Psychological Test and Assessment Modeling, Volume 62, 2020 (2), 147-177

# Rapid guessing rates across administration mode and test setting

*Ulf Kroehne*[1], *Tobias Deribo*[1], *& Frank Goldhammer*[1,2]

**Abstract**

Rapid guessing can threaten measurement invariance and the validity of large-scale assessments, which are often conducted under low-stakes conditions. Comparing measures collected under different administration modes or in different test settings necessitates that rapid guessing rates also be comparable. Response time thresholds can be used to identify rapid guessing behavior. Using data from an experiment embedded in an assessment of university students as part of the National Educational Panel Study (NEPS), we show that rapid guessing rates can differ across modes. Specifically, rapid guessing rates are found to be higher for un-proctored individual online assessment. It is also shown that rapid guessing rates differ across different groups of students and are related to properties of the test design. No relationship between dropout behavior and rapid guessing rates was found.

Keywords: rapid guessing, validity, assessment innovations, technology & assessment, test design & test construction, mode effects, test-taking behavior, log data

[1]*Correspondence concerning this article should be addressed to:* Ulf Kroehne, PhD, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Strasse 6, 60323 Frankfurt am Main, Germany; email: kroehne@dipf.de

[2]Centre for International Student Assessment, Germany (ZIB)

## Introduction

National and international large-scale assessments (LSAs), such as the *Programme for International Student Assessment* (PISA; OECD, 2017), the *Programme for the International Assessment of Adult Competencies* (PIAAC; OECD, 2016), and the *National Assessment of Educational Progress* (NAEP) are all low-stakes assessments. Current research raises concerns about test-taking effort in such low-stakes assessments (e.g., Soland, Wise, & Gao, 2019; Goldhammer, Martens, Christoph, & Lüdtke, 2016; Lee & Jia, 2014). It is assumed that reasonable test-taking effort is required to obtain valid scores (e.g., Wise, 2015). In recent years, many LSAs have changed their administration conditions, moving from paper-based assessment to computer-based assessment (e.g., PIAAC, Yamamoto, Shin, & Khorramdel, 2018; PISA, von Davier, Khorramdel, He, Shin, & Chen, 2019). In addition to such digitalization efforts, LSAs such as the *National Educational Panel Study* (NEPS) in Germany have also begun to investigate administering cognitive tests via un-proctored individual online assessment. Digital assessments make it possible to investigate test-taking effort in terms of rapid guessing rates using response time thresholds (e.g., Wise, 2017). However, little is known so far about the effects of different administration modes and test settings on rapid guessing rates. In this paper, we examine this issue for the population of university students. Using data from several different digital assessments, we compare rapid guessing rates, as an indicator for low test-taking effort, in a test measuring literacy in using information and communication technologies (*ICT literacy*, Senkbeil, Ihme, Wittwer, et al., 2013) and a test measuring *scientific literacy* (Hahn et al., 2013) across different test settings and administration modes. Rapid guessing is considered a possible cause or a mediating variable resulting in measurement non-invariance between groups or conditions defined by administration modes and test settings. Thus, the present study contributes to a better understanding of why establishing measurement invariance across modes and settings may fail. This improved understanding provides the basis for developing countermeasures to obtain comparability (e.g., adapted test designs or data analysis strategies).

### Theoretical background

#### Administration mode effects

Comparisons between different test administration modes (*mode effects*) are often discussed in terms of (measurement) invariance. Mode effects can be defined as differences with respect to different equivalence criteria resulting from properties of the test administration (e.g., Kroehne & Martens, 2011). Which equivalence criteria are appropriate to consider must be identified with respect to a particular purpose, such as comparisons at the item level (item difficulty, item discrimination; e.g., Buerger, Kroehne, Koehler, & Goldhammer, 2019) and at the scale level (e.g., construct equivalence; Kroehne et al., 2019a). The comparability of data across assessment modes is central to LSAs (e.g., von Davier et al., 2019) for reasons such as maintaining the validity of trend estimates (Robitzsch, Lüdtke, Goldhammer, Kroehne, & Köller, 2020) following the introduction of digital assessment methods.

Mode effects, such as item difficulty differences across modes, can also be of substantive interest in research fields such as reading assessment and research on text comprehension (Delgado, Vargas, Ackerman, & Salmerón, 2018). Researchers working in these fields could investigate potential mediator variables explaining the effect of different test administration properties. For instance, Zehner, Kroehne, Hahnel, and Goldhammer (2020) investigated differences in typed vs. handwritten short text responses, and Kroehne, Hahnel, and Goldhammer (2019) compared response processes to investigate the relationship between speededness and mode effects. This paper adds to this line of research by comparing rapid guessing rates as an additional criterion at the scale level.

### Test setting effects

Online data collection promises easy access to geographically diverse populations (e.g., Reips, 2000). However, differences in test administration settings raise potential issues with respect to comparability and measurement invariance (called *setting effects*, e.g., Kroehne, Gnambs, & Goldhammer, 2019). For instance, the presence of a test proctor can affect the results of standardized (large-scale) assessments (e.g., Lüdtke, Robitzsch, Trautwein, Kreuter, & Ihme, 2007) as well as unstandardized online assessments (e.g., Rios & Liu, 2017). Setting effects can affect item parameters and test-taking behavior as well as participation rates (responders vs. unit non-responders). Hence, while the mode (as the specific realization of multiple properties related to the test instrument) can be randomly assigned after test-takers agree to participate (resulting in random equivalent groups), setting-specific participation and drop-out rates could nevertheless lead to a confounding of selection effects and setting effects (e.g., Klausch, Hox, & Schouten, 2013), even if invitations to complete the online assessment are randomly assigned.

### Collecting timing data in paper-based assessments

Collecting response time information in paper-based assessments, either by having investigators use stop watches or by instructing test-takers to record the time after indicating their answer choice on the answer sheet (Blommers & Lindquist, 1944), was suggested in the early years of psychological and educational assessment (e.g., Ebel, 1953). Digital pens (such as Anoto Digital Pens, described by Steimle, 2012) make it possible to digitally collect response time data without specifically instructing test-takers. Digital pens can be used with regular paper booklets printed on top of a unique dot pattern. In addition to processing visible traces on the paper by scanning the booklets, these digital pens collect log data for paper-based assessments in the form of noting the pen's location coordinates (with a particular sampling rate) and the time stamps of all penstrokes made on the paper.

Hence, data collection with digital pens makes it possible to include time measures in the evaluation of mode effects. Previous modal comparisons using data from digital pens included response times in psychometric models such as the bivariate generalized IRT model, which was used to compare the relationship between speed and ability in reading assessments (Kroehne, Hahnel, & Goldhammer, 2019), or the diffusion model (Dirk

et al., 2017). In this paper, we will use log data on response times from a paper-based assessment with digital pens to investigate rapid guessing rates.

### Comparability of response time measures

Comparing rapid guessing rates across different versions of (digital) assessments requires the comparability of the operationalizations of guessing behavior. The method applied in this paper combines data on response times and time thresholds to create a *solution behavior index* (for a review, see, e.g., Lee & Chen, 2011). The comparability of response time measures is critical for this approach.

If the same computer-based assessment software is used in all settings, the identical processing of the log data ensures comparable time measures. Unfortunately, log data collected in paper-based assessments with digital pens differs from log data from computer-based assessments. Apart from the difference in format, the most prominent conceptual distinction is that the presentation of the item material (item stem, question, task) can not be logged in paper-based assessments with digital pens. In practice, this means that no log event exists indicating that a page was turned in the printed booklet, as only stroke data are collected. Hence, we used a theoretical framework that allows indicators to be defined not with respect to raw log events, but rather with respect to specific aspects of the interplay between the test-taker and assessment platform (termed as *states*; Kroehne & Goldhammer, 2018).

Two additional considerations were necessary when creating a comparable response time measure. Firstly, the primary log data transmitted by the digital pens (i.e., coordinate events) can be combined with information in the test booklets (i.e., metadata) to derive higher-level log events with more specific meanings. In this way, derived events can be created such as changing one's response to a question in the paper-based booklet. Secondly, we take advantage of the fact that these types of answer change events are also available for computer-based assessments (we ignored log events only available for computer-based assessments). Having answer change events for all platforms makes it possible to identify a simple state labeled as *working on task* $i$. This state represents the interaction between test-taker and assessment platform related to the particular item $i$ that the test-taker has in mind, before the answer change event for item $i$ triggers a transition to the next state, *working on task* $i + 1$.[1] The total time spent in this state can be interpreted as a measure of the processing time for one item (ignoring answer changes to the same question and summing up time differences due to multiple answer changes). In other words, the time difference between answers to different questions can be used to derive item-level response time measures based on decomposing the test-taking process into meaningful states (Kroehne & Goldhammer, 2018).

Following this approach leads to the generation of item-level response time measures that require taking the item context into account (i.e., the position of items relative to the preceding element of the assessment). In tests with a unit structure, the response time

[1] $i + 1$ refers to the test-taker's self-selected order of completing the assessment tasks.

for the first item within each unit contains a component related to the unit as a whole, for instance, the time required to read a stimulus text.

