pedocs
Open Access Erziehungswissenschaften

DIPF

Goldhammer, Frank; Kröhne, Ulf

Controlling individuals' time spent on task in speeded performance measures. Experimental time limits, posterior time limits, and response time modeling

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Applied Psychological Measurement 38 (2014) 4, S. 255-267

Bitte verwenden Sie beim Zitieren folgende URN /
Please use the following URN for citation:
urn:nbn:de:0111-pedocs-127839

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

Article

Applied Psychological Measurement
2014, Vol. 38(4) 255–267
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621613517164
apm.sagepub.com
SAGE

Controlling Individuals' Time Spent on Task in Speeded Performance Measures: Experimental Time Limits, Posterior Time Limits, and Response Time Modeling

Frank Goldhammer[1,2] and Ulf Kroehne[1]

Abstract

The speed-ability trade-off becomes a measurement problem if there is between-subject variation in the speed-ability compromise, as this may affect the comparability of ability estimates. To control individual speed differences, the response-signal (RS) paradigm was applied requiring an immediate response as soon as an acoustic signal is presented. A figural discrimination task and a word recognition task were completed both in an untimed condition allowing individual differences in time spent on task and in several timed conditions where the time available for item completion was limited using the RS paradigm. Thus, speed was manipulated by varying the available time between stimulus-onset and RS. A total of $N = 205$ high school students participated in the study. Results showed that across timed conditions with decreasing time on task, the ability level and ability variance decreased substantially. Ability correlations between timed conditions were high, whereas correlations between untimed and timed conditions were low. This finding suggested that ability differences being inconsistent to those found in the timed condition are due to individual differences in time on task in the untimed condition. To eliminate these differences, two ways were considered. First, untimed responses were recoded using two-tailed posterior time limits. As expected, correlations between timed and untimed conditions were increased. Second, the log-transformed item response times were included in the item response model, which led to even higher correlations between timed and untimed conditions. Validity and generalizability of the proposed testing procedure are discussed.

Keywords

speed-ability trade-off, time on task, experimental time limits, posterior time limits, response time modeling, item response modeling

[1]German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany
[2]Centre for International Student Assessment (ZIB), Frankfurt/Main, Germany

Corresponding Author:
Frank Goldhammer, German Institute for International Educational Research (DIPF), Schloßstr. 29, 60486 Frankfurt/Main, Germany.
Email: Goldhammer@dipf.de

From a measurement perspective, the speed-ability trade-off (SAT) represents a fundamental problem, as it may jeopardize comparability of performance measures if there is between-subject variation in the speed-ability compromise adopted by test takers (e.g., Dennis & Evans, 1996). Thus, the test taker's ability θ cannot be conceived as a single measurement, but has to be viewed as a monotonic decreasing function that defines the within-subject relation between ability θ and speed ζ (i.e., $\theta = f(\zeta)$; cf. van der Linden, 2009). From this, it can be assumed that when test takers p operate at different speeds, ability estimates $\hat{\theta}_p$ indicate individual ability differences, and these ability estimates are confounded with the test takers' decisions on speed. For example, if two monotonic decreasing speed-ability functions are assumed for two test takers (A and B; see Online Appendix A), it could be that A shows greater ability than B at all speed levels. However, B could obtain a higher ability estimate than A because A's speed may be much higher than B's. The SAT is a within-subject phenomenon: If test takers increase their speed, their ability decreases. However, even if test takers keep their speed constant (stationarity assumption, cf. van der Linden, 2007), the problem of comparing ability estimates still exists as long as test takers select very different levels of speed to complete the tasks (e.g., matrices tasks; cf. Goldhammer & Klein Entink, 2011).

Controlling Speed Differences Among Test Takers

To obtain comparable ability estimates that are not affected by individuals' differences in the speed-ability compromise, in the present study the speed control was removed from the test taker and given to the test developer (cf. Wainer et al., 2000). This was done by constraining the response period. According to van der Linden (2009), the expected log-transformed response time $\mathrm{E}(\ln(t_{pi}))$ is determined by person and item properties and can be broken down into the speed ζ_p of person p and the time intensity λ_i of item i to indicate how time-consuming the item is $\mathrm{E}(\ln(t_{pi}) = -\zeta_p + \lambda_i$. Test takers' levels of speed ζ_p, while completing an item i can be standardized by constraining the available time on task $\ln(t_{pi})$, that is, the time which can be spent on processing the stimulus and responding. Those able to meet the time constraint at item level have adapted their speed ζ_p to the same level. If time intensity is equal across items, speed is the same across items, given the same time constraint; however, if time intensity varies across items, the speed differs across items but not among test takers responding to a particular item.

There are various methods available to control time on task, which prevent too fast responses, too slow responses, or both (e.g., Davison, Semmes, Huang, & Close, 2012; Lien, Ruthruff, Remington, & Johnston, 2005; Reed, 1973; Semmes, Davison, & Close, 2011; Wright & Dennis, 1999). For the present study, the experimental response-signal (RS) paradigm was selected, as it enables control of the test taker's speed by restricting the time available to process the stimulus and give a response. To explore the SAT as a within-subject phenomenon, the RS paradigm was applied repeatedly within subjects with varying stimulus presentation times (timed conditions). To compare the results from the timed condition with results from a conventional administration, a condition without time constraints (i.e., an untimed condition without RS paradigm) was needed. The SAT is assumed to affect the results of both speed tests and power tests. For the present study, tasks from tests requiring figural discrimination and visual word recognition were selected, which could be assumed to be primarily speed tests. Given enough time, these items should be solved almost always correctly.

