prior test experience (Hagemeister, 2007). If a measure is significantly affected by retesting, and an individual's performance level before practice cannot be determined, it becomes impossible to separate potential practice effects from the individual's ability, which the test was constructed to measure (Cronbach, 1975, p. 310–312). Failure to use appropriate control techniques would then lead to erroneous inferences about the aptitude of the individual being tested. Of course, further research is needed to examine whether invariance to practice effects is a general property of RTCV or only specific to a certain category of tasks. According to our findings, RTCV might be a potential candidate for characterizing additional aspects of performance in a simple and efficient way, in both basic research and applied contexts. However, before applying RTCV in practical assessment settings, additional research is required to elucidate the impact of task-specific factors (i.e., optimizing item difficulty, item-set size, task length, etc.) on the reliability of this performance measure using the MTMM approach.

4.3. Conclusion

Among the various assessment instruments available, speed tests based on cultural techniques (e.g., coding, sorting, arithmetic) typically exhibit the highest degree of test reliability (Neubauer and Knorr, 1998; Stahl and Rammsayer, 2007; Steinborn et al., 2010a; Wuhr and Ansorge, 2020). In terms of psychometric characteristics, speed tests stand any comparison to virtually all popular test batteries aiming to assess executive functions with typical testing times of 60–90 min (Zimmermann and Fimm, 1993; Fan et al., 2002; Westhoff and Graubner, 2003;

Habekost et al., 2014). Self-paced tests are also frequently used to experimentally induce a state termed ego depletion (Hagger et al., 2010; Inzlicht and Schmeichel, 2012; Krishna and Strack, 2017; Inzlicht et al., 2018; Vohs et al., 2021), albeit there are substantial methodological weaknesses of many reported studies in this area, relating to aspects like the use of arbitrary tasks, unsound performance measures, or an insufficient number of trials. Some studies even concluded that ego depletion is particularly measurable in the variability of performance rather than in RT mean score (Massar et al., 2018; Satterfield et al., 2018; Kamza et al., 2019; Unsworth and Robison, 2020). The reason might lie in the nature of depletion phenomena. Given that it is not the specific process but the control of attention, that is, the superordinate process of adaptively regulating mental resources, one would predict that performance becomes unstable rather than simply slow. The key contribution of the present psychometric analysis therefore covers at least three aspects, knowledge and step-by-step guidance in terms of how to correctly analyze and evaluate the psychometric properties of multitasking measures, methodology of design and research logic within the framework of mental chronometry using multi-act mental arithmetic, and proper (simple but robust) measurement technology. The central message of our report is that individual-differences in multitasking as assessed *via* correlations are not interpretable by themselves but must be evaluated in the light of their *molecular* and *molar* preconditions that involve, according to our analysis, at least 7 *formal steps of evaluation* (**Table 1**) to be worked through one after another.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

FS, MS, and HF: concept and method. JK, RL, and LH: supervision and discussion. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

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

Ackerman, P. L., and Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *J. Exp. Psychol. Appl.* 15, 163–181. doi: 10.1037/a0015719

Altmann, E. M., and Gray, W. D. (2008). An integrated model of cognitive control in task switching. *Psychol. Rev.* 115, 602–639. doi: 10.1037/0033-295X.115.3.602

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition* 44, 75–106. doi: 10.1016/0010-0277(92)90051-i

Ashcraft, M. H., and Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *J. Exp. Psychol.* 4, 527–538. doi: 10.1037/0278-7393.4.5.527

Bates, M. E., and Lemay, E. P. Jr. (2004). The d2 Test of attention: Construct validity and extensions in scoring techniques. *J. Int. Neuropsychol. Soc.* 10, 392–400. doi: 10.1017/S135561770410307X

Bedi, A., Russell, P. N., and Helton, W. S. (2022). Go-stimuli probability influences response bias in the sustained attention to response task: A signal detection theory perspective. *Psychol. Res.* [Epub ahead of print]. doi: 10.1007/s00426-022-01679-7

Behringer, J. (1987). *Experimental Runtime System (ERTS)*. Frankfurt: BeRi Soft Cooperation Publishers.

Bilalic, M., Smallbone, K., McLeod, P., and Gobet, F. (2009). Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains. *Proc. Biol. Sci.* 276, 1161–1165. doi: 10.1098/rspb.2008.1576

Blankenberger, S. (2001). The arithmetic tie effect is mainly encoding-based. *Cognition* 82, 15–24. doi: 10.1016/s0010-0277(01)00140-8

Blotenberg, I., and Schmidt-Atzert, L. (2019a). On the Locus of the Practice Effect in Sustained Attention Tests. *J. Intell.* 7:12. doi: 10.3390/jintelligence7020012

Blotenberg, I., and Schmidt-Atzert, L. (2019b). Towards a Process Model of Sustained Attention Tests. *J. Intell.* 7:3. doi: 10.3390/jintelligence7010003