**Previous research on rapid guessing rates**

An exhaustive review of previous findings on rapid guessing goes beyond the scope of this paper. Rapid guessing has been investigated using data from different LSAs for secondary school students (e.g., NAEP, Lee & Jia, 2014; and PISA, Michaelides, Ivanova, & Nicolaou, 2020) and adult populations with heterogeneous educational attainment levels (e.g., PIAAC, Goldhammer et al., 2016). Rapid guessing has also been investigated by adding response times into latent variable measurement models, trying to identify mixtures of response time distributions (e.g., Schnipke & Pashley, 1997; Meyer, 2010; Pokropek, 2016; Ulitzsch, Davier, & Pohl, 2019). Beyond that, rapid guessing was investigated using methods such as multi-group Rasch models (e.g., Ranger & Kuhn, 2017), and approaches developed to detect differential item functioning (e.g., DeMars & Wise, 2010). While complex latent variable models are promising for investigating properties of the resulting measures, we treat rapid guessing indicators as observed variables in this paper. The approach, based on simple time thresholds, allows a pragmatic identification and treatment of rapid responses, which can be applied in large scale assessments without the need for pre-calibrated item parameters (cf. OECD, 2013; see also Goldhammer, Martens, & Lüdtke, 2017). Our method can, for instance, be used to examine the data quality with regard to rapid guessing before more complex models (such as scaling models) are used. In combination with the random assignment applied in our study, the manifest indicator variables allow us to make statements about the potential impact of rapid guessing on the measurement invariance of assessments conducted in different settings or modes. The remaining section reviews selected findings from previous research on rapid guessing that are related to this paper's research questions and hypotheses.

### Mode and setting

The literature on test-taking motivation assumes that computer-based assessment increases test-takers' motivation (e.g., Chua & Don, 2013; Jerrim, 2016). Accordingly, we expect lower rapid guessing rates for digital assessment compared to paper assessments in the same setting.

Test-taking motivation is also known to be related to proctoring by test administrators (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009), and higher rapid guessing rates are expected in un-proctored individual online assessments (Rios & Liu, 2017).

### Group differences

Previous research on rapid guessing behavior has revealed that males exhibit more rapid guessing behavior (e.g., DeMars, Bashkov, & Socha, 2013; Setzer, Wise, van den Heuvel, & Ling, 2013). This finding is consistent with previous findings on gender differences in self-reported effort (e.g., Butler & Adams, 2007).

Using data from PIAAC, Goldhammer et al. (2017) found higher rapid guessing rates when the test language was not the participant's native language. We expect to replicate this finding in the target population of university students studied in this paper.

#### Test properties and design

Rapid guessing rates have been found to be domain-specific (e.g.,Butler & Adams, 2007) and are expected to be related to the attractiveness of tests (Penk, Pöhlmann, & Roppelt, 2014) and item characteristics (e.g., Wise & Kong, 2005; Wise, 2006). It is known that rapid guessing is more severe for domains that require more reading and for items that are more difficult or appear to be more demanding (Setzer et al., 2013).

Short response times might indicate either hurrying-to-finish behavior (Schnipke & Pashley, 1997) or rapid guessing behavior (Wise & Kong, 2005). Accordingly, rapid guessing rates are related to test speededness (Lee & Chen, 2011, with higher rapid guessing rates expected for more speeded tests.

Limited research has focused on increases in rapid guessing within test sessions (Lindner, Lüdtke, & Nagy, 2019). Based on the literature on position effects (i.e., decreasing performance over the course of an assessment), it is expected that rapid guessing rates increase over the course of the test session.

#### Participation status and dropout

Finn (2015) identified student noncompliance as a potential threat to the validity of low-stakes assessments. Responding to the invitation to participate in a group test setting is expected to filter out students with low test-taking motivation. Accordingly, students who refuse to participate in group testing but do agree to participate in an un-proctored individual online assessment are expected to exhibit higher rapid guessing rates compared to students randomly selected to participate in this test setting.

Dropout (defined as aborting the assessment before reaching the last item) can be understood as an alternative way of expressing low test-taking motivation. Students who complete the test in the un-proctored individual online assessment condition are expected to exhibit lower rapid guessing rates compared to test-takers who drop out. In proctored group testing conditions, test-takers more likely continue until the end of the assessment, *thereby possibly contaminating the data* (Reips, 2000). Dropout rates might also interact with effects of the domain order (Weitensfelder, 2017).

## Research questions

Table 1 summarizes the different conditions examined in this study, which result from the combination of administration mode and test setting. Paper-based assessment was not administered in the un-proctored online setting. The following four research questions comparing rapid guessing rates across test administration modes and test settings were formulated:

**Table 1:**
Summary of the conditions examined in this study

| Setting | Administration Mode | |
|---|---|---|
| | **Computer-based** | **Paper-based** |
| Group (proctored) | proctored group digital assessment | proctored group paper assessment (digital pen) |
| Online (un-proctored) | un-proctored individual online assessment | |

(i) Are rapid guessing rates affected by administration mode and test setting?
(ii) Do different groups of students differ with respect to rapid guessing rates?
(iii) Are rapid guessing rates affected by test properties and test design?
(iv) Are rapid guessing rates related to participation status and observed dropout?

To answer these research questions, we derived the following hypotheses predicting differences in rapid guessing rates across factors (experimental conditions, groups and test-takers with particular behavior), which are summarized in Table 2. No specific expectations about interactions between factors were formulated.

**Table 2:**
Hypotheses about rapid guessing rate across the investigated factors for the four research questions

| RQ | H | Factor | Direction | Rationale | Identification |
|---|---|---|---|---|---|
| (i) | H1 | Mode | Higher rapid guessing rates for paper-based assessment | Rapid guessing related to test-taking motivation; Test-taking motivation affected by mode | Random Assignment |
| | H2 | Setting | Higher rapid guessing rates for un-proctored individual online assessment | Positive effect of interviewer in proctored group testing | Random Assignment (w/o Dropouts) |
| (ii) | H3 | Gender | Higher rapid guessing rates for male students | Replication of previous research | Observed Person-level Covariate |
| | H4 | First Language | Higher rapid guessing rates for non-native speakers | Replication of previous research | Observed Person-level Covariate |
| (iii) | H5 | Domain | Higher rapid guessing rates for the science test | Higher reading load and more demanding items; Higher test speededness | Two selected instruments |
| | H6 | Position | Higher rapid guessing rates for test in second position | Decreasing test-taking motivation | Randomized test position |
| (iv) | H7 | Preference | Higher rapid guessing rates for students non-randomly selected to test mode | Lower committment associated with lower test-taking motivation | Observed participation status |
| | H8 | Dropout | Higher rapid guessing rates for incomplete test sessions | Common source of rapid guessing and dropout behavior | Observed participation status |

*Note.* Hypotheses (H) are assigned to the four research questions (RQ).

## Methods

### Materials and participants

In this study, data from Starting Cohort 5 of an ongoing longitudinal large-scale assessment in Germany (National Educational Panel Study, NEPS; Blossfeld, Roßbach, & von Maurice, 2011) were analyzed for two instruments: a 30-item test measuring literacy in using information and communication technologies (*ICT literacy*, Senkbeil et al., 2013) and a 29-item test measuring *scientific literacy* (Hahn et al., 2013). Both tests use the multiple-choice (MC) and complex multiple-choice (CMC) item formats. In the MC item format, one out of four to five provided response alternatives must be selected, while in the CMC item format, several statements had to be answered with yes or no. The science test was composed of 14 units, each of which starting with a unit stimulus followed by 1 to 3 individual items.

Table 3 presents the sample sizes, mean age, percentage of female students, percentage of students whose first language is German, and dropout percentage. The column *Mode / Setting* refers to the experimental condition (described in the next section).

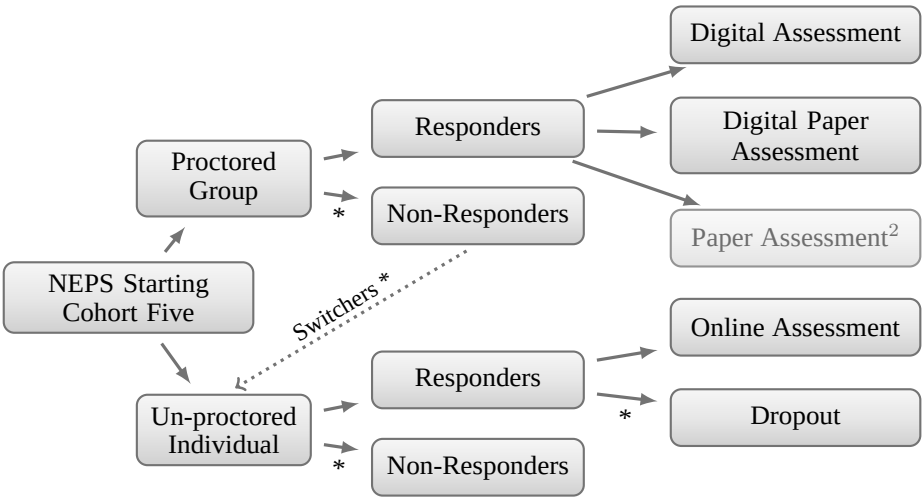
**Table 3:**
Sample size and demographic information about students in NEPS Starting Cohort 5

| Mode / Setting | N | Age | % Female | % Not German | % Dropout |
|---|---|---|---|---|---|
| Proctored Group Digital Assessment | 624 | 27.92 | 71.47 % | 3.69 % | - |
| Proctored Group Paper Assessment | 692 | 27.93 | 63.29 % | 5.78 % | - |
| Proctored Group Paper Assessment (Digital Pen) | 545 | 27.84 | 64.77 % | 4.77 % | - |
| Un-proctored Individual Online Assessment (Sample) | 4906 | 28.19 | 61.99 % | 7.36 % | 7.09 % |
| Un-proctored Individual Online Assessment (Switcher) | 1845 | 27.69 | 65.31 % | 5.15 % | 8.89 % |

### Experimental design

Data collection was carried out in two different administration modes and two different test settings, employing a between-subjects experimental design (see Figure 1). In a first step, several universities were chosen for conducting proctored group testing. All students from the remaining universities were assigned to unproctored online assessment. At the locations chosen for proctored group testing, conventional paper-based assessment, paper-based assessment with digital pens, and digital assessment were all administered.