A major assumption of the approach is that test takers react in a similar way to the introduced time constraints: There are no confounding differential effects across timed conditions due to the applied RS paradigm, which would be reflected by intersecting speed-ability functions. Nevertheless, it was expected that the RS paradigm would show differential effects

relative to the untimed condition in that test takers might be more or less required to adjust their (untimed) timing behavior to the experimental time limits in the timed conditions.

Hypotheses

In the present study, the following hypotheses based on the theoretical background on SAT were investigated. A precondition of the analyses was that the experimental manipulation by means of the RS paradigm succeeded in reducing differences among individuals' time spent on task in the timed conditions. In Hypothesis 1, it was assumed that by decreasing the available time on task, or more specifically, by shortening the time for item presentation, and in turn requiring participants to increase their speed ζ, the mean ability $\mu(\widehat{\theta}_t)$ would decrease in timed conditions, $t \geq 1$. In Hypothesis 2, it was expected that a very high speed would be associated with decreasing ability variance $\mathrm{Var}(\widehat{\theta}_t)$, as responses would become increasingly random regardless of participants' ability. For a speed test, it was assumed that this variance would decrease with very low levels of speed, as easy items could be solved by almost all test takers. In Hypothesis 3, it was supposed that significant speed differences in the untimed condition would result in low correlations $\mathrm{Cor}(\widehat{\theta}_0, \widehat{\theta}_t)$ between ability estimates of the untimed condition, $t = 0$, and any timed condition, $t \geq 1$. Following the same reasoning, the correlations between different timed conditions $t1$ and $t2$, $\mathrm{Cor}(\widehat{\theta_{t1}}, \widehat{\theta_{t2}})$, with $t1$ and $t2 \geq 1$, were expected to be at a much higher level, as test takers were required to adopt the speed-ability compromise in the same way in each condition. Differences in the correlations were not assumed to be due solely to bottom or ceiling effects. Low correlations between ability estimates in the untimed condition, $t = 0$, and any timed condition $\mathrm{Cor}(\widehat{\theta}_0, \widehat{\theta}_t)$, $t \geq 1$, were assumed to be due to individuals' differences in the time spent on task in the untimed condition. Therefore, correlations were expected to increase when the effect of individuals' differences in the time spent on a task was eliminated. In Hypothesis 4, it was assumed that individual differences could be equalized by recoding response data based on two-tailed posterior time limits. A response was scored as correct only if the correct answer was given *and* the response was given within the time limit; a response was considered incorrect if it was given outside the time limit or was incorrect within the time limit (cf. Partchev, De Boeck, & Steyer, 2012). For posterior time limits, one pair of upper and lower limits that were applied as experimental time limits of the RS paradigm was used. Moreover, it was systematically explored how the location of posterior time limits affected the results (see Online Appendix C). Although two-tailed posterior time limits could eliminate speed differences and increase correlations between untimed and timed conditions, an alternative was to incorporate directly the response time into the item response model, as proposed in Hypothesis 5 based on Roskam's (1987) approach (see also Roskam, 1997; for power tests, see Wang & Hanson, 2005). In the present study, this was done in a straightforward way by adding the log-transformed response time as a linear person-by-item covariate. Thus, on the person level the confounding of item response time and ability estimate was disentangled, and the accuracy at which a test taker operated depended on the composite of his or her ability and time spent on task.

Method

Participants

Of the 205 high school students (Grade 12) participated in this study, 57.6% were female and 42.4% male aged 15.5 to 21.75 ($M = 18.03$, $SD = .70$). Students completed the tasks in groups of up to 24 in a classroom setting, supervised by two test administrators.

Tasks

Figural discrimination task. This task required test takers to respond selectively to figural targets and non-targets by pressing one of two response buttons. The stimuli were geometrical figures similar to those on the Frankfurt Adaptive Concentration Test (FACT; cf. Goldhammer, Moosbrugger, & Krawietz, 2009) and differed in the following four ways: outer shape (circle vs. square), inner shape (circle vs. square), number of dots within the inner shape (two vs. three), and orientation of dots (diagonal vs. horizontal). Target items consisted of inner squares with two dots and inner circles with three dots, whereas non-target items consisted of inner squares with three dots and inner circles with two dots.

Visual word recognition task. This task (cf. Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Richter, Naumann, Isberner, & Kutzner, 2011) required participants to distinguish between words and equivalent non-words by pressing the corresponding response button. All the words were nouns, with their length varying between 3 and 10 letters and between one and three syllables (for further details on item development, see Richter et al., 2011).

In both tasks, stimuli appeared successively on the screen and in a random order, which was the same for all test takers. In the untimed condition, participants were asked to work as quickly as possible and to avoid making errors. In the timed conditions, they were instructed to press the correct response button once the RS was presented and to give as many correct responses as possible. In each of the conditions, participants judged eight stimuli of both response categories.

