PSYCHOMETRIKA—VOL. 87, NO. 2, 593–619
JUNE 2022
https://doi.org/10.1007/s11336-021-09817-7

A RESPONSE-TIME-BASED LATENT RESPONSE MIXTURE MODEL FOR IDENTIFYING AND MODELING CARELESS AND INSUFFICIENT EFFORT RESPONDING IN SURVEY DATA

ESTHER ULITZSCH

IPN–LEIBNIZ INSTITUTE FOR SCIENCE AND MATHEMATICS EDUCATION

STEFFI POHL

FREIE UNIVERSITÄT BERLIN

LALE KHORRAMDEL

BOSTON COLLEGE

ULF KROEHNE

DIPF–LEIBNIZ INSTITUTE FOR RESEARCH AND INFORMATION IN EDUCATION

MATTHIAS VON DAVIER

BOSTON COLLEGE

Careless and insufficient effort responding (C/IER) can pose a major threat to data quality and, as such, to validity of inferences drawn from questionnaire data. A rich body of methods aiming at its detection has been developed. Most of these methods can detect only specific types of C/IER patterns. However, typically different types of C/IER patterns occur within one data set and need to be accounted for. We present a model-based approach for detecting manifold manifestations of C/IER at once. This is achieved by leveraging response time (RT) information available from computer-administered questionnaires and integrating theoretical considerations on C/IER with recent psychometric modeling approaches. The approach a) takes the specifics of attentive response behavior on questionnaires into account by incorporating the distance–difficulty hypothesis, b) allows for attentiveness to vary on the screen-by-respondent level, c) allows for respondents with different trait and speed levels to differ in their attentiveness, and d) at once deals with various response patterns arising from C/IER. The approach makes use of item-level RTs. An adapted version for aggregated RTs is presented that supports screening for C/IER behavior on the respondent level. Parameter recovery is investigated in a simulation study. The approach is illustrated in an empirical example, comparing different RT measures and contrasting the proposed model-based procedure against indicator-based multiple-hurdle approaches.

Key words: careless responses, data screening, response times, item response theory, mixture modeling.

1. Introduction

Research in psychology, educational and social sciences heavily relies on questionnaire data.[1] Careless and insufficient effort responding (C/IER), referring to a "survey response set in which

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-021-09817-7.

Correspondence should be made to Esther Ulitzsch, IPN–Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118Kiel, Germany. Email: ulitzsch@leibniz-ipn.de

[1]Throughout this article, we employ the terms questionnaire, survey, and non-cognitive assessment data interchangeably.

© 2021 The Author(s)

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

a person responds to items without sufficient regard to the content of the items and/or survey instructions" (Huang, Liu, & Bowling, 2015, p. 828), may pose a major threat to data quality, and, as such, to validity of inferences drawn from questionnaire data. C/IE respondents are assumed to quickly proceed through the survey, and instead of providing high-quality data by attentively evaluating the item, retrieving relevant information, and selecting a relevant response, to choose response options that do not reflect the trait to be measured. Careless responses, although not reflecting respondents' trait levels, may not necessarily be random (as in Fig. 1a), but might follow distinct patterns Curran & Denison, 2019; DeSimone, DeSimone, Harms, & Wood, 2018; Kroehne, Buchholz, & Goldhammer, April 2019; Meade & Craig, 2012) such as straight lining (see Fig. 1b), diagonal lining (see Fig. 1c), or alternating extreme pole responses (see Fig. 1d).

When left unconsidered, C/IER can have detrimental effects on conclusions drawn from questionnaire data. These range from introducing systematic variance to—depending on dominant C/IER patterns—both attenuated or inflated associations among constructs of interests (Huang et al., 2015; McGrath, Mitchell, Kim, & Hough, 2010), and distorted psychometric properties such as reliability and factor structure (DeSimone et al., 2018; Huang, Curran, Keeney, Poposki, & DeShon, 2012; Schmitt & Stuits, 1985; Woods, 2006).