Botvinick, M. M., and Bylsma, L. M. (2005a). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychon. Bull. Rev.* 12, 1011–1017. doi: 10.3758/bf03206436

Botvinick, M. M., and Bylsma, L. M. (2005b). Regularization in short-term memory for serial order. *J. Exp. Psychol.* 31, 351–358. doi: 10.1037/0278-7393.31.2.351

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., et al. (2019). Measuring Adaptive Control in Conflict Tasks. *Trends Cogn. Sci.* 23, 769–783. doi: 10.1016/j.tics.2019.07.002

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

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

Bruning, J., Koob, V., Manzey, D., and Janczyk, M. (2022). Serial and parallel processing in multitasking: Concepts and the impact of interindividual differences on task and stage levels. *J. Exp. Psychol. Hum. Percept. Perform.* 48, 724–742. doi: 10.1037/xhp0001008

Bruning, J., and Manzey, D. (2018). Flexibility of individual multitasking strategies in task-switching with preview: Are preferences for serial versus overlapping task processing dependent on between-task conflict? *Psychol. Res.* 82, 92–108. doi: 10.1007/s00426-017-0924-0

Bruning, J., Muckstein, M., and Manzey, D. (2020). Multitasking strategies make the difference: Separating processing-code resources boosts multitasking efficiency when individuals prefer to interleave tasks in free concurrent dual tasking. *J. Exp. Psychol. Hum. Percept. Perform* [Epub ahead of print]. doi: 10.1037/xhp0000865

Bruning, J., Reissland, J., and Manzey, D. (2021). Individual preferences for task coordination strategies in multitasking: Exploring the link between preferred modes of processing and strategies of response organization. *Psychol. Res.* 85, 577–591. doi: 10.1007/s00426-020-01291-7

Bruyer, R., and Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychol. Belgica* 51, 5–13.

Calamia, M., Markon, K., and Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *Clin. Neuropsychol.* 27, 1077–1105. doi: 10.1080/13854046.2013.809795

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Cheyne, J. A., Carriere, J. S., Solman, G. J., and Smilek, D. (2011). Challenge and error: Critical events and attention-related errors. *Cognition* 121, 437–446. doi: 10.1016/j.cognition.2011.07.010

Compton, B. J., and Logan, G. D. (1991). The transition from algorithm to retrieval in memory-based theories of automaticity. *Mem. Cognit.* 19, 151–158. doi: 10.3758/bf03197111

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

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

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Crowe, E. M., and Kent, C. (2019). Evidence for short-term, but not long-term, transfer effects in the temporal preparation of auditory stimuli. *Q. J. Exp. Psychol.* 72, 2672–2679. doi: 10.1177/1747021819854044

Deutsch, R., Gawronski, B., and Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *J. Pers. Soc. Psychol.* 91, 385–405. doi: 10.1037/0022-3514.91.3.385

Deutsch, R., Kordts-Freudinger, R., Gawronski, B., and Strack, F. (2009). Fast and fragile: A new look at the automaticity of negation processing. *Exp. Psychol.* 56, 434–446. doi: 10.1027/1618-3169.56.6.434

Dignath, D., Eder, A. B., Steinhauser, M., and Kiesel, A. (2020). Conflict monitoring and the affective-signaling hypothesis-An integrative review. *Psychon. Bull. Rev.* 27, 193–216. doi: 10.3758/s13423-019-01668-9

Düker, H. (1949). Über ein Verfahren zur Untersuchung der geistigen Leistungsfähigkeit [A test for investigating general cognitive ability]. *Psychol. Res.* 23, 10–24.

Düker, H., and Lienert, G. A. (1949). *Konzentrations-Leistungs-Test (KLT).* Göttingen: Hogrefe.

Dunlap, W. P., Chen, R. S., and Greer, T. (1994). Skew reduces test-retest reliability. *J. Appl. Psychol.* 79, 310–313. doi: 10.1037/0021-9010.79.2.310

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., and Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *J. Cogn. Neurosci.* 14, 340–347. doi: 10.1162/089892902317361886

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

Fortenbaugh, F. C., DeGutis, J., and Esterman, M. (2017). Recent theoretical, neural, and clinical advances in sustained attention research. *Ann. N. Y. Acad. Sci.* 1396, 70–91. doi: 10.1111/nyas.13318

Gleser, G. C., Cronbach, L. J., and Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika* 30, 395–418. doi: 10.1007/BF02289531

Goodwin, G. P., and Johnson-Laird, P. N. (2011). Mental models of Boolean concepts. *Cogn. Psychol.* 63, 34–59. doi: 10.1016/j.cogpsych.2011.04.001

Goodwin, G. P., and Johnson-Laird, P. N. (2013). The acquisition of Boolean concepts. *Trends. Cogn. Sci.* 17, 128–133. doi: 10.1016/j.tics.2013.01.007