In a second step, the scheduled sessions at each testing location were assigned randomly to these conditions (i.e., digital assessment, paper-based assessment with digital pens and




*Paths marked with * were affected by self-selection.*

**Figure 1:**
Between-subjects design used to compare mode and setting effects with respect to rapid guessing rates

conventional paper-based assessment[2]). Digital assessment and assessment with digital pens were conducted with bring-in notebooks. While for the digital assessment items were presented and answered on computer screens, paper-based assessment used Anoto Digital Pens (ADP-301), connected with laptops via Bluetooth. Test-takers answered items in printed test booklets, while the digital pens recognized each stroke with a specific dot pattern and a built-in camera. Based on the transmitted coordinate events and events representing pen down and pen up, higher-level log events were derived. In all test sessions conducted as proctored group tests, all students completed one of the two booklets (containing ICT and Science or Science and ICT), and there was no dropout in the group tests. Not all members of the NEPS Starting Cohort 5 panel accepted the invitation and participated in their assigned test setting (responders), some of them refused (non-responders).

Students who refused to take the proctored group test were subsequently invited to participate in the un-proctored individual online assessment (see *switchers* in Figure 1). Again, only a subset of all students invited to participate in the individual online assessment started working on it (responders); some students refused to participate (non-responders). In addition, unlike in the proctored group tests, we observed dropout in the un-proctored individual online assessment, meaning that some students who began the

[2]Data from the random sub-sample of students completing paper-based assessments without digital pens were not considered in this paper.

assessment did not complete it in full (i.e., dropped out before time ran out).

This experimental design makes it possible to evaluate the differences addressed in the hypotheses listed in Table 2. Mode effects in rapid guessing rates can be estimated for the digital assessment and paper assessment (with digital pens) in the proctored group testing condition (H1). Setting effects (H2) can be obtained by comparing rapid guessing rates between the digital assessment in proctored group testing and the un-proctored individual online assessment. Students who completed the test as a paper assessment in a proctored group setting were not included in this comparison. To avoid confounding the setting effect with potential preference effects regarding a particular setting, *switchers* (i.e., students who were only invited to take the online test after not having accepted the invitation to take part in group testing) were treated as a separate group in the analysis. However, as response rates might differ between un-proctored individual online assessment and proctored group assessment, the setting effect is estimated both for complete cases only and for all cases (i.e., with and without dropout).

Gender differences in the rapid guessing rate (H3) and differences between students whose first language is German vs. a different language (H4) are not identified by the experimental design; they instead rest on the observed covariates. Moreover, since the proportions of female students and of students with a first language different than German are not equal across conditions (see Table 3), we will also plot condition-specific differences in rapid guessing rates.

The tests were administered in two booklets (Form A, which began with ICT followed by science, and Form B in the reverse order), which were randomly assigned to test-takers. This design makes it possible to estimate domain difference in rapid guessing rates (H5). Comparing the rapid guessing rates in block position one and two makes it possible to identify the position effect (H6). Note, however, that within each booklet the position of items was fixed (i.e., it is not possible to differentiate between position and block position effects; Rose, Nagy, Nagengast, Frey, & Becker, 2019).

Moreover, students who refused to participate in the proctored group testing but accepted the invitation to participate in the un-proctored individual online assessment (*switchers*) are compared to students originally assigned to the un-proctored individual online assessment. Although multiple reasons for refusing to participate in group testing are possible, we assume that switchers systematically prefer un-proctored individual online assessment over group testing (H7). Comparing rapid guessing rates between students who completed the un-proctored individual online assessment and students who did not reach the end of the assessment makes it possible to identify the effects of dropout behavior (H8). The indicator for completing the test was created in such a way that timeouts were not counted as dropouts. However, it should be noted that technical difficulties (i.e., getting disconnected from the internet) are included as cases of incomplete tests.

**Identification of response time thresholds**

The identification of rapid guessing is based on the assumption that disengaged test-takers are faster than engaged test-takers exhibiting solution behavior. Since, conceptually, solution behavior requires more time than rapid guessing behavior, two response time frequency distributions are expected if rapid guessing occurs. Consequently, Schnipke (1995) proposed determining threshold values that separate the bimodal response time distributions. This method is known as the *visual inspection (VI) method* (e.g., Wise, 2006) and has been frequently applied in previous research (Wise, 2019). The VI method can be used if the response time distribution is bimodal. The only prerequisite is that rapid guessing actually takes place and that the number of observations is sufficiently large, making it possible to identify the threshold value solely from the bimodal distribution.[3] Alternative approaches have been developed to overcome VI's limitation of requiring bimodal response time distributions. A recent overview by Wise (2019) provides operational rules for applying the threshold identification methods listed in Table 4.

**Table 4:**
Threshold identification methods (see Wise, 2019, for details)

| Threshold Method | Brief Description |
|---|---|
| Visual Inspection (VI) | Time between the two modes of response time distribution (only defined for bimodal distributions) |
| Visual Inspection with conditional response accuracy (VITP) | Time after response accuracy fluctuated around the chance level of accuracy (not defined if accuracy higher than chance level) |
| Modified VITP (VITP-M) | Similar to VITP, but use VI when VITP is not defined and response time distribution is bimodal |
| Cumulative Proportion (CUMP) | Time after the cumulative proportion correct is above the chance level of accuracy |
| Change in Information (ChInf) | Time after item-total correlation exceeds 0.20 for multiple seconds |
| Change in Information and Accuracy (ChIA) | Average of ChInf and an additional time threshold, after which response accuracy shows a sustained increase towards the value characteristic of solution behavior |
| Normative Threshold 10% (NT10), 15% (NT15), 20% (NT20) | 10 (15 or 20) percent of the average time measure (with an upper limit of 10, 15 or 20 seconds) |

The methods incorporating the use of response accuracy (VITP and VITP-M) or cumulated response accuracy (CUMP) require that the proportion correct for rapid guessing differs from the chance level and can not be applied equally well to items of all difficulty levels / to all scores for polytomous items (with multiple thresholds). Using the item-total correlation (ChInf and ChIA) adds additional complexity, as the item-total correlation

[3] The VI method identifies the upper bounds of the effect of engagement (see also Lee & Jia, 2014), since the overlap of response time distributions of engaged and disengaged responses is not considered.

cannot necessarily be estimated very stably using the small number of responses given in a particular response time segment. Accordingly, Wise (2019) emphasizes that *the estimation of response time thresholds is not an exact science* and that *goals about threshold identification may vary with our measurement needs* (p. 335).

When comparing rapid guessing rates across groups of unequal size, the required sample size for threshold identification becomes an additional selection criterion (see Table 3). In particular, the response accuracy conditional on response time fluctuates by chance in smaller samples. Defining a *common threshold* for every item (e.g., 3 sec) is a more basic option that could be applied (Kong, Wise, & Bhola, 2007). However, if test-takers complete the test with different speeds, a common threshold would bias the comparisons in the direction of lower rapid guessing rates in conditions with overall slower test-taking speeds. Following this reasoning, the *normative threshold method* (*NT*) is considered more appropriate for our intended use of the thresholds. This method identifies the time threshold using a maximum value of $x$ seconds (e.g., $x = 10$), as $x$% of the average time required to respond to an item. In addition, the maximum value of the threshold is restricted to be $x$ seconds. The NT15 threshold, for instance, is obtained by taking 15 % of the mean response time for a particular item, if it is below 15 sec. (or 15 sec. otherwise). By definition, NT thresholds can be computed without visually inspecting the response time distribution, which must not necessarily be bimodal. Moreover, the NT method can be applied to smaller sample sizes, as only the average response time needs be estimated from the data. The method was suggested by Wise and Ma (2012) as a method for establishing time thresholds for item pools. The thresholds produced via the NT method have been found to be conservative (Wise, 2019), which can be compensated for by choosing a higher percentage value (e.g., *NT15* or *NT20*).

**Operationalization of rapid guessing rates**

Item-specific response time threshold $T_i$ obtained from one of the methods listed in Table 4 can be used to identify test-taker $j$'s responses to item $i$ that exhibit rapid guessing instead of *solution behavior*:

$$SB_{ij} = \begin{cases} 1 & \text{if } TM_{ij} \geq T_i \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

The time measure $TM_{ij}$ in Equation 1 used in this paper refers to the differences between responses as described above.

The responses classified as solution behavior at the item level are aggregated, either across all items in a particular test for each test-taker or across all test-takers for a considered factor for a particular item. Aggregated across all $n$ test-takers, the *response time fidelity* (RTF, Wise, 2006) can be computed for each item $i$

$$RTF_i = \frac{\sum_j SB_{ij}}{n}, \tag{2}$$

and *rapid guessing rates* for single items can be described as $1 - RTF_i$. Aggregated across all $k$ items in a test, *response time effort* measures (RTE, Wise & Kong, 2005) can be computed for each test-taker $j$

$$RTE_j = \frac{\sum_i SB_{ij}}{k}, \tag{3}$$

and the *rapid guessing rates* across factors can be described as $1 - RTE_j$. Confidence intervals, simple mean value comparisons and effect sizes for the rapid guessing rates are reported.[4] Taking the clustered data structure in group testing into account, standard errors, z-values and p-values are computed with the t-test for clustered data from the R package `Hmisc` (Harrel et al., 2020).