Conceptually, C/IER can be understood as a special case of response style behavior. Response styles refer to a systematic response tendency irrespective of the item content (Baumgartner & Steenkamp, 2001). Vast literature exists proposing sophisticated model-based solutions for identifying and modeling response styles (see Böckenholt & Meiser, 2017; Khorramdel, Jeon, & Leigh Wang, 2019, for overviews over current solutions). Usually, in approaches for identifying and modeling response styles, observed responses are allowed to be affected by both the respondents' content trait and their response styles. Under C/IER, in contrast, responses may not be reflective of the respondents' trait levels whatsoever. What is more, response style approaches have commonly been tailored to detecting and modeling specific types of response styles, such as mid point, extreme, or acquiescent response styles. Nevertheless, recent approaches allow for modeling and detecting multiple types of response styles simultaneously (Adams, Bolt, Deng, Smith, & Baker, 2019; Bolt, Lu, & Kim, 2014; Takagishi, van de Velden, & Yadohisa, 2019). Similar to these approaches, researchers may not have presumptions on the specific type of C/IER in their data due to its various possible manifestations.

Previous approaches that specifically aim at detecting C/IER usually support the detection of some types of C/IER behavior, but are insensitive to others (see Curran, 2016; Niessen, Meijer, & Tendeiro, 2016, for overviews and comparisons). In this article, we present a model-based approach for detecting manifold manifestations of C/IER at once. This is achieved by leveraging response time (RT) information available from computer-administered questionnaires and integrating theoretical considerations on C/IER with recent psychometric modeling approaches. We specifically make use of psychometric models that have been developed in the context of low effort on cognitive assessments and adapt them to the case of C/IER in non-cognitive assessments.

In the following, we first briefly review previous procedures for C/IER detection. We then delineate how drawing on recent method developments for detecting low effort on cognitive assessments can assist overcoming some of the limitations of these previous procedures, and present a model-based approach that leverages RTs for detecting and modeling multiple types of C/IER. When questionnaire items are administered as item batteries with multiple questions on one screen, timing data is oftentimes recorded at the screen-level. If further log data such as time-stamped log events are available, item-level RTs can be reconstructed (see Kroehne et al., April 2019; Kroehne & Goldhammer, 2018). To equip researchers with tools for both types of timing data, we first introduce the model-based approach considering RTs on the item level. We then provide an adapted version for RTs aggregated on the screen-level. Parameter recovery is investigated in a simulation study. We illustrate the approach using data from the Programme

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

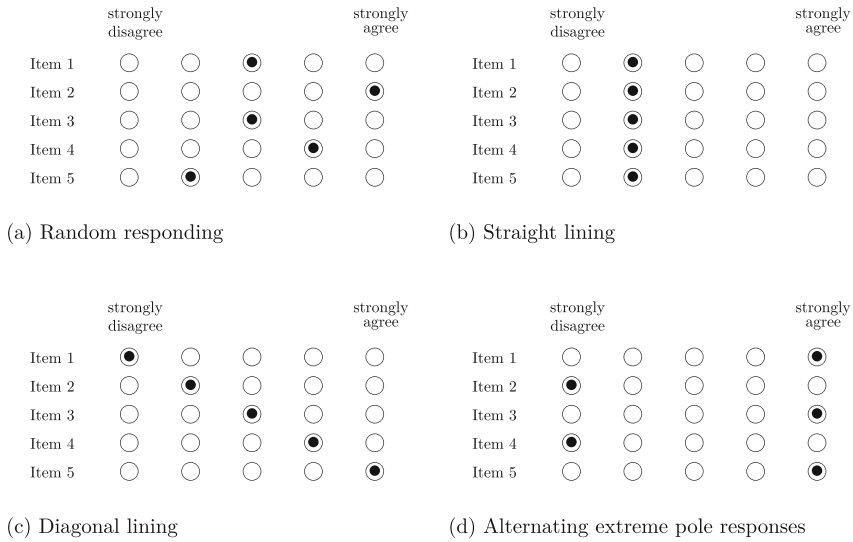


FIGURE 1.
Schematic illustration of different careless and insufficient effort response patterns

for International Student Assessment 2015 (PISA, OECD, 2017) background questionnaire and compare and contrast it against previous indicator-based procedures to C/IER.