RS Paradigm

In the RS procedure (e.g., Miller, Sproesser, & Ulrich, 2008; Reed, 1973), participants were given a signal when they were supposed to respond. Thus, the RS allowed to control effectively the time used to perform the task on a given trial and thereby to trace the SAT. As blocks of trials with constant RS lags were used, test takers knew in advance the relative amount of time available to perform the task and therefore could try to respond as accurately as possible in the time given. Previous research (e.g., Miller et al., 2008) has shown that test takers are able to respond within 300 ms after the RS. Each trial began with a 500-ms presentation of a centered fixation cross. When the fixation cross disappeared from the screen, the stimulus was presented. For the untimed condition task, test takers decided individually when to respond. After responding, a blank screen appeared for 500 ms. In the timed conditions, the stimulus disappeared once the predefined presentation time elapsed, which was indicated by the RS. The signal indicating that it was time to respond was a 150-ms beep of 900 Hz. The participants were required to give a response as soon as he or she heard the RS via earphone and no later than 300 ms after the onset of the RS. Feedback on timing was provided in all timed trials: If the response was given in time, a happy face was presented for 800 ms; if the response was too early or too late, an unhappy face was presented for 1,200 ms as well as the message ''too early'' or ''too late,'' respectively. After presenting the feedback, a blank screen was shown for 500 ms.

The stimulus presentation times (i.e., RS lags) were derived from quantiles of the response time distributions, which had been obtained in previous studies without time constraints. For figural discrimination, the values of the 10th, 30th, 50th, 70th, and 90th percentiles were used; for word recognition, the 20th, 40th, 60th, and 80th percentiles were used. Based on pilot studies, the experimental RS lags were further adjusted by a linear transformation. This enabled to define a range of RS lags that were appropriate for the assessment of individuals' differences within the group of participants.

Design

First, participants completed in the untimed condition the visual word recognition test, which consisted of a block of 16 trials. Then, they completed the test in five timed conditions with decreasing RS lags of 1,069, 741, 393, and 189 ms. Similarly, the test takers first completed the figural discrimination test in the untimed condition, and then they did the test in five timed conditions with decreasing RS lags of 970, 571, 440, 325, and 184 ms. The order of conditions was the same for all participants. To avoid carryover effects by presenting a particular word repeatedly, five parallel test forms including different words and non-words were administered for the word recognition test. These test forms were constructed by matching word properties and empirical item properties (i.e., item difficulty, discrimination) across test forms. Figural discrimination test forms differed only in the order of geometrical figures. At the beginning, participants completed 14 untimed practice trials and received immediate error feedback on the screen. If more than five errors were made, practice trials were repeated once, which happened rarely. Before each block of timed trials, participants were made familiar with the RS paradigm and the stimulus presentation time. To practice responding to the RS in time, six trials with neutral stimuli were used.

Data Analysis

The Generalized Linear Mixed Model (GLMM) framework (e.g., De Boeck et al., 2011; De Boeck & Wilson, 2004) was used to model responses given in the timed and untimed conditions. Assuming that there was one observation per person p and item i, the observed dichotomous response Y_{pi} follows a binominal distribution, $Y_{pi} \sim \text{binominal}\left(1, \pi_{pi}\right)$ with π_{pi} as success parameter and expected value, respectively. The expected value is linked to a continuous range from $-\infty$ to $+\infty$ by means of the logit function, $\eta_{pi} = \ln\left(\pi_{pi} / \left(1 - \pi_{pi}\right)\right)$. For the one-parameter logistic model (1PL or Rasch model), η_{pi} is a linear combination of the person parameter θ_p as a random effect and the item easiness parameter β_i as a fixed effect. For Hypotheses 1 to 4, this item response theory (IRT) model was extended to a multidimensional 1PL model by introducing item partition covariates indicating whether an item belonged to a particular dimension. For each (un)timed condition, one dimension, $t = 0, \ldots, T$, was assumed with $T = 5$ for figural discrimination and $T = 4$ for word recognition. The effects of the item partition covariates were modeled to be random effects varying among persons p, that is, θ_{pt}. A categorical covariate c_t was added to represent the fixed effects of the conditions. Thereby, the mean structure of the person parameter was provided. The component θ_{pt} then represented the individual ability not explained by the condition. To obtain measurement model invariance, the easiness parameters β_i of the items were constrained, so that they were equal across untimed and timed conditions for matched items, meaning that the item indicator covariate showed the same values $i = 0, \ldots, 15$ across conditions for those items that were matched across test forms. Assuming that this level of measurement invariance held, changes in test takers' response behavior could be explained with the structural part of the model. The following linear components of the GLMM resulted as follows: $\eta_{pit} = \theta_{pt} + \beta_i + c_t$, with $\boldsymbol{\theta}_p \sim \text{MVN}(0, \Sigma_\theta)$ and Σ_θ as the covariance matrix of the random effects. To address Hypothesis 5, the model was extended by $\ln\left(t_{\text{pit}}\right)$, representing the log-transformed response time: $\eta_{pit} = \theta_{pt} + \beta_i + c_t + \omega \ln\left(t_{\text{pit}}\right)$, with ω as the fixed effect. The random intercept θ_{pt} at person level then represented the ability without reflecting the time spent on task. The model suggests that if the response time is increased, the probability of a correct response approaches 1, regardless of the difficulty of the item. To estimate the GLMMs, the lmer function of the lme4 package (Bates, Maechler, & Bolker, 2011) for the R environment (R Development Core Team, 2011) was used; for classical reliability analysis

Table 1. Lower and Upper Limits of the Response Window as well as Means and Standard Deviations of Response Times in the Untimed and Timed Conditions (in ms).