**Comparability of rapid guessing rates**

Previous research has investigated the relationship between thresholds identified using different methods (e.g., Kong et al., 2007; Wise, 2019), with the agreement in thresholds identified using different criteria interpreted as a cross-validation (e.g., Lindner et al., 2019). In this paper, we extend this strategy by conducting a robust comparison of rapid guessing rates across modes and settings while acknowledging different time intensities and group sizes across the conditions.

The following strategy was applied to achieve comparability in the rapid guessing rates: Firstly, the different threshold identification methods described in Table 4 were applied to un-proctored individual online assessment, as the condition with the largest sample size. Secondly, the normative threshold method was applied also to the data gathered in the proctored group testing conditions. Spearman rank correlations and descriptive statistics for the thresholds are reported to illustrate the NT approach's ability to adapt to item-specific and factor-specific differences in response times. Thirdly, rapid guessing rates were compared across modes and settings using the condition-specific NT threshold. This procedure makes it possible to investigate the robustness of all findings with respect to the choice of the threshold method. When necessary, rapid guessing rates are reported using condition-specific NT thresholds for all conditions and using VI thresholds for un-proctored individual online assessment. If the pattern of results is identical, only the rapid guessing rates obtained with the NT20 thresholds are used for plots.

It is important to note that thresholds are used in this paper to compare rapid guessing rates across conditions and between groups. We do not interpret the absolute amount of rapid guessing, as this would require more certainty about the classification of solution behavior.

[4] The results are reported without sampling weights to ensure a consistent presentation of the results for randomized and non-randomized factors.

**Table 5:**
Descriptive statistics for the threshold identification methods applied to 30 ICT literacy and 29 science items.

| Threshold Method | ICT | | | Science | | |
|---|---|---|---|---|---|---|
| | # | Mean | SD | # | Mean | SD |
| *Un-proctored Individual Online Assessment* | | | | | | |
| Visual Inspection (VI) | 30 | 8.12 | 3.13 | 29 | 15.97 | 6.28 |
| Visual Inspection with Response Accuracy (VITP) | 25 | 10.86 | 4.86 | 22 | 20.43 | 10.66 |
| Modified VITP (VITP-M) | 30 | 10.58 | 4.63 | 29 | 19.67 | 9.90 |
| Cumulative Proportion (CUMP) | 18 | 8.95 | 5.53 | 22 | 9.02 | 7.55 |
| Change in Information (ChInf) | 30 | 9.53 | 4.95 | 29 | 10.93 | 5.99 |
| Change in Information and Accuracy (ChIA) | 30 | 9.27 | 4.48 | 29 | 14.78 | 6.86 |
| Normative Threshold 10% (NT10) | 30 | 3.67 | 0.95 | 29 | 5.81 | 2.09 |
| Normative Threshold 15% (NT15) | 30 | 5.51 | 1.42 | 29 | 8.71 | 3.13 |
| Normative Threshold 20% (NT20) | 30 | 7.35 | 1.89 | 29 | 11.61 | 4.17 |
| *Proctored Group Digital Assessment* | | | | | | |
| Normative Threshold 10% (NT10) | 30 | 3.65 | 1.25 | 29 | 5.96 | 2.08 |
| Normative Threshold 15% (NT15) | 30 | 5.47 | 1.88 | 29 | 8.95 | 3.12 |
| Normative Threshold 20% (NT20) | 30 | 7.30 | 2.51 | 29 | 11.93 | 4.16 |
| *Proctored Group Paper Assessment (Digital Pens)* | | | | | | |
| Normative Threshold 10% (NT10) | 30 | 4.40 | 1.22 | 29 | 6.56 | 1.85 |
| Normative Threshold 15% (NT15) | 30 | 6.59 | 1.83 | 29 | 9.84 | 2.78 |
| Normative Threshold 20% (NT20) | 30 | 8.79 | 2.44 | 29 | 13.12 | 3.71 |

## Results

### Thresholds

All methods listed in Table 4 were applied to identify thresholds for all items in the ICT and science tests using the data from the un-proctored individual online assessment. As shown in Table 5, the VI method worked[5], as all items had bimodal response time distributions. VITP and CUMP could only be applied to a subset of items. The methods incorporating visual inspection (VI, VITP and VITP-M) resulted in higher thresholds on average compared to the methods incorporating response accuracy information (CUMP, ChInf and ChIA).

The NT thresholds reflect the scaled differences in the average response time across administrations. The results in Table 5 confirm that the NT10 and NT15 thresholds are most conservative (i.e., resulted in the shortest thresholds). Moreover, the NT thresholds for proctored group paper assessment with digital pens are systematically larger than the NT thresholds for digital assessment (proctored group and un-proctored individual

[5]The Inter-rater agreement between three raters for the VI thresholds was 0.86 (ICC2 statistic, see Shrout & Fleiss, 1979) only for one item no threshold was identified by one rater.

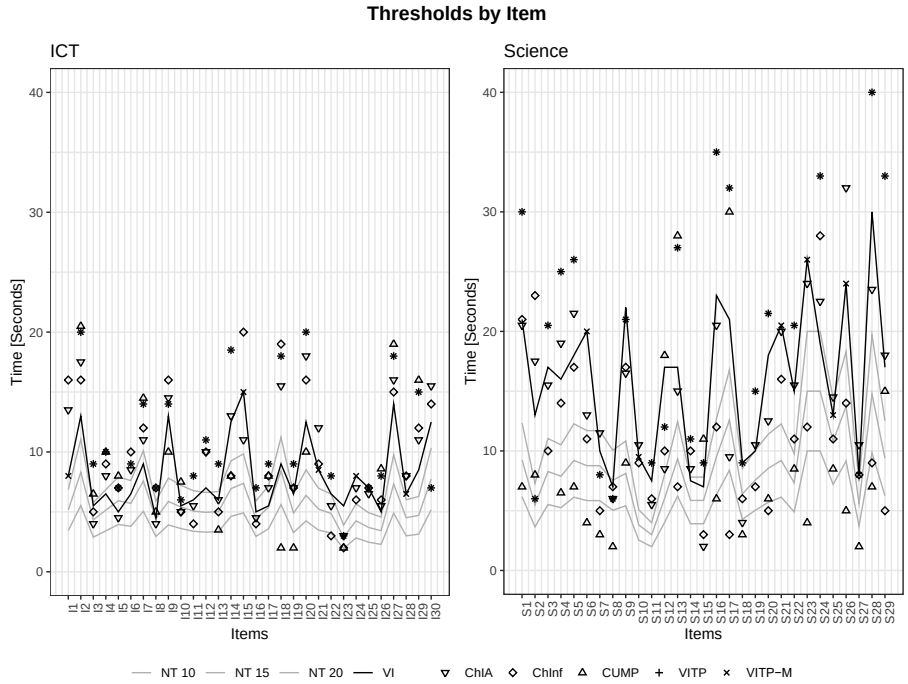



**Figure 2:**
Response time thresholds identified using the different methods (see Table 4) for un-proctored individual online assessment

online assessment). As shown in the upper part of Figure 2, the VI thresholds are above NT10 for all items and above NT15 and NT20 for most items.

No systematic general trend of increasing thresholds across item positions within tests can be observed (see Figure 2). Thresholds for the science test are higher overall and exhibit larger differences between items compared to the thresholds for the ICT test (see also mean and standard deviation in Table 5).

The number of omitted responses is highest in both domains for the proctored group paper assessment (see upper part of Figure 3). For the proctored group digital assessment and the un-proctored individual online assessment, the frequencies of omitted responses across items are similar across conditions.

The lowest number of not reached items in both domains was observed for proctored group digital assessment. However, there is a noticeable difference between the two tests in terms of the number of not reached items (the science test was found to be severely speeded). Even in the proctored group digital assessment, more than half of students do

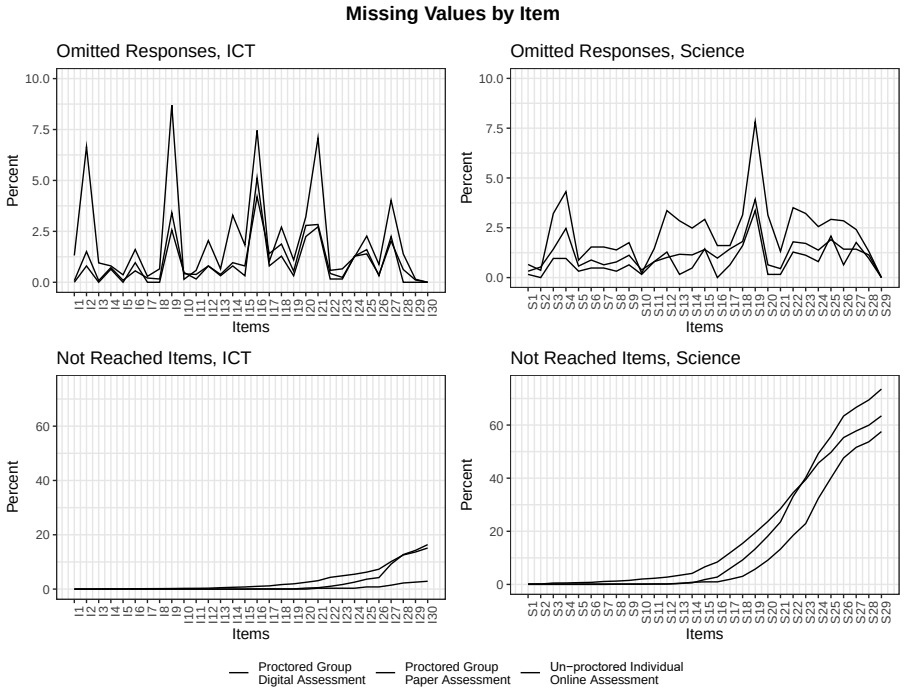



**Figure 3:**
Omitted responses and not reached items for the two investigated domains

not reach the last item (see lower part of Figure 3).