2. Previous Approaches for Detecting and Dealing with Careless and Insufficient Effort Responding

2.1. Attention Check/Bogus/Instructional Manipulation Check Items

Oftentimes, researchers administer attention check, instructional manipulation check, or bogus items for drawing inferences on the attentiveness of respondents. These are items that researchers presume attentive respondent will answer in the same way (e.g., disagreement with "All my friends are aliens"; Curran, 2016; Meade & Craig, 2012). A response other than the expected one is taken as an indicator of C/IER. Such items, however, have to be used with precaution since extensive use might confuse attentive respondents (Meade & Craig, 2012). In addition, Curran and Hauser (2019) found that some respondents do not provide the expected answer to those items even after reading them aloud, suggesting that (attentively) reading the item does not inevitably lead to choosing the intended response.

2.2. Response Pattern Analyses

Since C/IER is assumed to result in respondents choosing response options by mechanisms other than according to their trait levels, C/IER should come with response patterns that differ from attentive response patterns. Response–pattern-based approaches to C/IER have been designed for detecting different possible patterns arising from C/IER and involve a) assessing invariability of response patterns, b) assessing individual consistency in responses, c) outlier detection, or d) person-fit statistics. Only recently, machine learning methods have been suggested, which, however, require access to an adequate training data set. We will shortly review response–pattern-based methods based on a prominent example for each. Exhaustive overviews and discussions

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

of other response–pattern-based indicators are given in Curran (2016), Meade and Craig (2012), and Niessen et al. (2016).

2.2.1. Response Invariability The assumption underlying response invariability indicators as, for instance, derived from long string analyses is that C/IE respondents straight line. The long string index is constructed by examining the longest sequence of subsequently occurring identical responses for each respondent (Johnson, 2005). In order to disentangle extreme trait levels from straight lining, this approach requires differently pooled items or to mix items that refer to different traits.

2.2.2. Response Consistency Response consistency indicators are built on the rationale that attentive response patterns are internally consistent, while C/IE response patterns are dominated by random responses (Curran, 2016; Jackson, 1976). A fairly simple measure of response consistency is the even–odd index, given by the within-person correlation between the responses to odd-numbered and even-numbered items belonging to the same scale. When multiple scales are administered, these correlations are averaged across scales. Large values indicate higher within-person consistency of response patterns and, as such, presumably lower levels of C/IER (Curran, 2016; Huang et al., 2012).

2.2.3. Outlier Analysis Since the majority of responses is likely to stem from attentive response processes (see Hauser & Schwarz, 2016; Thomas & Clifford, 2017, for comparisons of different samples), C/IE responses can also be seen as outliers that deviate from typical response patterns. Mahalanobis distance (Mahalanobis, 1936) provides a measure of the multivariate distance between the respondent's response vector and the vector of sample means (Ehlers, Greene-Shortridge, Weekley, & Zajack, 2009; Maniaci & Rogge, 2014; Meade & Craig, 2012). However, Mahalanobis distance can be influenced by too much normality in C/IE responses (arising when respondents randomly choose categories around the mid point, Curran, 2016) and thus performs well for detecting uniformly distributed but fails detecting normally distributed random responses (Meade & Craig, 2012).

2.2.4. Person-Fit Statistics Person-fit statistics can be used to identify response patterns that are unlikely to be observed given an assumed statistical model for item responses. A common person-fit statistic employed for C/IER is the number of Guttman errors in item response theory (IRT) models. In the case of dichotomously scored responses, the number of Guttmann errors is given by the number of item pairs ordered by difficulty with a 0 on the item that is easier to endorse and a 1 on the item that is more difficult to endorse (Meijer, 1994). An extension to ordered polytomous data can be found in Emons (2008). In a similar vein, the l_z^p statistic has been employed (developed by Drasgow, Levine, & Williams, 1985; employed for C/IER by Niessen et al., 2016). This statistic quantifies the likelihood of observing a response vector under a given IRT model. In the case that the l_z^p statistic is very low, the response pattern strongly deviates from what could be expected based on the employed IRT model, and the pattern is classified as inconsistent. In the context of C/IER, person-fit statistics have predominantly been evaluated for identifying uniform random responses (Niessen et al., 2016).

2.2.5. Combining Multiple Response–Pattern-Based Indicators A major limitation of response–pattern-based methods for detecting C/IER is that different measures have been designed for detecting different types of C/IER. For instance, long string analysis has been designed to detect C/IER in terms of straight lining, it is, however, insensitive to other forms of C/IER such as random responding or diagonal lining. Conversely, consistency indicators are insensitive to straight

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

lining since this results in consistency of response patterns (Curran, 2016). Accordingly, when applied to both empirical and simulated data, different methods may show positive, negative, or no agreement (Meade & Craig, 2012; Niessen et al., 2016). Within one survey, different types of C/IER are likely to be present (Meade & Craig, 2012) and need to be dealt with.