Gopher, D. (1996). Attention control: Explorations of the work of an executive controller. *Brain Res. Cogn. Brain. Res.* 5, 23–38. doi: 10.1016/s0926-6410(96)00038-9

Greenwald, A. G. (1970). Sensory feedback mechanisms in performance control: With special reference to the ideo-motor mechanism. *Psychol. Rev.* 77, 73–99. doi: 10.1037/h0028689

Greenwald, A. G. (1972). On doing two things at once: Time sharing as a function of ideomotor compatibility. *J. Exp. Psychol.* 94, 52–57. doi: 10.1037/h0032762

Greer, T., Dunlap, W. P., Hunter, S. T., and Berman, M. E. (2006). Skew and internal consistency. *J. Appl. Psychol.* 91, 1351–1358. doi: 10.1037/0021-9010.91.6.1351

Groen, G. J., and Parkman, J. M. (1972). A chronometric analysis of simple addition. *Psychol. Rev.* 79, 329–343. doi: 10.1037/h0032950

Gropel, P., Baumeister, R. F., and Beckmann, J. (2014). Action versus state orientation and self-control performance after depletion. *Pers. Soc. Psychol. Bull.* 40, 476–487. doi: 10.1177/0146167213516636

Grosjean, M., Rosenbaum, D. A., and Elsinger, C. (2001). Timing and reaction time. *J. Exp. Psychol.* 130, 256–272. doi: 10.1037//0096-3445.130.2.256

Habekost, T., Petersen, A., and Vangkilde, S. (2014). Testing attention: Comparing the ANT with TVA-based assessment. *Behav. Res. Methods* 46, 81–94. doi: 10.3758/s13428-013-0341-2

Hagemeister, C. (2007). How Useful is the Power Law of Practice for Recognizing Practice in Concentration Tests? *Eur. J. Psychol. Assess.* 23, 157–165. doi: 10.1027/1015-5759.23.3.157

Hagemeister, C., and Kronmaier, M. (2017). Alcohol consumption and cycling in contrast to driving. *Accid. Anal. Prev.* 105, 102–108. doi: 10.1016/j.aap.2017.01.001

Hagger, M. S., Wood, C., Stiff, C., and Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychol. Bull.* 136, 495–525. doi: 10.1037/a0019486

Han, T., and Proctor, R. W. (2022a). Change of variable-foreperiod effects within an experiment: A Bayesian modeling approach. *J. Cogn.* 5:40. doi: 10.5334/joc.235

Han, T., and Proctor, R. W. (2022b). Effects of a neutral warning signal on spatial two-choice reactions. *Q. J. Exp. Psychol.* 75, 754–764. doi: 10.1177/17470218211037604

Han, T., and Proctor, R. W. (2022c). Revisiting variable-foreperiod effects: Evaluating the repetition priming account. *Atten. Percept. Psychophys.* 84, 1193–1207. doi: 10.3758/s13414-022-02476-5

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186. doi: 10.3758/s13428-017-0935-1

Herbort, O., and Rosenbaum, D. A. (2014). What is chosen first, the hand used for reaching or the target that is reached? *Psychon. Bull. Rev.* 21, 170–177. doi: 10.3758/s13423-013-0488-y

Hommel, B. (1998a). Automatic stimulus-response translation in dual-task performance. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1368–1384. doi: 10.1037//0096-1523.24.5.1368

Hommel, B. (1998b). Event Files: Evidence for automatic integration of stimulus-response episodes. *Visual Cogn.* 5, 183–216. doi: 10.1080/713756773

Huber, S., Nuerk, H. C., Willmes, K., and Moeller, K. (2016). A general model framework for multisymbol number comparison. *Psychol. Rev.* 123, 667–695. doi: 10.1037/rev0000040

Huestegge, L., and Koch, I. (2013). Constraints in task-set control: Modality dominance patterns among effector systems. *J. Exp. Psychol. Gen.* 142, 633–637. doi: 10.1037/a0030156

Huestegge, L., Pieczykolan, A., and Koch, I. (2014). Talking while looking: On the encapsulation of output system representations. *Cogn. Psychol.* 73, 72–91. doi: 10.1016/j.cogpsych.2014.06.001

Humphreys, M. S., and Revelle, W. (1984). Personality, motivation, and performance - A theory of the relationship between individual-differences and information-processing. *Psychol. Rev.* 91, 153–184. doi: 10.1037/0033-295x.91.2.153

Imbo, I., and Vandierendonck, A. (2008a). Effects of problem size, operation, and working-memory span on simple-arithmetic strategies: Differences between children and adults? *Psychol. Res.* 72, 331–346. doi: 10.1007/s00426-007-0112-8

Imbo, I., and Vandierendonck, A. (2008b). Practice effects on strategy selection and strategy efficiency in simple mental arithmetic. *Psychol. Res.* 72, 528–541. doi: 10.1007/s00426-007-0128-0