Rank correlations for all thresholds in the un-proctored individual online assessment and correlations between thresholds derived using the NT method for the proctored group testing conditions are shown in Table 6. The thresholds are generally highly correlated, with the exception of ChInf. The VI thresholds derived for un-proctored individual online assessment are correlated between 0.82 and 0.85 with the NT thresholds derived for the group testing conditions. The rank correlation between the NT thresholds for proctored group digital assessment and proctored group paper assessment with digital pens is 0.90, slightly lower than the estimated correlations with the NT threshold from un-proctored individual online assessment.

**Table 6:**
Spearman rank correlation of thresholds identified in un-proctored individual online assessment, proctored group digital and paper assessment

| | Un-proctored individual online assessment | | | | | Proctored Group | |
|---|---|---|---|---|---|---|---|
| **Threshold Method** | **VI** | **NT** | **VITP** | **VITP-M** | **ChInf** | **Digital NT** | **Paper NT** |
| VI | | | | | | 0.85 | 0.82 |
| NT | 0.83 | | | | | 0.94 | 0.95 |
| VITP | 0.83 | 0.75 | | | | 0.75 | 0.70 |
| VITP-M | 0.83 | 0.76 | 1.00 | | | 0.75 | 0.70 |
| ChInf | 0.55 | 0.43 | 0.43 | 0.44 | | 0.52 | 0.47 |
| ChIA | 0.86 | 0.76 | 0.76 | 0.75 | 0.76 | 0.79 | 0.73 |

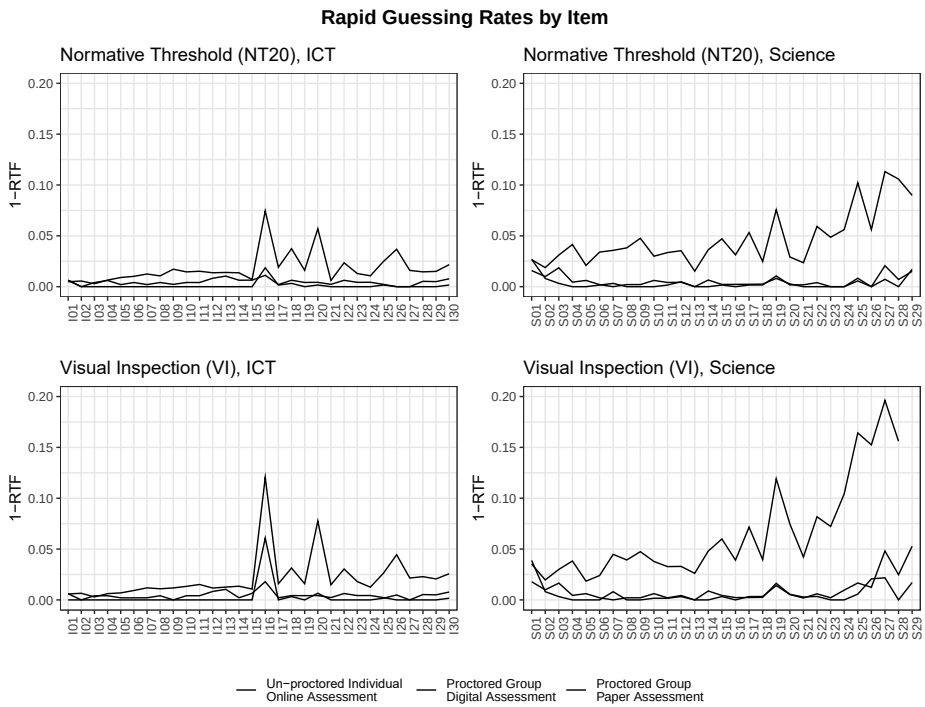



**Figure 4:**
Rapid Guessing Rates by Item using Condition-specific NT20 Threshold and using the VI Threshold from Un-proctored Individual Online Assessment

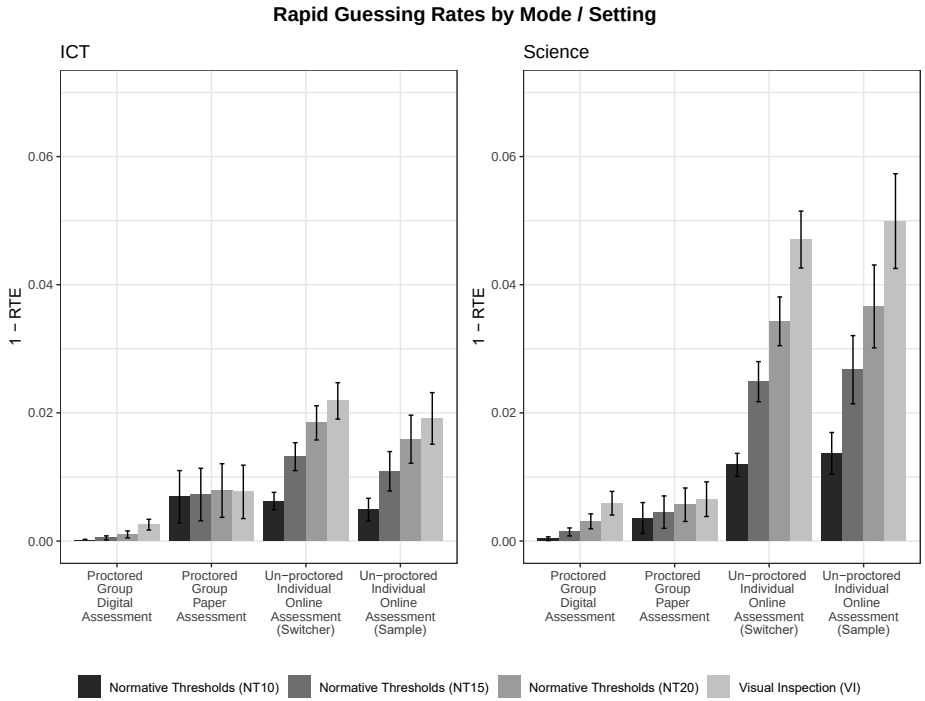



**Figure 5:**
Rapid guessing rates ($1 - RTE$) for the two investigated domains by administration mode and test setting for selected threshold methods

### Mode and setting effects

In this section, we report the results for comparisons identified by random assignment in the experimental study design. Rapid guessing could be identified using the normative thresholds (NT) and the thresholds derived via visual inspection (VI). Figure 5 shows the rapid guessing rates for both domains using the three condition-specific NT thresholds and the VI threshold obtained for the un-proctored individual online assessment. It is possible to compare rapid guessing rates across conditions for each threshold method. Within each administration condition, rapid guessing rates increase as the thresholds increase. Only for the ICT test administered as a proctored group paper assessment are the rapid guessing rates mostly unaffected by the choice of the threshold method.

Table 7 presents the univariate statistics for testing Hypotheses H1 (mode effect) and H2 (setting effect). For ICT, the expectation that computer-based testing will lead to lower rapid guessing rates compared to paper-based assessment was confirmed. However, a similar mode effect in rapid guessing rates was not found for the science test for all threshold methods. The effect size measures provided in Table 7 indicate small mode effects for ICT. For science, the mode effect is not robust against the choice of the threshold and only statistically significant for shorter thresholds (not for VI and NT20).

**Table 7:**
Confidence intervals, z-statistics and effect sizes for the differences in rapid guessing rates across modes (H1) and settings (H2)

| Domain | Hypothesis | Threshold | CI (lower) | CI (upper) | z | p | Cohen's d |
|---|---|---|---|---|---|---|---|
| ICT | H1 | NT10 | 0.002 | 0.010 | 3.30 | 0.001 | 0.205 |
| ICT | H1 | NT15 | 0.002 | 0.010 | 3.29 | 0.001 | 0.204 |
| ICT | H1 | NT20 | 0.003 | 0.010 | 3.26 | 0.001 | 0.203 |
| ICT | H1 | VI | 0.001 | 0.008 | 2.32 | 0.020 | 0.144 |
| ICT | H2 | NT10 | 0.002 | 0.009 | 3.14 | 0.002 | 0.132 |
| ICT | H2 | NT15 | 0.006 | 0.018 | 4.05 | 0.000 | 0.170 |
| ICT | H2 | NT20 | 0.010 | 0.024 | 4.58 | 0.000 | 0.192 |
| ICT | H2 | VI | 0.011 | 0.026 | 4.77 | 0.000 | 0.200 |
| Science | H1 | NT10 | 0.001 | 0.006 | 2.91 | 0.004 | 0.181 |
| Science | H1 | NT15 | 0.001 | 0.006 | 2.53 | 0.011 | 0.157 |
| Science | H1 | NT20 | -0.000 | 0.005 | 1.78 | 0.075 | 0.110 |
| Science | H1 | VI | -0.003 | 0.005 | 0.28 | 0.777 | 0.021 |
| Science | H2 | NT10 | 0.007 | 0.017 | 4.55 | 0.000 | 0.191 |
| Science | H2 | NT15 | 0.015 | 0.033 | 5.33 | 0.000 | 0.224 |
| Science | H2 | NT20 | 0.021 | 0.043 | 5.83 | 0.000 | 0.245 |
| Science | H2 | VI | 0.029 | 0.054 | 6.61 | 0.000 | 0.278 |

*Note. Visual inspection (VI) based on un-proctored individual online assessment, normative thresholds (NT) mode- and setting-specific.*

Regarding the setting effect, the results are in the expected direction and consistent across the threshold method. Significant setting effects are found for both ICT and science, with a small effect size.