Due to the different performance of C/IER measures under different C/IE response patterns, it is commonly recommended to draw conclusions on C/IER based on multiple measures (Curran, 2016; Meade & Craig, 2012; Niessen et al., 2016). Curran (2016) suggested a multiple-hurdle approach that filters out respondents with the most extreme values on each indicator considered based on conservative cut-off values and provided guidelines on how to decide on these cut-off values. Curran (2016), however, also noted that any cut-off setting will to some degree falsely classify attentive as C/IE respondents and/or vice versa.[2] Since in a multiple-hurdle approach multiple cut-off values need to be set, setting cut-offs too high for some indicators and too low for others may result in a complex interplay of different misclassifications that is yet not well understood. Further, depending on the indicators employed, the multiple-hurdle approach may also be affected by the order in which the indicators are considered. This is due to the fact that some indicators, such as Mahalanobis distance or person-fit statistics, are affected by which respondents have been filtered out in preceding hurdles.

As an alternative, multiple measures can be aggregated. Huang et al. (2015), for instance, performed principal component analysis on multiple measures for C/IER and subsequently employed the first factor extracted as a measure of C/IER. Since the first factor extracted from principal component analysis might only capture the dominant C/IE response patterns, this procedure may not control for all types of C/IER.

Note that, again, approaches that combine information from different indicators only support the detection of C/IER patterns to which the employed indicators are sensitive to. As such, these approaches considerably alleviate, but not entirely eradicate the issue of focusing on specific patterns of C/IER behavior.

2.2.6. Employing Supervised Machine Learning To avoid making assumptions concerning the specific types of C/IER patterns—or attentive response patterns, for that matter—Schmidt, and Gnambs (2020) suggested to employ supervised machine learning techniques, with the algorithm being trained on a data set for which it is known which respondents displayed attentive and C/IER behavior, e.g., on a data set stemming from an experiment manipulating instructions on how to approach the questionnaire. Note that the training and test data sets need to be based on the same questionnaire. The approach is limited in that it requires access to an adequate training data set and is based on the assumption that both attentive and C/IE responses follow a structure that is comparable to the respective structures in the training data. This assumption may be violated when respondents do not comply with instructions in the study for obtaining the training data or when respondents being instructed to show C/IER behavior do not behave in a comparable manner to those displaying C/IER behavior in out-of-lab conditions.

2.3. Response Time Analyses

Due to the absence of cognitive processing required for attentively evaluating the item, retrieving relevant information, and selecting a relevant response, short RTs spent on single items can be

[2]Note that both filtering out too few or too many respondents can jeopardize validity of inferences. When too few respondents are filtered out, attentive responses are left confounded with C/IE responses. The effects of this are well studied (DeSimone et al., 2018; Huang et al., 2012; Schmitt & Stuits, 1985; Woods, 2006). When too many respondents are filtered out, valuable information contained in attentive responses is discarded. Further, the filtering procedure may systematically exclude specific subgroups of respondents, e.g., those with high trait levels when all items are worded in the same direction and the threshold for the long string index is set too low. This, too, is likely to impact conclusions on associations between the measured traits.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

seen as indicators of C/IER. Since in computer-administered questionnaires, multiple items are oftentimes displayed on one screen, item-level RTs might not always be at hand and time spent on screen or the survey as a whole may be used as an aggregated proxy (Huang et al., 2012). Previous research utilizing RTs has focused on time spent on screen and on the whole survey, classifying respondents with screen or completion times below a pre-defined threshold as showing C/IER. The thresholds are commonly defined either based on an educated guess on the minimum amount of time required for an attentive response (Huang et al., 2012; Meade & Craig, 2012) or are created using visual inspection of the RT distribution (Kroehne et al., April 2019; Wise, 2017).