Task	Condition	Untimed 0	Timed 1	Timed 2	Timed 3	Timed 4	Timed 5
Word recognition	Lower limit	NA	1,569	1,241	893	689	NA
	Upper limit	NA	1,869	1,541	1,193	989	NA
	$M(t_t)$	1,415	1,755	1,477	1,133	950	NA
	$SD(t_t)$	530	225	121	112	102	NA
	% too early	NA	.14	.06	.07	.02	NA
	% too late	NA	.08	.07	.12	.17	NA
Figural discrimination	Lower limit	NA	1,470	1,071	940	825	684
	Upper limit	NA	1,770	1,371	1,240	1,125	984
	$M(t_t)$	1,457	1,702	1,289	1,165	1,058	928
	$SD(t_t)$	673	138	165	137	142	141
	% too early	NA	.07	.10	.09	.07	.04
	% too late	NA	.08	.08	.10	.13	.19

Note. The time of the limits is from the onset of the fixation cross; the response time is from the onset of the fixation cross to the response. The item presentation time is the lower limit minus the 500-ms presentation time of the fixation cross.
NA = not applicable.

(Cronbach's α), the ltm package (Rizopoulos, 2006) was used. Based on the parameters of the IRT model, the ICC(k) coefficient as the reliability of the sum of all items was computed (cf. DeBoeck, 2008; Semmes et al., 2011) as follows: $\mathrm{ICC}(k) = \mathrm{Var}(\theta)/(\mathrm{Var}(\theta) + \mathrm{Var}(\varepsilon)/n)$, where $\mathrm{Var}(\theta)$ is the estimated ability variance, $\mathrm{Var}(\varepsilon)$ is the error variance which is for the logit link function $\mathrm{Var}(\varepsilon) = \pi^2/3 = 3.29$, and n is the number of items. Two subjects were excluded from the data analysis, as their proportion of correct responses in the untimed condition was more than three standard deviations below the mean.

Results

Manipulation Check

Table 1 shows the mean response time $M(t_t)$ and standard deviation $SD(t_t)$. For each timed condition, $M(t_t)$ is located as required by the RS lag and the imposed upper and lower time limits for a response. Most important, individual differences in response time t_t were substantial in the untimed condition, whereas variability was remarkably reduced in the timed conditions. The standard deviation $SD(t_t)$ decreased substantially from the untimed condition to the Timed 1 condition (see Table 1; see also Online Appendix B for a plot of response time distributions). For word recognition, a further decrease was observed from Timed 1 condition to Timed 2 condition, which might indicate that test takers were becoming more familiar with the RS paradigm. For figural discrimination, which was administered after word recognition, such a decrease could not be observed. Table 1 also provides the percentage of responses that were too early or too late for each timed condition. There was a trend that as RS lags decreased, the percentage of delayed responses increased; however, overall the percentage of responses that were too early or too late was low. These responses were not excluded from data analysis, as test takers usually only just missed the response window. Overall, the data presented in Table 1 clearly suggest that the RS paradigm succeeded in reducing individuals' differences in the time spent on task. Hence, persons adapted their speed to the same level when completing an item.

Reliability

For figural discrimination, Cronbach's α of the accuracy scores obtained for the untimed condition was very low, $\alpha_0 = .40$; for the timed conditions, it was reasonably high in the first three timed conditions, $\alpha_1 = .71$, $\alpha_2 = .71$, and $\alpha_3 = .66$. As expected, for the fastest conditions, in which test takers were forced to show more random response behavior, reliability dropped substantially, $\alpha_4 = .55$ and $\alpha_5 = .40$. For word recognition, the pattern was similar: $\alpha_0 = .35$ for the untimed condition and $\alpha_1 = .77$, $\alpha_2 = .80$, $\alpha_3 = .65$, and $\alpha_4 = .47$ for the timed conditions. The ICC(k) coefficients representing the reliability of the sum of all items were similar, although the ICC(k) values were a bit higher than the Cronbach's α values. For figural discrimination, the results were $\text{ICC}(k)_0 = .64$, $\text{ICC}(k)_1 = .90$, $\text{ICC}(k)_2 = .83$, $\text{ICC}(k)_3 = .75$, $\text{ICC}(k)_4 = .60$, and $\text{ICC}(k)_5 = .30$; for word recognition, the ICC(k) coefficients were $\text{ICC}(k)_0 = .67$, $\text{ICC}(k)_1 = .88$, $\text{ICC}(k)_2 = .90$, $\text{ICC}(k)_3 = .70$, and $\text{ICC}(k)_4 = .42$.