### Gender effects and effects of students' first language

In the following section, non-experimental results are reported that incorporate comparisons of students based on observed covariates. Table 8 contains the comparisons addressing hypotheses H3 (gender effect) and H4 (first language). Hypothesis H3 predicted gender differences in the form of lower rapid guessing rates among female students. Averaged across all conditions, we find gender differences in rapid guessing rates, with male students exhibiting higher rapid guessing rates. The gender effects are robust

against the choice of the threshold method.

**Table 8:**
Confidence intervals, z-statistics and effect sizes for the differences in rapid guessing rates between male and female students (H3) and students whose first language is German vs. another language (H4)

| Domain | Hypothesis | Threshold | CI (lower) | CI (upper) | z | p | Cohen's d |
|---|---|---|---|---|---|---|---|
| ICT | H3 | NT10 | 0.006 | 0.011 | 6.70 | 0.000 | 0.187 |
| ICT | H3 | NT15 | 0.010 | 0.017 | 7.30 | 0.000 | 0.193 |
| ICT | H3 | NT20 | 0.012 | 0.021 | 7.16 | 0.000 | 0.195 |
| ICT | H3 | VI | 0.012 | 0.022 | 6.81 | 0.000 | 0.191 |
| ICT | H4 | NT10 | 0.006 | 0.011 | 6.70 | 0.000 | 0.171 |
| ICT | H4 | NT15 | 0.010 | 0.017 | 7.30 | 0.000 | 0.296 |
| ICT | H4 | NT20 | 0.012 | 0.021 | 7.16 | 0.000 | 0.332 |
| ICT | H4 | VI | 0.012 | 0.022 | 6.81 | 0.000 | 0.347 |
| Science | H3 | NT10 | 0.010 | 0.017 | 6.93 | 0.000 | 0.219 |
| Science | H3 | NT15 | 0.016 | 0.029 | 6.51 | 0.000 | 0.217 |
| Science | H3 | NT20 | 0.016 | 0.033 | 5.83 | 0.000 | 0.195 |
| Science | H3 | VI | 0.015 | 0.035 | 5.06 | 0.000 | 0.172 |
| Science | H4 | NT10 | 0.010 | 0.017 | 6.93 | 0.000 | 0.288 |
| Science | H4 | NT15 | 0.016 | 0.029 | 6.51 | 0.000 | 0.365 |
| Science | H4 | NT20 | 0.016 | 0.033 | 5.83 | 0.000 | 0.389 |
| Science | H4 | VI | 0.015 | 0.035 | 5.06 | 0.000 | 0.389 |

*Note. Visual inspection (VI) based on un-proctored individual online assessment, normative thresholds (NT) mode- and setting-specific.*

Figure 6 illustrates the descriptive finding that the observed differences in rapid guessing rates between male and female students are only clearly pronounced in the un-proctored individual assessment. The rapid guessing rates for male and female students are similar in both domains in proctored group assessment, either digital or paper-based, with slightly higher rapid guessing rates for males in the paper assessment condition.

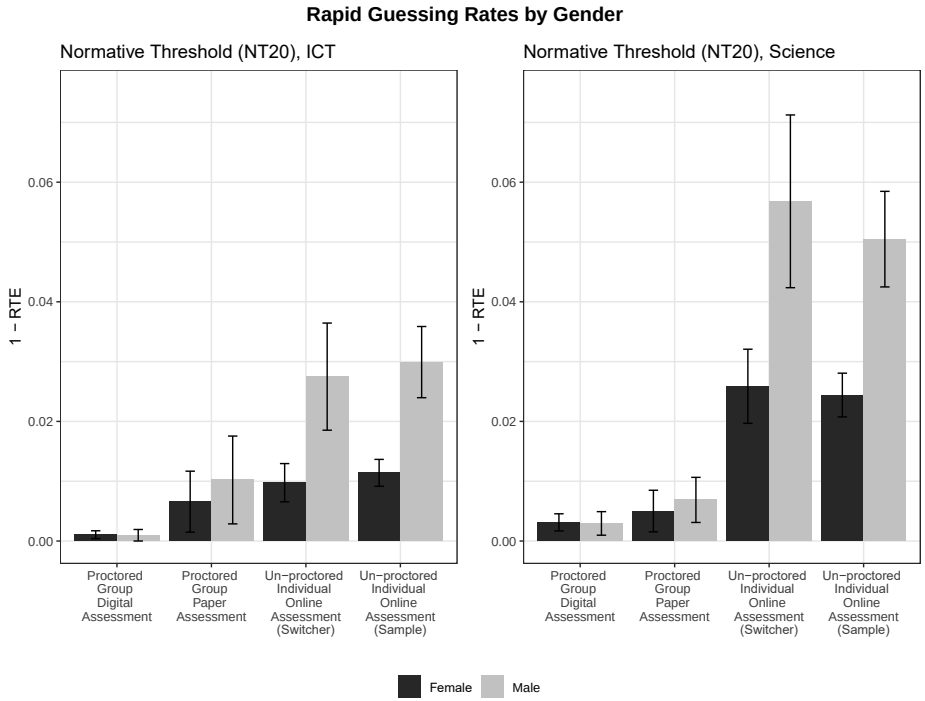



**Figure 6:**
Rapid guessing rates ($1 - RTE$) for the two investigated domains by administration mode and test setting for female and male students

The results regarding Hypothesis H4, shown in Table 8, confirm the lower rapid guessing rates for students whose first language is German. Although only a small number of students differed along this variable (see Table 3), the effect sizes (see Table 8) are small to medium. Just like for the gender effect, the overall difference in rapid guessing rates between students whose first language is German vs. another language is only clearly present for un-proctored individual online assessment (see Figure 7). However, the differences for proctored group testing were in the same direction as for un-proctored online assessment.

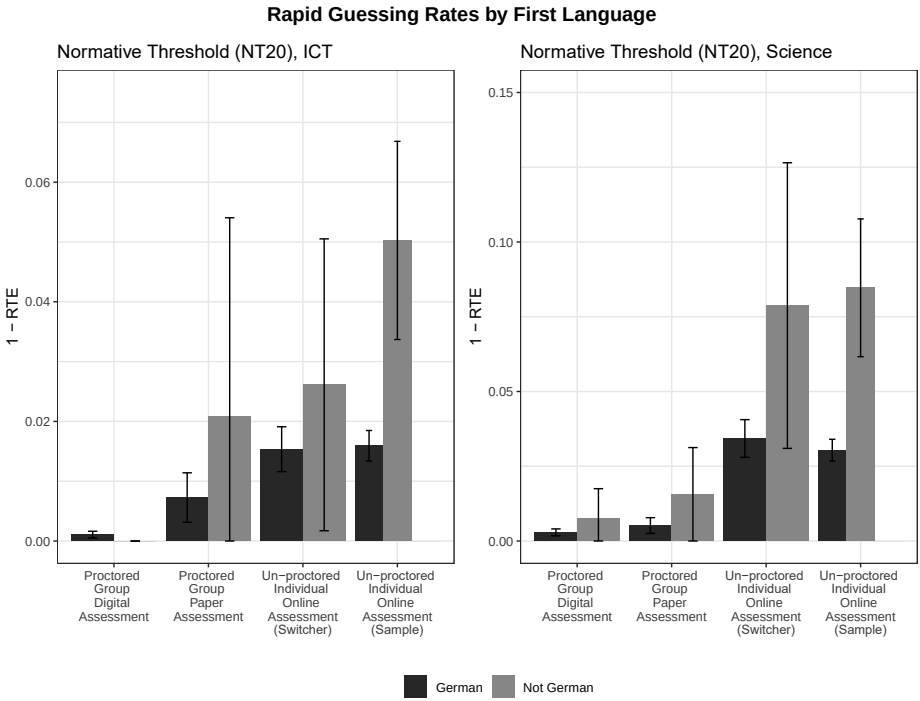



**Figure 7:**
Rapid guessing rates ($1 - RTE$) for the two investigated domains by administration mode and test setting for students whose first language is German vs. another language

### Domain differences and position effects

The results reported in the remaining two sections are ensured by random assignment, but are particularly specific to the used instruments. Rapid guessing rates aggregated across the different assessment conditions were computed to investigate domain differences (H5) and effects of the domain's position within the test booklet (H6). The results are shown in Table 9.

The findings aggregated over all administration conditions confirm the presence of higher rapid guessing rates for the science test (H5) and for the test administered in the second position (H6). This conclusion is robust against the choice of the threshold method, with effect sizes increasing for larger thresholds.