One of the major advantages of RT-based over response–pattern-based indicators is that these do not entail presumptions on the specific C/IER patterns. In support of this, Huang et al. (2012) and Huang et al. (2015) found high agreement between RT-based and various other, response–pattern-based indicators. Niessen et al. (2016) compared different methods using both empirical and simulated data. RT-based indicators outperformed response–pattern-based indicators in terms of sensitivity to different C/IER patterns. For validating different measures of C/IER, Meade and Craig (2012) performed factor mixture modeling analyses on facet-score level data and regressed class membership on different indicators of C/IER. Meade and Craig (2012) reported two classes: one class with high and one with low factor loadings, with the latter being interpreted as a response class with high prevalence of C/IER. Meade and Craig (2012) found that completion times could well predict latent class membership, that is, could well predict whether respondents generated facet scores showing high or low association with the construct to be measured.

Nevertheless, Meade and Craig (2012) argued against the use of RTs as single indicators of C/IER due to practical considerations concerning thresholds. While very short RTs or very few time spent on the survey can well be seen as indicators of C/IER, RTs above a set threshold may or may not stem from C/IER. That is, attentive and inattentive RT distributions are likely to overlap, potentially resulting in misclassifications by RT-based threshold methods (Curran, 2016; Meade & Craig, 2012). It has therefore been recommended to apply a sequential approach that classifies C/IER first, based on RTs and second, using response–pattern-based indicators for respondents with longer RTs (Maniaci & Rogge, 2014; Meade & Craig, 2012).

2.4. Dealing with Careless and Insufficient Effort Responding

Previous approaches that deal with C/IER by excluding either responses or cases from the analyses may yield biased conclusions on item and structural parameters (i.e., variances, covariances, or regression coefficients of the traits to be measured), especially when the mechanisms underlying C/IER are not distinct from the traits to be measured (see Köhler, Pohl, & Carstensen, 2017; Pohl, Gräfe, & Rose, 2014; Rose, 2013; Rose, von Davier, & Xu, 2010, for related research on the treatment of nonignorable missing responses). Commonly, respondents with C/IER indicator values falling below a certain cut-off value are eliminated from further analyses. Removing presumable C/IE respondents from further analyses comes with the assumption that the missing values induced by this procedure are ignorable, implying that the constructs to be measured and the mechanisms underlying C/IER are unrelated. Empirical research, however, has found the extent to which C/IER behavior is shown to be related to person characteristics and common constructs of interest such as education (Kim, Dykema, Stevenson, Black, & Moberg, 2018) or personality (Bowling et al., 2016; Huang et al., 2015; Maniaci & Rogge, 2014), rendering this assumption likely to be violated. In this case, filtering can yield biased conclusions (Deribo, Kroehne, & Goldhammer, 2021; Ulitzsch, von Davier, & Pohl, 2020). In addition, C/IER may vary across the assessment and respondents who display C/IER on some parts of the assessment might still provide valid responses to others. Indeed, respondents are more likely to respond randomly toward the middle or end of long questionnaires (Baer, Ballenger, Berry, & Wetter, 1997; Berry et al., 1992), while probably providing valid responses at the beginning. Discarding all responses of

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

respondents who have been identified to show C/IER at some point of the questionnaire thus also discards their valid responses.

3. Approaches for Disengaged Responding Developed in the Context of Cognitive Assessments

The distinction between attentive response behavior and C/IER in non-cognitive assessments shows many parallels to the distinction between solution and disengaged rapid guessing behavior in cognitive assessment. In cognitive assessments, solution behavior is assumed to result in item responses reflecting "what the test taker knows and can do" (Wise, 2017, p. 52). Its counterpart, non-effortful test-taking behavior, is defined as "quickly proceeding through the test without applying [...] knowledge, skills, and abilities" (Wise & Gao, 2017, p. 384). Research on non-effortful test-taking behavior has predominantly focused on rapid guessing as one possible manifestation of non-effortful test-taking behavior. Hence, veins of research on C/IER on the one hand and rapid guessing behavior on the other hand both assume disengaged, respectively inattentive, responding to require less time for its execution than engaged responding (Kroehne et al., April 2019). Sophisticated models for detecting disengaged responding in cognitive assessments have been developed. These could be adapted to non-cognitive assessments and may thereby enhance identification of C/IER.