Inzlicht, M., and Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspect. Psychol. Sci.* 7, 450–463. doi: 10.1177/1745691612454134

Inzlicht, M., Shenhav, A., and Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends Cogn. Sci.* 22, 337–349. doi: 10.1016/j.tics.2018.01.007

Janczyk, M., and Kunde, W. (2010). Stimulus-response bindings contribute to item switch costs in working memory. *Psychol. Res.* 74, 370–377. doi: 10.1007/s00426-009-0259-6

Janczyk, M., and Kunde, W. (2020). Dual tasking from a goal perspective. *Psychol. Rev.* 127, 1079–1096. doi: 10.1037/rev0000222

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Pers. Individ. Differ.* 13, 869–881. doi: 10.1016/0191-8869(92)90004-9

Jensen, A. R., and Rohwer, W. D. Jr. (1966). The Stroop color-word test: A review. *Acta Psychol.* 25, 36–93. doi: 10.1016/0001-6918(66)90004-7

Jentzsch, I., and Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *Q. J. Exp. Psychol.* 62, 209–218. doi: 10.1080/17470210802240655

Jentzsch, I., and Leuthold, H. (2006). Control over speeded actions: A common processing locus for micro- and macro-trade-offs? *Q. J. Exp. Psychol.* 59, 1329–1337. doi: 10.1080/17470210600674394

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A flaw in human judgement*. London, UK: HarperCollins.

Kamza, A., Molinska, M., Skrzypska, N., and Dlugiewicz, P. (2019). Can sustained attention adapt to prior cognitive effort? An evidence from experimental study. *Acta. Psychol.* 192, 181–193. doi: 10.1016/j.actpsy.2018.11.007

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., et al. (2010). Control and interference in task switching - A review. *Psychol. Bull.* 136, 849–874. doi: 10.1037/a0019842

Klein, R. J., and Robinson, M. D. (2019). Neuroticism as mental noise: Evidence from a continuous tracking task. *J. Pers.* 87, 1221–1233. doi: 10.1111/jopy.12469

Krishna, A., and Strack, F. (2017). "Reflection and impulse as determinants of human behavior," in *Knowledge and Action, Knowledge and Space*, eds P. Meusburger, B. Merlen, and L. Suarsana (Berlin: Springer), 145–167.

Kunchulia, M., Parkosadze, K., Lomidze, N., Tatishvili, T., and Thomaschke, R. (2022). Children with developmental dyslexia show an increased variable foreperiod effect. *J. Cogn. Psychol.* 34, 563–574. doi: 10.1080/20445911.2022.2060989

Kunde, W., Skirde, S., and Weigelt, M. (2011). Trust my face: Cognitive factors of head fakes in sports. *J. Exp. Psychol.* 17, 110–127. doi: 10.1037/a0023756

Langner, R., and Eickhoff, S. B. (2013). Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychol. Bull.* 139, 870–900. doi: 10.1037/a0030694

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

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

Langner, R., Steinborn, M. B., Eickhoff, S. B., and Huestegge, L. (2018). When specific action biases meet nonspecific preparation: Event repetition modulates the variable-foreperiod effect. *J. Exp. Psychol.* 44, 1313–1323. doi: 10.1037/xhp0000561

Lemay, S., Bedard, M. A., Rouleau, I., and Tremblay, P. L. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *Clin. Neuropsychol.* 18, 284–302. doi: 10.1080/13854040490501718

Leth-Steensen, C., Elbaz, Z. K., and Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychol.* 104, 167–190. doi: 10.1016/s0001-6918(00)00019-6

Lienert, G. A. (1969). *Testaufbau und Testanalyse [test construction and test analysis]*, 1. Edn. Weinheim: Beltz.

Liesefeld, H. R., and Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behav. Res. Methods* 51, 40–60. doi: 10.3758/s13428-018-1076-x

Logan, G. D. (1979). On the use of a concurrent memory load to measure attention and automaticity. *J. Exp. Psychol. Hum. Percept. Perform.* 5, 189–207. doi: 10.1037/0096-1523.5.2.189

Logan, G. D. (1988). Toward and instance theory of automatization. *Psychol. Rev.* 95, 492–527. doi: 10.1037//0033-295x.95.4.492

Logan, G. D. (2002). An instance theory of attention and memory. *Psychol. Rev.* 109, 376–400. doi: 10.1037//0033-295x.109.2.376

Logan, G. D. (2004). Cumulative progress in formal theories of attention. *Annu. Rev. Psychol.* 55, 207–234. doi: 10.1146/annurev.psych.55.090902.141415

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

Los, S. A., Kruijne, W., and Meeter, M. (2014). Outlines of a multiple trace theory of temporal preparation. *Front. Psychol.* 5:1058. doi: 10.3389/fpsyg.2014.01058

Los, S. A., Kruijne, W., and Meeter, M. (2017). Hazard versus history: Temporal preparation is driven by past experience. *J. Exp. Psychol. Hum. Percept. Perform.* 43, 78–88. doi: 10.1037/xhp0000279