Results: Hypotheses 1 to 5

Mean structure. As expected in Hypothesis 1, the mean ability $\mu(\hat{\theta}_p) = c_t$ decreased across timed conditions, that is, with decreasing item presentation time. For figural discrimination (see Table 2, Model M0), there was a continuous decrease across timed conditions and an increase from the untimed to the first timed condition, in which test takers were forced to take more time than they did on average in the untimed condition. For word recognition (see Table 3, Model M0), the decrease was only minor for the first two timed conditions and means were as high as the mean of the untimed condition. However, for the last two timed conditions the mean ability was reduced substantially. This decrease in ability occurred only in relatively fast timed conditions, as test takers were forced to take less time for task completion than on average in the untimed condition, that is, for figural discrimination $M(t_2) < M(t_0)$ and for word recognition $M(t_3) < M(t_0)$ (cf. Table 1).

Variance structure. As assumed in Hypothesis 2, for figural discrimination, the ability variance $\text{Var}(\hat{\theta}_t)$ decreased dramatically as speed increased (see Table 2, Model M0). Similarly, for word recognition, $\text{Var}(\hat{\theta}_t)$ was reduced during the two fastest timed conditions (see Table 3, Model M0). As expected for a test with a considerable speed component, the ability variance of word recognition increased across the first two timed conditions, suggesting that individuals' ability can be distinguished more easily at medium speed than at extremely high or low speeds. However, this could not be observed for figural discrimination ability variance, which decreased continuously across timed conditions. For both figural discrimination and word recognition, $\text{Var}(\hat{\theta}_t)$ at low speed was higher than the variance observed for the untimed administration. Differences in $\text{Var}(\hat{\theta}_t)$ reflected exactly the differences in Cronbach's α and ICC(k) between timed and untimed conditions.

Correlational structure. From the perspective of measuring individual differences, it was most interesting to compare correlations between conditions. As assumed in Hypothesis 3, comparatively low ability correlations were found between the untimed condition and timed conditions: $.22 \leq \text{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .53$ for figural discrimination (see Table 2, Model M0) and $.24 \leq \text{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .43$ for word recognition (see Table 3, Model M0). This supports the assumption that the SAT gives rise to disordered ability estimates. In contrast, when speed was controlled experimentally, the correlations between timed conditions were about twice as high: $.67 \leq \text{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .95$ for figural discrimination (see Table 2, Model M0) and $.71 \leq \text{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .93$ for word recognition (see Table 3, Model M0). Correlations between timed conditions in the main diagonal were comparatively high, which may reflect the typical

Table 2. Figural Discrimination Parameters From Model Fitting.

Model	Condition	$\mu(\hat{\theta}_t)$	$\mathrm{Var}(\hat{\theta}_t)$	$\mathrm{Cor}(\hat{\theta}_t, \hat{\theta}_t)$				
				U0	T1	T2	T3	T4
M0	Untimed 0	2.17 (0.10)	0.37					
	Timed 1	2.99 (0.15)	1.94	.53				
	Timed 2	1.53 (0.11)	1.01	.22	.81			
	Timed 3	0.90 (0.10)	0.62	.45	.79	.95		
	Timed 4	0.31 (0.09)	0.31	.49	.68	.79	.93	
	Timed 5	−0.09 (0.08)	0.09	.37	.69	.67	.78	.92
M1	Untimed 0	−0.54 (0.10)	0.66					
	Timed 1	3.00 (0.15)	1.93	.59				
	Timed 2	1.55 (0.11)	0.98	.51	.83			
	Timed 3	0.93 (0.10)	0.61	.60	.78	.96		
	Timed 4	0.33 (0.09)	0.31	.44	.67	.81	.92	
	Timed 5	−0.06 (0.08)	0.09	.39	.70	.66	.79	.94
M2	Untimed 0	1.97 (0.10)	0.32					
	Timed 1	2.41 (0.16)	1.93	.70				
	Timed 2	1.44 (0.11)	0.89	.36	.79			
	Timed 3	0.98 (0.10)	0.55	.63	.80	.93		
	Timed 4	0.54 (0.09)	0.27	.67	.67	.78	.91	
	Timed 5	0.32 (0.09)	0.07	.65	.65	.62	.74	.93

Note. Standard errors for fixed effects are given in brackets; Model M0 = baseline model; Model M1 = model using untimed responses recoded by means of two-tailed posterior time limits as applied for Timed 2 condition; Model M2 = model including $\ln(t_{pit})$, $\ln(t_{pit})^2$, and $\ln(t_{pit})^3$ as predictors.

finding that successive measures are more highly correlated than those showing greater temporal distance.