With the rise of computer-based assessment and the related availability of log data, a rapidly growing body of methods emerged aiming at identification of rapid guessing behavior in cognitive assessments. Primarily, these methods leverage RT data either by defining RT-based scoring rules, with responses associated with RTs below a pre-defined threshold being classified as rapid guesses (Goldhammer, Martens, Christoph, & Lüdtke, 2016; Guo et al., 2016; Lee & Jia, 2014; Wise, Kingsbury, Thomason, & Kong, April 2004; Wise & Ma, April 2012; Wise, Pastor, & Kong, 2009) or by utilizing RT information in mixture modeling approaches, explicating different data-generating processes for RTs and responses associated with solution and (rapid) guessing behavior (Nagy & Ulitzsch, 2021; Schnipke & Scrams, 1997; Ulitzsch et al., 2020; Wang & Xu, 2015).

A recent example for mixture modeling approaches is the speed-accuracy+engagement (SA+E) model developed by Ulitzsch et al. (2020). The SA+E model allows for rapid guessing behavior to vary at the item-by-person level. For the probability of observing a correct response under solution behavior, the SA+E model assumes an IRT model to hold. Probability correct for rapid guesses is assumed to correspond to the probability of guessing correct at chance level. RTs associated with solution behavior are modeled as a function of person speed and the item's time intensity (see also van der Linden, 2007). RTs associated with rapid guessing are assumed not to depend on person or item characteristics and to be shorter than those associated with valid responses. Item-by-person mixing proportions are modeled with a latent response approach as a function of person engagement and item engagement difficulty employing an IRT model. By doing so, the model allows assessing how the tendency to show rapid guessing behavior relates to ability and speed as well as identifying items that are likely to evoke rapid guessing behavior. The SA+E model overcomes major limitations of previously developed approaches for the identification of rapid guessing behavior. First, as a purely model-based approach, the SA+E model does not require setting an RT threshold and allows for overlapping RT distributions, potentially resulting in fewer misclassifications when there is strong overlap of RT distributions associated with solution and rapid guessing behavior. Second, the model allows for rapid guessing behavior to vary across both items and persons and does, as such, not discard valid responses from test takers rapidly guessing only on some parts of the test. Third, since the tendency to show rapid guessing behavior is modeled jointly with ability and speed, the model does not rely on the assumption that ability and the mechanism underlying rapid guessing behavior are unrelated.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

Given the parallels of C/IER and rapid guessing behavior as processes resulting in (possibly) fast responses not reflecting the traits to be measured, it deems promising to build on these recent developments in cognitive assessments to improve the identification of C/IER. Indeed, the literature on the identification of rapid guessing behavior has already stimulated research on the identification of C/IER. Huang et al. (2012), for instance, built their rationale for classifying respondents with completion times below a pre-defined threshold on methods for detecting rapid guessing behavior developed by Wise and DeMars (2006). Nevertheless, concepts and methods developed in the context of cognitive assessments are not directly applicable to the context of non-cognitive assessments. First, methods for identifying rapid guessing behavior have been developed in the context of dichotomously scored responses. Non-cognitive assessments, however, primarily rely on Likert scales for measuring constructs of interests, that is, most often entail (ordered) polytomous response data. Second, and more importantly, methods for non-effortful responding developed in the context of cognitive assessment are concerned with probability correct, the analysis of responses for detecting C/IER, however, is concerned with the chosen response option itself. Third, in non-cognitive assessment data, the relationship between RTs and the trait to be measured is likely to deviate from the linear relationship commonly assumed in models that integrate RT information with IRT models in the context of cognitive assessment. One example is the distance–difficulty hypothesis (Ferrando & Lorenzo-Seva, 2007; Kuncel & Fiske, 1974), assuming that responses take more time when an item is well targeted to the trait of a person, while responses can be given rather quickly when the item thresholds deviate strongly from the trait level of the person. This mimics the fact that statements can be faster endorsed (or not endorsed), when persons are sure of their response. While mixture models for identifying rapid guessing behavior in cognitive assessments are very promising, they need to be adapted to suit the specifics of response behavior in non-cognitive assessments.