Los, S. A., Nieuwenstein, J., Bouharab, A., Stephens, D. J., Meeter, M., and Kruijne, W. (2021). The warning stimulus as retrieval cue: The role of associative memory in temporal preparation. *Cogn. Psychol.* 125:101378. doi: 10.1016/j.cogpsych.2021.101378

Lupker, S. J., Brown, P., and Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 570–590. doi: 10.1037//0278-7393.23.3.570

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

Manstead, A. S. R., and Semin, G. R. (1980). Social facilitation effects: Mere enhancement of dominant response? *Br. J. Soc. Clin. Psychol.* 19, 119–136. doi: 10.1111/j.2044-8260.1980.tb00937.x

Massar, S. A. A., Sasmita, K., Lim, J., and Chee, M. W. L. (2018). Motivation alters implicit temporal attention through sustained and transient mechanisms: A behavioral and pupillometric study. *Psychophysiology* 55:e13275. doi: 10.1111/psyp.13275

Mattiesing, R. M., Kruijne, W., Meeter, M., and Los, S. A. (2017). Timing a week later: The role of long-term memory in temporal preparation. *Psychon. Bull. Rev.* 24, 1900–1905. doi: 10.3758/s13423-017-1270-3

Meyer, D. E., and Kieras, D. E. (1997a). A computational theory of executive cognitive processes and multiple-task performance: 1. Basic mechanisms. *Psychol. Rev.* 104, 3–65. doi: 10.1037//0033-295x.104.1.3

Meyer, D. E., and Kieras, D. E. (1997b). A computational theory of executive cognitive processes and multiple-task performance: 2. Accounts of psychological refractory-period phenomena. *Psychol. Rev.* 104, 749–791. doi: 10.1037/0033-295x.104.4.749

Meyerhoff, H. S., Papenmeier, F., and Huff, M. (2017). Studying visual attention using the multiple object tracking paradigm: A tutorial review. *Atten. Percept. Psychophys.* 79, 1255–1274. doi: 10.3758/s13414-017-1338-1

Miller, J., and Ulrich, R. (2003). Simple reaction time and statistical facilitation: A parallel grains model. *Cogn. Psychol.* 46, 101–151. doi: 10.1016/s0010-0285(02)00517-0

Miller, J., and Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychon. Bull. Rev.* 20, 819–858. doi: 10.3758/s13423-013-0404-5

Miller, J., and Ulrich, R. (2021). A simple, general, and efficient method for sequential hypothesis testing: The independent segments procedure. *Psychol. Methods* 26, 486–497. doi: 10.1037/met0000350

Miller, J., and Ulrich, R. (2022). Optimizing Research Output: How Can Psychological Research Methods Be Improved? *Annu. Rev. Psychol.* 73, 691–718. doi: 10.1146/annurev-psych-020821-094927

Moeller, B., and Frings, C. (2019). From simple to complex actions: Response-response bindings as a new approach to action sequences. *J. Exp. Psychol. General* 148, 174–183. doi: 10.1037/xge0000483

Navon, D., and Gopher, D. (1979). On the economy of the human-processing system. *Psychol. Rev.* 86, 214–255. doi: 10.1037/0033-295x.86.3.214

Neubauer, A. C., and Fink, A. (2009). Intelligence and neural efficiency. *Neurosci. Biobehav. Rev.* 33, 1004–1023. doi: 10.1016/j.neubiorev.2009.04.001

Neubauer, A. C., and Knorr, E. (1998). Three paper-and-pencil tests for speed of information processing: Psychometric properties and correlations with intelligence. *Intelligence* 26, 123–151. doi: 10.1016/s0160-2896(99)80058-0

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

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *J. Exp. Psychol. Learn Mem. Cogn.* 28, 411–421. doi: 10.1037//0278-7393.28.3.411

Oberauer, K. (2003). Selective attention to elements in working memory. *Exp. Psychol.* 50, 257–269. doi: 10.1026//1618-3169.50.4.257

Paas Oliveros, L. K., Pieczykolan, A., Plaschke, R. N., Eickhoff, S. B., and Langner, R. (2022). Response-code conflict in dual-task interference and its modulation by age. *Psychol. Res.* [Epub ahead of print]. doi: 10.1007/s00426-021-01639-7

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Hum. Fact.* 52, 381–410. doi: 10.1177/0018720810376055

Pashler, H. (1994a). Dual-task interference in simple tasks - Data and theory. *Psychol. Bull.* 116, 220–244. doi: 10.1037/0033-2909.116.2.220

Pashler, H. (1994b). Overlapping mental operations in serial performance with preview. *Q. J. Exp. Psychol. A.* 47, 161–191; discussion193–169; 201–165. doi: 10.1080/14640749408401148