Item response modeling using two-tailed posterior time limits. To test Hypothesis 4, responses in the untimed condition were recoded by applying two-tailed posterior time limits. Responses were considered correct if they were given within these limits, whereas responses were considered incorrect if they were given beyond the limits or were incorrect within the time limits (cf. time-accuracy data as used by Partchev et al., 2012). To compare untimed and timed procedures, posterior time limits were set as the experimental time limits used in timed conditions. Medium-fast time limits to maximize the number of responses within the limits were selected. For figural discrimination, the limits of Timed 2 condition were used, covering 47.86% of untimed responses. For word recognition, the limits of Timed 3 condition were chosen including 36.83% of untimed responses. Recoding substantially increased Cronbach's α for figural discrimination from $\alpha_0 = .40$ to .70 and for word recognition from $\alpha_0 = .35$ to .81. Reliability as assessed by ICC(k) also increased for figural discrimination from $\mathrm{ICC}(k)_0 = .64$ to .76 and for word recognition from $\mathrm{ICC}(k)_0 = .67$ to .83. Table 2 (Model M1) shows that for figural discrimination the mean ability $\mu(\hat{\theta}_0)$ in the untimed condition decreased substantially due to the stricter scoring rule. The ability variance $\mathrm{Var}(\hat{\theta}_0)$ almost doubled, reflecting increased reliability. Most importantly, as assumed in Hypothesis 4, the correlations between untimed and timed conditions increased: $.39 \leq \mathrm{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .60$. As shown in Table 3 Model M1, results obtained for word recognition were similar. The mean ability $\mu(\hat{\theta}_0)$ was much smaller, and the ability variance $\mathrm{Var}(\hat{\theta}_0)$ more than doubled. Again, the correlations between untimed and timed conditions clearly increased $.44 \leq \mathrm{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .60$ (to see how the location of posterior time limits affected results, see Online Appendix C).

Table 3. Word Recognition Parameters From Model Fitting.

Model	Condition	$\mu(\hat{\theta}_t)$	$\mathrm{Var}(\hat{\theta}_t)$	$\mathrm{Cor}(\hat{\theta}_t, \hat{\theta}_t)$			
				U0	T1	T2	T3
M0	Untimed 0	4.21 (0.16)	0.43				
	Timed 1	4.21 (0.17)	1.49	.34			
	Timed 2	4.13 (0.18)	1.87	.43	.87		
	Timed 3	2.33 (0.14)	0.47	.43	.85	.93	
	Timed 4	1.23 (0.13)	0.15	.24	.75	.71	.91
M1	Untimed 0	−0.36 (0.12)	1.04				
	Timed 1	4.03 (0.16)	1.51	.48			
	Timed 2	3.95 (0.16)	1.90	.51	.86		
	Timed 3	2.12 (0.12)	0.49	.60	.83	.91	
	Timed 4	0.99 (0.10)	0.17	.44	.70	.65	.89
M2	Untimed 0	4.16 (0.16)	0.40				
	Timed 1	3.57 (0.18)	1.35	.55			
	Timed 2	3.48 (0.18)	1.77	.68	.91		
	Timed 3	2.68 (0.14)	0.41	.73	.90	.94	
	Timed 4	2.01 (0.21)	0.11	.54	.82	.73	.91

Note. Standard errors for fixed effects are given in brackets; Model M0 = baseline model; Model M1 = model using untimed responses recoded by means of two-tailed posterior time limits as applied for Timed 2 condition; Model M2 = model including $\ln(t_{\mathrm{pit}})$ as predictor.

Item response modeling incorporating item response times. Hypothesis 5 assumed that by introducing the covariate $\ln(t_{\mathrm{pit}})$ the confounding of ability and time spent on task could be removed, which should increase correlations between untimed and timed conditions. The response time $\ln(t_{\mathrm{pit}})$ was modeled as covariate for both untimed and timed conditions with $SD(t_t)$ being high for the untimed condition and low for the timed conditions (cf. Table 1). For figural discrimination, the effect was positive and significant, $\omega_1 = 0.65 (z = 7.65, p < .01)$, suggesting that spending more time on a task is associated with a higher probability of giving a correct response. However, the correlations between untimed and timed conditions rose only moderately, $.23 \leq \mathrm{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .63$. To explore whether item response time showed a non-linear functional relationship, the second- and third-order polynomial were modeled successively by including $\omega_2 \ln(t_{\mathrm{pit}})^2$ and $\omega_3 \ln(t_{\mathrm{pit}})^3$. To counter multicollinearity, for all analyses $\ln(t_{\mathrm{pit}})$ was centered by item. In the final model, positive significant effects were obtained for the linear component, $\omega_1 = 1.40 (z = 10.41, p < .01)$, quadratic component, $\omega_2 = 0.26 (z = 7.18, p < .01)$, and cubic component, $\omega_3 = 0.05 (z = 2.64, p < .01)$. As a result, the correlations between untimed and timed conditions increased further, $.36 \leq \mathrm{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .70$, with four of five correlations $\geq .63$ (see Table 2, Model M2). The fourth-order polynomial model showed convergence problems. For word recognition, the revealed effect of the covariate $\ln(t_{\mathrm{pit}})$ was positive and highly significant, $\omega_1 = 2.42 (z = 11.42, p < .01)$. Higher order polynomials did not show statistical significance or were not stable. As shown in Table 3 Model M2, correlations $\mathrm{Cor}(\hat{\theta}_0, \hat{\theta}_t)$ between untimed and timed conditions increased substantially, $.54 \leq \mathrm{Cor}(\hat{\theta}_0, \hat{\theta}_t) \leq .73$. Interestingly, for word recognition, the correlations between timed conditions were also slightly higher, especially for correlations with Timed 1 condition. This suggests that the remaining response time variation in the timed conditions, which was highest for Timed 1 condition (cf. Table 1), had a slight effect on ability. This means that the ability estimates obtained with the RS paradigm were still slightly confounded with the self-selected time spent on task.