4. Proposed Approach

The presented approach for identifying and modeling C/IER behavior is a latent response model for computer-administered questionnaires in which item-level RTs are available. Building on Ulitzsch et al. (2020), the approach a) takes the specifics of attentive response behavior in non-cognitive assessments into account by incorporating the distance–difficulty hypothesis, b) allows for attentiveness to vary on the screen-by-respondent level, c) allows for respondents with different trait and speed levels to differ in their attentiveness, and d) can deal with various response patterns arising from C/IER. The approach assumes that respondents have a constant probability to provide either attentive or C/IE responses on all items on a screen and that respondents do not switch between response modes on a given screen. They can, however, switch from C/IE to attentive responding and vice versa between screens.[3]

In the presented approach, latent attentiveness indicators Δ_{is} denote whether respondent i, $i = 1, \ldots, N$, was attentive when approaching screen s, $s = 1, \ldots S$, ($\Delta_{is} = 1$) or not ($\Delta_{is} = 0$). While the attentiveness status itself is not observable, it is assumed to be associated with different data-generating processes underlying responses and RTs. When approaching a screen attentively, respondents are assumed to generate responses according to their trait levels on all item administered on the screen. When showing C/IER, respondents are assumed to choose response options that do not reflect their trait level. C/IER behavior can have various manifestations, including choosing randomly, marking patterns, such as straight or diagonal lines, or alternating extreme pole responses.

[3]This assumption is reasonable in most computer-administered questionnaires where the number of items presented per screen is not very high. In the PISA 2015 background questionnaire for instance, a median of 4 items was presented per screen (range: 1 to 16 items; OECD, 2017).

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

For reason of simplicity, however, without loss of generality, we present the approach assuming the same number of response options for all items, and that all items measuring a trait are displayed on one screen, and that each screen contains items measuring one trait only. Concerning the relationship between RTs and trait levels, we focus on the distance–difficulty hypothesis as a special case.

4.1. Attentive Behavior

4.1.1. Item Responses When being attentive, respondents are assumed to respond to all items displayed on screen s according to their trait levels. Different IRT models such as the graded response model (Samejima, 2016) or the generalized partial credit model (Muraki, 1997) can be employed to model attentive responses. Here, we present the model with a generalized partial credit model for item responses $x_{ijs} \in \{0. \ldots, K\}$, containing person i's response to the jth item, $j = 1, \ldots, J_s$, displayed on screen s, with K giving the highest possible response category for the considered items. That is, under $\Delta_{is} = 1$, we model the probability of respondent i to choose category k, $k = 1, \ldots, K$, on the jth item displayed on screen s as

$$p(x_{ijs} = k|\Delta_{is} = 1) = \frac{\exp\left(\sum_{l=0}^{k} v_{js}\eta_{is} - b_{jsl}\right)}{\sum_{r=0}^{K} \exp\left(\sum_{l=0}^{r} v_{js}\eta_{is} - b_{jsl}\right)} \quad \text{with} \sum_{l=0}^{0} v_{js}\eta_{is} - b_{jsl} \equiv 0. \quad (1)$$

Here, η_{is} denotes respondents i's level on the sth trait. The parameters b_{jsl} and v_{js} give the lth step difficulty and discrimination of item j measuring latent trait s, respectively.

4.1.2. Response Times When associated with attentive responses, RTs t_{ijs}, denoting the time respondent i spent on the jth item displayed on screen s, are assumed to follow a lognormal distribution governed by the respondent's speed τ_i and the item's time intensity β_{js} (see Ulitzsch et al., 2020; van der Linden, 2007). The distance–difficulty relationship between the respondents' trait levels and their RTs is incorporated following Molenaar, Tuerlinckx, and van der Maas (2015) by regressing log RTs on the absolute weighted distance between the respondent's trait level and the middle step difficulty parameter o_{js}. In the case of four response categories with the three step difficulty parameters b_{js1}, b_{js2}, and b_{js3}, for instance, o_{js} is given by b_{js2}.[4] That is, attentive RTs are modeled as

$$\ln\left(t_{ijs}|\Delta_{is} = 1\right) \sim \mathcal{N}\left(\beta_{js} - \tau_i - \gamma|v_{js}\eta_{is} - o_{js}|, \sigma_A^2\right), \quad (2)$$

with γ denoting the distance–difficulty parameter. Note that different approaches exist for incorporating the distance–difficulty relationship between traits and RTs (see Ranger, 2013, for an overview). Further, the relationship of the distance between the respondent's trait level and the middle step difficulty parameter and RTs must not necessarily be linear but may take other functional forms.