Pashler, H., and Baylis, G. C. (1991a). Procedural learning: I. Locus of practice effects in speeded choice tasks. *J. Exp. Psychol. Learn. Mem. Cogn.* 17, 20–32. doi: 10.1037/0278-7393.17.1.20

Pashler, H., and Baylis, G. C. (1991b). Procedural learning: II. Intertrial repetition effects in speeded-choice tasks. *J. Exp. Psychol. Learn. Mem. Cogn.* 17, 33–48. doi: 10.1037/0278-7393.17.1.33

Pedraza-Ramirez, I., Musculus, L., Raab, M., and Laborde, S. (2020). Setting the scientific stage for esports psychology: A systematic review. *Int. Rev. Sport Exerc. Psychol.* 13, 319–352. doi: 10.1080/1750984X.2020.1723122

Phillips, L. H., and Rabbitt, P. M. A. (1995). Impulsivity and speed-accuracy strategies in intelligence-test performance. *Intelligence* 21, 13–29. doi: 10.1016/0160-2896(95)90036-5

Pieters, J. P. M. (1983). Sternberg additive factor method and underlying psychological processes - Some theoretical considerations. *Psychol. Bull.* 93, 411–426. doi: 10.1037/0033-2909.93.3.411

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

Polzien, A., Guldenpenning, I., and Weigelt, M. (2022). Repeating head fakes in basketball: Temporal aspects affect the congruency sequence effect and the size of the head-fake effect. *J. Exp. Psychol.* [Epub ahead of print]. doi: 10.1037/xap0000419

Rabbitt, P. M. A., and Banerji, N. (1989). How does very prolonged practice improve decision speed? *J. Exp. Psychol. Gen.* 118, 338–345. doi: 10.1037/0096-3445.118.4.338

Raettig, T., and Huestegge, L. (2021). Representing action in terms of what not to do: Evidence for inhibitory coding during multiple action control. *J. Exp. Psychol. Hum. Percept. Perform.* 47, 1253–1273. doi: 10.1037/xhp0000943

Rajaratnam, N., Cronbach, L. J., and Gleser, G. C. (1965). Generalizability of Stratified-Parallel Tests. *Psychometrika* 30, 39–56. doi: 10.1007/BF02289746

Restle, F. (1970). Speed of adding and comparing numbers. *J. Exp. Psychol.* 83, 274–278. doi: 10.1037/h0028573

Rickard, T. C. (2005). Revised identical elements model of arithmetic fact representation. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 250–257. doi: 10.1037/0278-7393.31.2.250

Salvucci, D. D., and Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychol. Rev.* 115, 101–130. doi: 10.1037/0033-295x.115.1.101

Salvucci, D. D., and Taatgen, N. A. (2011). Toward a unified view of cognitive control. *Top Cogn. Sci.* 3, 227–230. doi: 10.1111/j.1756-8765.2011.01134.x

Sanabria, D., Capizzi, M., and Correa, A. (2011). Rhythms that speed you up. *J. Exp. Psychol.* 37, 236–244. doi: 10.1037/a0019956

Satterfield, K., Harwood, A. E., Helton, W. S., and Shaw, T. H. (2018). Does depleting self-control result in poorer vigilance performance? *Hum. Fact.* 61:0018720818806151. doi: 10.1177/0018720818806151

Schaaf, M., Kunde, W., and Wirth, R. (2022). Evidence for initially independent monitoring of responses and response effects. *J. Exp. Psychol. Hum. Percept. Perform.* 48, 128–138. doi: 10.1037/xhp0000979

Scharfen, J., Blum, D., and Holling, H. (2018a). Response time reduction due to retesting in mental speed tests: A meta-analysis. *J. Intell.* 6:6. doi: 10.3390/jintelligence6010006

Scharfen, J., Jansen, K., and Holling, H. (2018b). Retest effects in working memory capacity tests: A meta-analysis. *Psychon. Bull. Rev.* 25, 2175–2199. doi: 10.3758/s13423-018-1461-6

Scharfen, J., Peters, J. M., and Holling, H. (2018c). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence* 67, 44–66.

Schmidt, J. R. (2017). Time-out for conflict monitoring theory: Preventing rhythmic biases eliminates the list-level proportion congruent effect. *Can. J. Exp. Psychol.* 71, 52–62. doi: 10.1037/cep0000106

Schmidt, J. R., De Houwer, J., and Rothermund, K. (2016). The Parallel Episodic Processing (PEP) model 2.0: A single computational model of stimulus-response binding, contingency learning, power curves, and mixing costs. *Cogn. Psychol.* 91, 82–108. doi: 10.1016/j.cogpsych.2016.10.004

Schumann, F., Steinborn, M. B., Kürten, J., Cao, L., Händel, B., and Huestegge, L. (2022). Restoration of attention by rest in a multitasking world: Theory, methodology, and empirical evidence. *Front. Psychol.* 13:867978. doi: 10.3389/fpsyg.2022.867978