Discussion

The presented findings show that individuals' differences in the SAT, that is, in the chosen time on task, can be substantially reduced by means of the RS paradigm. As suggested by the assumed speed-ability functions, $\theta = f(\zeta)$, this results in more consistent differences in ability estimates as shown by high ability correlations across timed conditions. Most importantly, ability correlations between untimed and timed conditions could be increased by applying posterior time limits and by including item response times in the model. The results generalize remarkably well across the different types of tasks.

Reliability analyses showed very low Cronbach's α and low ICC(k) values for the untimed condition, whereas reliability was much higher in the timed conditions as long as the time limit was not too short. If the timed condition became shorter, responses were given more and more randomly regardless of the individual's ability, which reduced the estimated ability variance and reliability of ability scores. Low reliability in the untimed condition could not be explained by the low level of item difficulty and the potentially associated variance restriction. In the very slow timed conditions, estimates for the shown ability were comparably high (word recognition) or even higher (figural discrimination), but the reliabilities of these timed conditions clearly exceeded the ones of the untimed condition. This suggests that differences in the time spent on task and in balancing the SAT heavily impair reliability in the untimed condition. This interpretation is supported by the observation that by introducing posterior time limits, and thereby eliminating speed differences, reliability increased substantially. Differences in time on task and speed can occur at the between- and within-subject levels. Reliability seems to be threatened especially by within-subject variation in time on task, which is associated with changes in how individuals differ in time on task. If a test taker's time on task relative to that of others' varies from item to item, the relative success rate is also expected to vary, reducing intercorrelations of item scores. This interpretation is supported by only small to moderate correlations between item response times in the untimed condition.

As expected, with increasing speed across the timed conditions, the mean ability level decreases. Most interestingly, it declines substantially once the presentation time falls below the average self-determined response time level. Moreover, the ability variance was reduced in extreme timed conditions including long and short stimulus presentation times; for figural discrimination, this decrease could be observed only in the fastest timed condition. Thus, the word recognition task was more like a speed test, which was also indicated by higher ability levels, that is, word recognition items were easier than figural discrimination items.

Regarding the consistency of individual differences across the various conditions, low correlations were found between untimed and timed conditions, while the timed conditions were highly correlated. This suggests that ability levels in the untimed condition were inconsistent with those found in the timed conditions due to speed differences in the untimed condition. To support further this conclusion, individual differences in time on task in the untimed condition were reduced post hoc in two ways. As expected, when using two-tailed posterior time limits, the correlations between ability estimates from untimed and timed conditions increased substantially and were even higher when incorporating item response times into the 1PL model. Notably, incorporating item response times had less effect on mean and variance of the ability distribution in the untimed condition than it did in either posteriori time limits or experimental time limits.

An important assumption of the experimental approach was that the test takers were equally able to adapt their timing and response behavior to the introduced time constraints. This means that the time constraint on the item level in the timed conditions, which was introduced to avoid confounding ability with time on the task, did not evoke confounding with another dimension,

for instance, the ability to deal with the time limits (cf. the speed dimension shown by Semmes et al., 2011, to explain timed item performance in a reasoning test). Some test takers might perform equally well at their ability limits, whereas others might be affected more and operate below their ability limits. However, the observed low ability correlations between untimed and timed conditions are not interpreted as a consequence of confounding with another dimension in the timed conditions. If there were differential effects of the time constraint across timed conditions, this should have lowered the correlation between untimed and timed conditions, as well as that between timed conditions, as the level of time constraint varied substantially from very generous to strict time limits. Direct empirical evidence was provided by applying posterior time limits and including the item response time as covariate. Thereby, correlations between untimed and timed conditions could be raised substantially, suggesting that differences in the time spent on task in the untimed condition lowered the correlations to a large extent. Nevertheless, further research is needed to address the question of whether the timed procedure affects the (construct) validity of the measure.

What are the implications of the results for applied measurement? The results suggest that item time limits are a suitable way to deal with the SAT in speeded types of measurements. Test takers were able to give responses in time, and there was no evidence for confounding factors. Moreover, the results suggest that time limits allow test developers to manipulate easily item difficulty, which would be beneficial for adaptive testing. Items of varying levels of difficulty could be generated automatically by increasing or decreasing item presentation time (cf. Goldhammer et al., 2009; Wainer et al., 2000). Applying posterior time limits to recode untimed items requires some caution, as individual differences and reliability depend on the location of time limits (see Online Appendix C). However, this approach offers the possibility to increase reliability substantially. Finally, incorporating item response times into IRT models enables explanation of differences in item responses. However, this approach is less straightforward to implement for measurement practice. Interpreting such fixed response time effects also requires consideration of the correlation structure of response time-related as well as response-related item and person parameters (cf. van der Linden, 2009; see also Goldhammer et al., 2014). Regarding limitations of the present study and related future research goals, it must be borne in mind that for figural discrimination, higher order polynomial terms were added in an exploratory way. This research needs to be replicated with another sample. In addition, the obtained result pattern should be tested to determine whether it depends on the order of conditions, which was fixed among test takers. In the present study, the influence of time limits was investigated at the latent structural level and restrictions were introduced into the measurement models. However, it would also be interesting to reverse the perspective by fixing parameters in the structural model and investigate how freely estimated item parameters change across conditions. For this, also more liberal item response models with more item parameters could be tested.