**Table 9:**
Confidence intervals, z-statistics and effect sizes for the differences in rapid guessing rates between domains (H5), domain positions (H6), online preference (H7) and dropout (H8)

| Hypothesis | Threshold | CI (lower) | CI (upper) | z | p | Cohen's d |
|---|---|---|---|---|---|---|
| H5 | NT10 | 0.003 | 0.008 | 4.20 | 0.000 | 0.100 |
| H5 | NT15 | 0.006 | 0.015 | 4.72 | 0.000 | 0.122 |
| H5 | NT20 | 0.009 | 0.020 | 5.18 | 0.000 | 0.136 |
| H5 | VI | 0.016 | 0.029 | 6.89 | 0.000 | 0.188 |
| H6 | NT10 | 0.002 | 0.008 | 3.77 | 0.000 | 0.095 |
| H6 | NT15 | 0.005 | 0.014 | 4.14 | 0.000 | 0.108 |
| H6 | NT20 | 0.007 | 0.018 | 4.30 | 0.000 | 0.115 |
| H6 | VI | 0.009 | 0.022 | 4.52 | 0.000 | 0.124 |
| H7 | NT10 | -0.003 | 0.004 | 0.48 | 0.631 | 0.016 |
| H7 | NT15 | -0.005 | 0.007 | 0.30 | 0.766 | 0.010 |
| H7 | NT20 | -0.006 | 0.008 | 0.29 | 0.772 | 0.010 |
| H7 | VI | -0.007 | 0.010 | 0.25 | 0.802 | 0.009 |
| H8 | NT10 | 0.000 | 0.012 | 1.99 | 0.046 | 0.115 |
| H8 | NT15 | 0.000 | 0.019 | 2.00 | 0.045 | 0.111 |
| H8 | NT20 | -0.000 | 0.022 | 1.91 | 0.057 | 0.102 |
| H8 | VI | -0.005 | 0.020 | 1.19 | 0.233 | 0.063 |

*Note. Visual inspection (VI) based on un-proctored individual online assessment, normative thresholds (NT) mode- and setting-specific.*

**Setting preference and dropout**

Two groups of students participated in the un-proctored individual online assessment. Hypothesis H7 (Preference) investigates the extent to which students preferred this setting over proctored group testing. As Table 9 reveals, there was no statistically significant difference in rapid guessing rates between the two subsamples who completed the test as an un-proctored individual online assessment. This rejection of H7 is not affected by the choice of the threshold method.

In the present study, aborted test sessions (dropout) only occurred in the un-proctored individual online assessment. Hypothesis H8 (Dropout) predicted higher rapid guessing rates for incomplete test sessions. However, as summarized in Table 9, directly comparing the rapid guessing rates for students with versus without dropout does not confirm the hypothesis of higher rapid guessing rates for incomplete test sessions (i.e., H8 is rejected), regardless of which threshold method is used.

## Summary and discussion

Digital assessment and paper-based assessment with digital pens can be used to quantify rapid guessing rates. In this paper, we compared rapid guessing rates across administration modes and test settings using response time measures extracted from log data. For this purpose, we altered the definition of response times to measure time differences between subsequent answers, which are available for paper assessment with digital pens as well as for digital assessment. As illustrated in this paper, the procedure allows the quantification of rapid guessing rates in different administration modes or test settings. The method can also be used to identify rapidly guessed responses as a first step, which can be combined with further treatments and psychometric methods (e.g., filtering or mixture modeling).

The identification of rapid guessing using response times required defining threshold values for separating solution behavior from rapid guessing behavior. Visual inspection of the response time distribution worked for all items in the two domains but was only applied to the un-proctored individual online assessment data. Since only smaller samples were available for the two conditions involving proctored group testing, the normative threshold method with condition-specific thresholds was primarily used. Two empirical findings underpin this decision to use normative thresholds to compare rapid guessing rates. Firstly, we found high correlations between the normative thresholds from the proctored group testing conditions with both the normative thresholds and thresholds identified using visual inspection from the un-proctored individual online assessment. Secondly, we found mean differences in the normative thresholds for the proctored group assessment in the direction of slower test-taking in the paper compared to digital assessment. This finding is in line with a previous study (Kroehne, Hahnel, & Goldhammer, 2019) and illustrates that the normative threshold method is capable of adjusting for mode-specific speed differences. In summary, our procedure illustrates how comparisons of rapid guessing rates between conditions with different sample sizes are possible by combining different threshold identification methods. In practice, one of the advantages of using response times for identifying rapid guessing is, that once the thresholds are identified in a large sample, the gained thresholds can be used for smaller groups and even for responses from individual cases.

The first substantial research question, whether rapid guessing rates are affected by mode and setting, can be answered by this paper's empirical findings. We found evidence of the hypothesized mode effect, with lower rapid guessing rates for digital assessment across all threshold methods in one of the two domains. In addition to the higher rapid guessing rates for paper-based assessment, we also observed higher rates of omitted responses in paper-based assessment (compared to digital assessment in proctored group testing). We also confirmed the expected setting effect, with higher rapid guessing rates for un-proctored individual online assessment. Both findings are relevant for assessment programs considering changing their administration mode or setting, as potential differences in rapid guessing rates could bias comparability and violate measurement

invariance across administration conditions. Since the results differ to some extent between the two investigated domains, more research is necessary to determine the generalizability of the findings. Further research integrating rapid guessing and rapid omission is also needed, as both behaviors might be related to low test-taking effort (as suggested, for instance, by Wise & Gao, 2017).

The second research question concerning differences in rapid guessing rates between different groups of students can also be answered. We found differences in rapid guessing rates in the expected directions for two selected person-level variables. This implies that comparisons across sub-groups could be biased, because rapid guessing behavior, as an indicator of low test-taking effort, can be related to ability (e.g., Goldhammer et al., 2017). In light of this finding, future research might wish to investigate the interaction between person-level variables and test administration mode and test setting. Moreover, it would be interesting to examine whether these findings generalize with respect to different student groups and to study the performance of adjustment methods such as motivation filtering (Wise, 2006) as a way of accounting for inter-individual differences in rapid guessing rates (DeMars et al., 2013).

With respect to the third research question regarding the influence of test properties and test design on rapid guessing rates, we can confirm that in our study setting rapid guessing rates depend on the domain and the position of the test within the test session. As only instruments from two domains were considered, we cannot judge the generalizability of these findings. Moreover, the speededness of the instruments (i.e., the number of items that must be answered within a fixed time limit) might explain the domain differences, because the number of not reached items for the science test was high across all conditions. Further experimental investigations regarding the influence of time limits and test speededness on rapid guessing rates will be necessary to verify this proposed explanation.

The fourth research question regarding the influence of participation status and dropout behavior must be answered negatively. We found no evidence of the expected differences resulting from participation status and dropout behavior. This might be interpreted as an indication that the mechanism underlying dropout behavior differs from rapid guessing behavior.

For geographically dispersed samples, as is often the case for panel studies with multiple waves, unproctored online assessment can be a cost-effective form of data collection for both surveys and cognitive tests. Considering all research questions together, we have shown that the amount of rapid guessing differs across administration modes and test settings, even with identical test designs. This finding contributes to the empirical knowledge about the advantages and disadvantages of different ways of implementing competence assessment.

Concerning the practical transition of assessments from paper-based to computer-based testing, our results illustrate that mode effects (as defined by Kroehne & Martens, 2011) do not only impact item parameters, such as item difficulties (e.g., Buerger et al., 2019):

Additional properties, such as, for instance, information quantity and relevance of text responses (Zehner et al., 2020) and test-taking speed (e.g., Kroehne, Hahnel, & Goldhammer, 2019) can also be affected by the test administration mode. This paper adds to the available evidence regarding mode effects in the form that the missing value patterns and the rapid guessing rates can be affected by test administration mode. For practical assessments, construct-irrelevant differences caused by the mode can be corrected using psychometric methods (e.g., linking as applied by Robitzsch et al., 2020). While digital-based assessments seem to have an advantage by showing lower guessing rates compared to paper-based assessments, the multitude of potential mode effects emphasizes the importance of investigating construct-equivalence across administration modes (see, for instance, Kroehne, Buerger, Hahnel, & Goldhammer, 2019) as an essential pre-requisite for maintaining the interpretation of (adjusted) scores after changing to computer-based testing.

**Limitations and implications for future research**

This study has several limitations that require further research and replications of the results. A replication with larger sample sizes would be beneficial regarding the identification of rapid guessing behavior. In the current study, only the normative threshold method could be applied to all conditions due to the limited sample size. Accordingly, the results rest on the assumption that the normative thresholds capture rapid guessing equally well in different conditions.

There is also a need for further research on the interpretation of fast responses as rapid guessing in paper assessments. In the current study, we modified the well-known approach of classifying responses by focusing on time differences between subsequently answered items rather than response times measured between item onset and response.

Moreover, as we used data from NEPS Starting Cohort 5, future research might wish to explore which additional demographic variables, attitudinal variables, and additional covariates available for this starting cohort are related to rapid guessing behavior. In addition to explaining rapid guessing through personal variables and considering the stability of rapid guessing rates over time, future research could more closely investigate differences across items and domains. In this respect, the correlation between rapid guessing rates at the item level and task characteristics such as item difficulty remains of particular interest. Moreover, in the current study, one of the two tests was administered with time limits that prevented more than half of students from finishing all items. Accordingly, domain differences were confounded by different speededness levels, and further research seems necessary to disentangle potential test-related differences.

For the generalization of the results, it must be taken into account that experimental and non-experimental comparisons have been reported in this paper. Especially with regard to the reported differences for gender and first language, it must be considered that these comparisons were not identified by random assignment. Although manifest group differences in rapid guessing rates were observed, further research is needed to

identify mediating variables.

In this study, we have provided clear evidence of a mode effect in favor of lower rapid guessing rates in digital assessments, at least in one test. This finding might contribute to the empirical evidence concerning the potential consequences of mode changes in (large-scale educational) assessments. It is expected that an understanding of the factors influencing rapid guessing rates will help to establish measurement invariance in large-scale assessments. For example, notification on test-taking engagement, as already investigated by Wise, Kuhfeld, and Soland (2019), could potentially be used to increase the measurement invariance of un-proctored individual online assessment and proctored group digital assessment. The effect of altered test conditions is hypothesized to be the combination of many small effects (e.g., Kroehne & Martens, 2011), and the use of technology-based assessments with digital pens in the present study helped shed additional light on one of these components.

## Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort First-Year Students, doi:10.5157/NEPS:SC5:12.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). Since 2014, the NEPS has been carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

## References

Blommers, P., & Lindquist, E. F. (1944). Rate of comprehension of reading; its measurement and its relation to comprehension. *Journal of Educational Psychology*, *35*(8), 449–473. doi: 10.1037/h0054306

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a Lifelong Process – The German National Educational Panel Study (NEPS)* (Vol. [Special Issue] Zeitschrift für Erziehungswissenschaft: 14.).

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, *62*, 1–9. doi: 10.1016/j.stueduc.2019.04.005

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*(3), 279.

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior*, *29*(5), 1889–1895. doi: 10.1016/j.chb.2013.03.008

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed

books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38. doi: 10.1016/j.edurev.2018.09.003

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and practice in assessment*, *8*, 69—82.

DeMars, C. E., & Wise, S. L. (2010). Can Differential Rapid-Guessing Behavior Lead to Differential Item Functioning? *International Journal of Testing*, *10*(3), 207–229. doi: 10.1080/15305058.2010.496347

Dirk, J., Kratzsch, G. K., Prindle, J. P., Kröhne, U., Goldhammer, F., & Schmiedek, F. (2017). Paper-Based Assessment of the Effects of Aging on Response Time: A Diffusion Model Analysis. *Journal of Intelligence*, *5*(2), 12. doi: 10.3390/jintelligence5020012

Ebel, R. L. (1953). The Use of Item Response Time Measurements in the Construction of Educational Achievement Tests. *Educational and Psychological Measurement*, *13*(3), 391–401. doi: 10.1177/001316445301300303

Finn, B. (2015). *Measuring Motivation in Low-Stakes Assessments: Measuring Motivation in Low-Stakes Assessment* (Research Report No. RR-15-19). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12067.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers No. 133). doi: 10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, *5*(1), 18. doi: 10.1186/s40536-017-0051-9

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., … others (2013). Assessing scientific literacy over the lifespan-a description of the NEPS science framework and the test development. *Journal for educational research online*, *5*(2), 110–138.

Harrel, F., et al. (2020). Hmisc: Harrell miscellaneous, R package version 4.3-1. URL: http://cran.r-project.org/package=Hmisc.

Jerrim, J. (2016). *PISA 2012: How do results for the paper and computer tests compare?* (Tech. Rep.). Department of Quantitative Social Science-UCL Institute of Education, University College London.

Klausch, T., Hox, J. J., & Schouten, B. (2013). Assessing the mode-dependency of sample selectivity across the survey response process. *Statistics Netherlands: The Hague*.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. doi: 10.1177/0013164406294779

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct Equivalence of PISA Reading Comprehension Measured With Paper-Based and Computer-Based Assessments. *Educational Measurement: Issues and Practice*, emip.12280. doi: 10.1111/emip.12280

Kroehne, U., Gnambs, T., & Goldhammer, F. (2019). Disentangling Setting and Mode Effects for Online Competence Assessment. In *Education as a Lifelong Process* (pp. 171–193).

Wiesbaden: Springer VS. doi: 10.1007/978-3-658-23162-0_10

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*(2), 527–563. doi: 10.1007/s41237-018-0063-y

Kroehne, U., Hahnel, C., & Goldhammer, F. (2019). Invariance of the Response Processes Between Gender and Modes in an Assessment of Reading. *Frontiers in Applied Mathematics and Statistics*, *5*, 2. doi: 10.3389/fams.2019.00002

Kroehne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, *14*(S2), 169–186. doi: 10.1007/s11618-011-0185-4

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education*, *58*(3), 196–217.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*(1), 8. doi: 10.1186/s40536-014-0008-1

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The Onset of Rapid-Guessing Behavior Over the Course of Testing Time: A Matter of Motivation and Cognitive Resources. *Frontiers in Psychology*, *10*, 1533. doi: 10.3389/fpsyg.2019.01533

Lüdtke, O., Robitzsch, A., Trautwein, U., Kreuter, F., & Ihme, J. M. (2007). Are There Test Administrator Effects in Large-Scale Educational Assessments? *Methodology*, *3*(4), 149–159. doi: 10.1027/1614-2241.3.4.149

Meyer, J. P. (2010). A Mixture Rasch Model With Item Response Time Components. *Applied Psychological Measurement*, *34*(7), 521–538. doi: 10.1177/0146621609355451

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items. *International Journal of Testing*, 1–19. doi: 10.1080/15305058.2019.1706529

OECD. (2013). *Technical Report of the Survey of Adult Skills (PIAAC).* Paris: OECD Publishing.

OECD. (2016). *Skills Matter: Further Results from the Survey of Adult Skills*. Author. doi: 10.1787/9789264258051-en

OECD. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. Author. doi: 10.1787/9789264281820-en

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education*, *2*(1), 5. doi: 10.1186/s40536-014-0005-4

Pokropek, A. (2016). Grade of Membership Response Time Model for Detecting Guessing Behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325. doi: 10.3102/1076998616636618

Ranger, J., & Kuhn, J.-T. (2017). Detecting unmotivated individuals with a new model-selection approach for Rasch models. *Psychological Test and Assessment Modeling*, *59*, 269–295.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychology Experiments on the Internet* (pp. 89–117). San Diego, CA: Academic Press.

Rios, J. A., & Liu, O. L. (2017). Online Proctored Versus Unproctored Low-Stakes Internet Test Administration: Is There Differential Test-Taking Behavior and Performance? *American Journal of Distance Education*, 1–14. doi: 10.1080/08923647.2017.1258628

Robitzsch, A., Lüdtke, O., Goldhammer, F., Kroehne, U., & Köller, O. (2020). Reanalysis of the German PISA data: A comparison of different approaches for trend estimation with a particular emphasis on mode effects. *Frontiers in Psychology*, *11*(884). doi: http://dx.doi.org/10.3389/fpsyg.2020.00884

Rose, N., Nagy, G., Nagengast, B., Frey, A., & Becker, M. (2019). Modeling Multiple Item Context Effects With Generalized Linear Mixed Models. *Frontiers in Psychology*, *10*, 248. doi: 10.3389/fpsyg.2019.00248

Schnipke, D. L. (1995). Assessing Speededness In Computer-Based Tests Using Item Response Times. In *Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).*

Schnipke, D. L., & Pashley, P. J. (1997). Assessing Subgroup Differences in Item Response Times. In *Paper presented at the Annual Meeting of the American Educational Research Association Annual Meeting (Chicago, IL, March 24-28, 1997).*

Senkbeil, M., Ihme, J. M., Wittwer, J., et al. (2013). The Test of Technological and Information Literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for educational research online*, *5*(2), 139–161.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment. *Applied Measurement in Education*, *26*(1), 34–49. doi: 10.1080/08957347.2013.739453

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi: 10.1037/0033-2909.86.2.420

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying Disengaged Survey Responses: New Evidence Using Response Time Metadata. *Applied Measurement in Education*, *32*(2), 151–165. doi: 10.1080/08957347.2019.1577244

Steimle, J. (2012). *Pen-and-paper User Interfaces: Integrating Printed and Digital Documents*. Heidelberg: Springer.

Ulitzsch, E., Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British*

*Journal of Mathematical and Statistical Psychology*. doi: 10.1111/bmsp.12188

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in Psychometric Population Models for Technology-Based Large-Scale Assessments: An Overview of Challenges and Opportunities. *Journal of Educational and Behavioral Statistics*. doi: 10.3102/1076998619881789

Weitensfelder, L. (2017). Test order effects in an online self- assessment: An experimental study. *Psychological Test and Assessment Modeling*, *59*(2), 229–243.

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education*, *19*(2), 95–114. doi: 10.1207/s15324818ame1902_2

Wise, S. L. (2015). Effort Analysis: Individual Score Validation of Achievement Test Data. *Applied Measurement in Education*, *28*(3), 237–252. doi: 10.1080/08957347.2015.1042155

Wise, S. L. (2017). Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. doi: 10.1111/emip.12165

Wise, S. L. (2019). An Information-Based Approach to Identifying Rapid-Guessing Thresholds. *Applied Measurement in Education*, *32*(4), 325–336. doi: 10.1080/08957347.2019.1660350

Wise, S. L., & Gao, L. (2017). A General Approach to Measuring Test-Taking Effort on Computer-Based Tests. *Applied Measurement in Education*, *30*(4), 343–354. doi: 10.1080/08957347.2017.1353992

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. doi: 10.1207/s15324818ame1802_2

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The Effects of Effort Monitoring With Proctor Notification on Test-Taking Engagement, Test Performance, and Validity. *Applied Measurement in Education*, *32*(2), 183–192. doi: 10.1080/08957347.2019.1577248

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. In *Presented at the Paper presented at the Annual Meeting of the National Council on Measurement in Education (Vancouver, Canada, April 12-16, 2012).*

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage Adaptive Testing Design in International Large-Scale Assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16–27. doi: 10.1111/emip.12226

Zehner, F., Kroehne, U., Hahnel, C., & Goldhammer, F. (2020). PISA Reading: Mode Effects Unveiled in Short Text Responses. *Psychological Test and Assessment Modeling*, *62*(1), 85–105.