Seli, P., Jonker, T. R., Solman, G. J., Cheyne, J. A., and Smilek, D. (2013). A methodological note on evaluating performance in a sustained-attention-to-response task. *Behav. Res. Methods* 45, 355–363. doi: 10.3758/s13428-012-0266-1

Soveri, A., Lehtonen, M., Karlsson, L. C., Lukasik, K., Antfolk, J., and Laine, M. (2018). Test–retest reliability of five frequently used executive tasks in healthy adults. *Appl. Neuropsychol. Adult* 25, 155–165. doi: 10.1080/23279095.2016.1263795

Stahl, J., and Rammsayer, T. H. (2007). Identification of sensorimotor components accounting for individual variability in Zahlen–Verbindungs-Test (ZVT) performance. *Intelligence* 35, 623–630. doi: 10.1016/j.intell.2006.12.001

Steinborn, M. B., Flehmig, H. C., Bratzke, D., and Schroter, H. (2012). Error reactivity in self-paced performance: Highly-accurate individuals exhibit largest post-error slowing. *Q. J. Exp. Psychol.* 65, 624–631. doi: 10.1080/17470218.2012.660962

Steinborn, M. B., Flehmig, H. C., Westhoff, K., and Langner, R. (2010a). Differential effects of prolonged work on performance measures in self-paced speed tests. *Adv. Cogn. Psychol.* 5, 105–113. doi: 10.2478/v10053-008-0070-8

Steinborn, M. B., Rolke, B., Bratzke, D., and Ulrich, R. (2010b). The effect of a cross-trial shift of auditory warning signals on the sequential foreperiod effect. *Acta Psychol.* 134, 94–104. doi: 10.1016/j.actpsy.2009.12.011

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

Steinborn, M. B., and Huestegge, L. (2017). Phone conversation while processing information: Chronometric analysis of load effects in everyday-media multitasking. *Front. Psychol.* 8:896. doi: 10.3389/fpsyg.2017.00896

Steinborn, M. B., and Huestegge, L. (2020). Socially alerted cognition evoked by a confederate's mere presence: Analysis of reaction-time distributions and delta plots. *Psychol. Res.* 84, 1424–1439. doi: 10.1007/s00426-019-01143-z

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

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

Steinborn, M. B., Langner, R., Flehmig, H. C., and Huestegge, L. (2016). Everyday life cognitive instability predicts simple reaction time variability:

Analysis of reaction time distributions and delta plots. *Appl. Cogn. Psychol.* 30, 92–102. doi: 10.1002/acp.3172

Steinborn, M. B., Langner, R., Flehmig, H. C., and Huestegge, L. (2018). Methodology of performance scoring in the d2 sustained-attention test: Cumulative-reliability functions and practical guidelines. *Psychol. Assess.* 30, 339–357. doi: 10.1037/pas0000482

Steinborn, M. B., Langner, R., and Huestegge, L. (2017). Mobilizing cognition for speeded action: Try-harder instructions promote motivated readiness in the constant-foreperiod paradigm. *Psychol. Res.* 81, 1135–1151. doi: 10.1007/s00426-016-0810-1

Steinborn, M. B., Rolke, B., Bratzke, D., and Ulrich, R. (2009). Dynamic adjustment of temporal preparation: Shifting warning signal modality attenuates the sequential foreperiod effect. *Acta Psychol.* 132, 40–47. doi: 10.1016/j.actpsy.2009.06.002

Steinhauser, M., Ernst, B., and Ibald, K. W. (2017). Isolating component processes of posterror slowing with the psychological refractory period paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* 43, 653–659. doi: 10.1037/xlm0000329

Strack, F., and Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Pers. Soc. Psychol. Rev.* 8, 220–247. doi: 10.1207/s15327957pspr0803_1

Strobach, T., and Huestegge, L. (2017). Evaluating the effectiveness of commercial brain game training with working-memory tasks. *J. Cogn. Enhance.* 1, 539–558. doi: 10.1007/s41465-017-0053-0

Strobach, T., Liepelt, R., Schubert, T., and Kiesel, A. (2012). Task switching: Effects of practice on switch and mixing costs. *Psychol. Res.* 76, 74–83. doi: 10.1007/s00426-011-0323-x

Stroebe, W., and Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9, 59–71. doi: 10.1177/1745691613514450

Stuss, D. T. (2006). Frontal lobes and attention: Processes and networks, fractionation and integration. *J. Int. Neuropsychol. Soc.* 12, 261–271. doi: 10.1017/S1355617706060358

Stuss, D. T., Alexander, M. P., Shallice, T., Picton, T. W., Binns, M. A., Macdonald, R., et al. (2005). Multiple frontal systems controlling response speed. *Neuropsychologia* 43, 396–417. doi: 10.1016/j.neuropsychologia.2004.06.010

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

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

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

Thomaschke, R., and Dreisbach, G. (2015). The time-event correlation effect is due to temporal expectancy, not to partial transition costs. *J. Exp. Psychol. Hum. Percept. Perform.* 41, 196–218. doi: 10.1037/a0038328

Thomaschke, R., Hopkins, B., and Miall, R. C. (2012). The planning and control model (PCM) of motorvisual priming: Reconciling motorvisual impairment and facilitation effects. *Psychol. Rev.* 119, 388–407. doi: 10.1037/a0027453

Thorndike, E. L. (1971). Concepts of culture fairness. *J. Educ. Measure.* 8, 63–70.