In this study, task material from tests was used, which were primarily speed tests. Accordingly, very high rates of correct responses were obtained for the untimed condition and for the slow timed conditions. For both researchers and practitioners, it would be very interesting if the timed procedure at the item level also could be applied to power tests (cf. the approach by Wright & Dennis, 1999). There is evidence that administering power tests under time constraints at test level increases the shared variance with mental speed (e.g., Preckel, Wermer, & Spinath, 2011). Semmes et al. (2011) introduced speededness at item level by setting a one-tailed upper time limit at the median item response time obtained from untimed administration. They provided evidence for the existence of a speed dimension underlying timed item performance. Thus, to prevent speededness and construct irrelevant variance, the time constraint at item level should not be too strict. For instance, Walczyk, Kelly, Meche, and

Braud (1999) allowed adults to read texts under various time pressure conditions. Under mild time pressure, reading comprehension was improved, probably due to increased effort and motivation; under severe time pressure, participants displayed reduced performance and increased stress level. Instead of introducing strict time limits at task level, an interesting option to standardize timing behavior would be to provide feedback on the elapsed time and whether test takers proceeded too quickly or too slowly to the next item without forcing them to work further on the current item. In the present study, the approach reducing individual differences in the speed-ability compromise was to keep the time spent on a task constant among test takers. Assuming that ability is a function of speed, $\theta = f(\zeta)$, this seems to be reasonable. However, at least for speed tests the SAT appears to be a symmetrical problem, that is, it is possible to control speed and assess ability, or to control ability and assess speed. Controlling ability to a particular level among test takers could be done by adapting the allowed time spent on individual items (cf. Goldhammer et al., 2009). For power tests, however, this would not be feasible, as the effective ability level is not fully under the control of the testing procedure, but depends also on the individual's maximum ability level that can be achieved.

Acknowledgments

The authors express their gratitude to two anonymous reviewers for their thoughtful comments. They are also grateful to Johannes Naumann and Tobias Richter for making the word recognition items available for this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Federal Ministry for Education and Research (BMBF).

Supplemental Materials

The online appendices are available at http://apm.sagepub.com/supplemental.

References

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283-316.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes [Computer software] (R package version 0.999375-42). Retrieved from http://CRAN.R-project.org/package=lme4

Davison, M. L., Semmes, R., Huang, L., & Close, C. (2012). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement*, *72*, 245-263.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533-559.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1-28.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Dennis, I., & Evans, J. St. B. T. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology*, *87*, 105-129.

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*, 108-119.

Goldhammer, F., Moosbrugger, H., & Krawietz, S. (2009). FACT-2-The Frankfurt Adaptive Concentration Test. Convergent validity with self-reported cognitive failures. *European Journal of Psychological Assessment*, *25*, 73-82.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*.

Lien, M.-C., Ruthruff, E., Remington, R. W., & Johnston, J. C. (2005). On the limits of advance preparation for a task switch: Do people prepare all the task some of the time or some of the task all the time? *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 299-315.

Miller, J., Sproesser, G., & Ulrich, R. (2008). Constant versus variable response signal delays in speed–accuracy trade-offs: Effects of advance preparation for processing time. *Perception & Psychophysics*, *70*, 878-886.

Partchev, I., De Boeck, P., & Steyer, R. (2012). How much power and speed is measured in this test? *Assessment*, *20*, 242-252.

Preckel, F., Wermer, C., & Spinath, F. M. (2011). The interrelationship between speeded and unspeeded divergent thinking and reasoning, and the role of mental speed. *Intelligence*, *39*, 378-388.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*, 574-576.

Richter, T., Naumann, J., Isberner, M.-J., & Kutzner, Y. (2011). *Diagnostik von Lesefähigkeiten bei Grundschulkindern: Eine prozessorientierte Alternative zu produktorientierten Tests* [Assessment of reading skills in primary school children: A process-oriented alternative to product-oriented tests]. *Diskurs Kindheits- und Jugendforschung*, *6*, 479-486.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1-25.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151-171). Amsterdam, the Netherlands: North-Holland.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New York, NY: Springer.

Semmes, R., Davison, M. L., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement*, *35*, 433-446.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *73*, 287-308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247-272.

Wainer, H., Dorans, N., Green, B., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). Future challenges. In H. Wainer, N. Dorans, D. Eignor, R. Flaugher, B. Green, R. Mislevy, & . . . D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 231-270). Hillsdale, NJ: Erlbaum.

Walczyk, J., Kelly, K., Meche, S., & Braud, H. (1999). Time limitations enhance reading comprehension. *Contemporary Educational Psychology*, *24*, 156-165.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323-339.

Wright, D. E., & Dennis, I. (1999). Exploiting the speed-accuracy trade-off. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 231-248). Washington, DC: American Psychological Association.