Thorne, D. R. (2006). Throughput: A simple performance index with desirable characteristics. *Behav. Res. Methods* 38, 569–573. doi: 10.3758/bf03193886

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Unsworth, N. (2015). Consistency of attentional control as an important cognitive trait: A latent variable analysis. *Intelligence* 49, 110–128. doi: 10.1016/j.intell.2015.01.005

Unsworth, N., and Robison, M. K. (2020). Working memory capacity and sustained attention: A cognitive-energetic perspective. *J. Exp. Psychol. Learn. Mem. Cogn.* 46, 77–103. doi: 10.1037/xlm0000712

Vallesi, A., Arbula, S., and Bernardis, P. (2014). Functional dissociations in temporal preparation: Evidence from dual-task performance. *Cognition* 130, 141–151. doi: 10.1016/j.cognition.2013.10.006

Vallesi, A., and Shallice, T. (2007). Developmental dissociations of preparation over time: Deconstructing the variable foreperiod phenomena. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 1377–1388. doi: 10.1037/0096-1523.33.6.1377

Van Breukelen, G. J., Roskam, E. E., Eling, P. A., Jansen, R. W., Souren, D. A., and Ickenroth, J. G. (1995). A model and diagnostic measures for response time series on tests of concentration: Historical background, conceptual framework, and some applications. *Brain Cogn.* 27, 147–179. doi: 10.1006/brcg.1995.1015

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behav. Res. Methods* 49, 653–673. doi: 10.3758/s13428-016-0721-5

Vandierendonck, A., Liefooghe, B., and Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychol. Bull.* 136, 601–626. doi: 10.1037/a0019791

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., et al. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychol. Sci.* 32, 1566–1581. doi: 10.1177/0956797621989733

Wagenmakers, E. J., and Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychol. Rev.* 114, 830–841. doi: 10.1037/0033-295X.114.3.830

Watson, J. M., and Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychon. Bull. Rev.* 17, 479–485. doi: 10.3758/PBR.17.4.479

Wehrman, J. J., and Sowman, P. (2019). Time in the motor cortex: Motor evoked potentials track foreperiod duration without concurrent movement. *Neurosci. Lett.* 698, 85–89. doi: 10.1016/j.neulet.2019.01.012

Wehrman, J. J., and Sowman, P. (2021). Oddball onset timing: Little evidence of early gating of oddball stimuli from tapping, reacting, and producing. *Atten. Percept. Psychophys.* 83, 2291–2302. doi: 10.3758/s13414-021-02257-6

Wessel, J. R. (2018). An adaptive theory of error processing. *Psychophysiology* 55:e13041. doi: 10.1111/psyp.13041

Westhoff, K., and Graubner, J. (2003). Konstruktion eines Komplexen Konzentrationstests [construction of a complex concentration test battery]. *Diagnostica* 49, 110–119. doi: 10.1026//0012-1924.49.3.110

Wickens, C. D. (2008). Multiple resources and mental workload. *Hum. Fact.* 50, 449–455. doi: 10.1518/001872008x288394

Williams, D. R., Martin, S. R., and Rast, P. (2022). Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. *Behav Res Methods* 54, 1272–1290. doi: 10.3758/s13428-021-01646-x

Willmes, K. (1985). An approach to analyzing a single subject's scores obtained in a standardized test with application to the Aachen Aphasia Test (AAT). *J. Clin. Exp. Neuropsychol.* 7, 331–352. doi: 10.1080/01688638508401268

Wuhr, P., and Ansorge, U. (2020). Do left-handers outperform right-handers in paper-and-pencil tests of attention? *Psychol. Res.* 84, 2262–2272. doi: 10.1007/s00426-019-01224-z

Zbrodoff, N. J., and Logan, G. D. (1986). On the autonomy of mental processes: A case study of arithmetic. *J. Exp. Psychol. Gen.* 115, 118–130. doi: 10.1037//0096-3445.115.2.118

Zbrodoff, N. J., and Logan, G. D. (1990). On the relation between production and verification tasks in the psychology of simple arithmetic. *J. Exp. Psychol.* 16, 83–97. doi: 10.1037/0278-7393.16.1.83

Zimmermann, P., and Fimm, B. (1993). *Testbatterie zur Aufmerksamkeitsprüfung TAP [test battery of attention functions]*. Herzogenrath, Germany: Psytest.