We assume a common residual variance σ_A^2 (see van der Linden, 2007). Note that a common speed factor is assumed across all measured traits, that is, it is assumed that respondents approach all screens to which they respond attentively with the same speed level.

[4]For an uneven number of answer categories, Molenaar et al. (2015) suggested to taken the average of the two middle step difficulties.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

4.2. Careless and Insufficient Effort Behavior

4.2.1. Item Responses Category probabilities that are not reflective of person or item characteristics are estimated for inattentive responses, that is,

$$p(x_{ijs} = k|\Delta_{is} = 0) = \kappa_k \qquad \text{with } \sum_{k=0}^{K} \kappa_k = 1. \tag{3}$$

Note that κ_k gives the marginal probability over all types of C/IER patterns of inattentively choosing category k. Hence, the model is capable of capturing various types of C/IER patterns that all result in no relationship with the measured trait. The model does, however, not allow disentangling groups of respondents with different C/IER patterns.

4.2.2. Response Times In line with mixture modeling approaches for rapid guessing behavior in cognitive assessments (Schnipke & Scrams, 1997; Ulitzsch et al., 2020; Wang & Xu, 2015), we assume RTs associated with C/IE responses to be unaffected by person or item characteristics. Hence, for RTs associated with C/IER, we assume a lognormal distribution governed by a common mean β_C and a common variance σ_C^2:

$$\ln\left(t_{ijs}|\Delta_{is} = 0\right) \sim \mathcal{N}\left(\beta_C, \sigma_C^2\right). \tag{4}$$

It is further assumed that C/IER requires less time than evaluating the item, retrieving relevant information, and selecting a relevant response. This mirrors the assumption of rapid, disengaged guesses in cognitive assessment to be shorter than responses stemming from good faith attempts to solve an item (Wise, 2017). Hence, following Ulitzsch et al. (2020), time intensities for attentive RTs β_{js} are defined as the sum of the C/IER mean β_C and an item-specific, positive offset parameter β_{js}^*. That is,

$$\beta_{js} = \beta_C + \beta_{js}^* \qquad \text{where } \beta_{js}^* \geq 0. \tag{5}$$

The offset parameter β_{js}^* indicates how much longer respondents commonly require to generate an attentive response to the $j$the item presented on screen s rather than showing C/IER. Note that RT distributions associated with attentive and careless responses are allowed to overlap, such that also responses associated with longer RTs may be classified as C/IER.

4.3. Higher-Order Structures

The attentiveness status Δ_{is} of respondent i on screen s is not observable. It, however, determines the measurement properties of the observed responses and associated RTs and thus represents a latent response variable (see Maris, 1995). In line with Ulitzsch et al. (2020), latent response variables Δ_{is} are modeled using a Rasch model as a function of the respondent's attentiveness ψ_i and the screen's attentiveness difficulty ι_s, that is

$$p\left(\Delta_{is} = 1\right) = \frac{\exp\left(\psi_i - \iota_s\right)}{1 + \exp\left(\psi_i - \iota_s\right)}. \tag{6}$$

This supports investigating respondent characteristics associated with low attentiveness and allows identifying screens evoking C/IER behavior. For instance, if screens administered at the end of the survey are more likely to evoke C/IER behavior, they can be expected to have higher attentiveness difficulties.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

Person parameters are assumed to be multivariate normally distributed with mean vector and covariance matrix

$$\boldsymbol{\mu} = \left(\mu_{\psi}, \mu_{\tau}, \mu_{\eta_1}, \ldots, \mu_{\eta_S}\right) \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{\psi} & \sigma_{\psi\tau} & \sigma_{\psi\eta_1} & \ldots & \sigma_{\psi\eta_S} \\ \sigma_{\psi\tau} & \sigma^2_{\tau} & \sigma_{\tau\eta_1} & \ldots & \sigma_{\tau\eta_S} \\ \sigma_{\psi\eta_1} & \sigma_{\tau\eta_1} & \sigma^2_{\eta_1} & \ldots & \sigma_{\eta_1\eta_S} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{\psi\eta_S} & \sigma_{\tau\eta_S} & \sigma_{\eta_1\eta_S} & \ldots & \sigma^2_{\eta_S} \end{pmatrix}. \tag{7}$$

For identifying the model, we set person parameter means to zero. When a generalized partial credit model is employed for attentive responses, the model can be identified by setting trait variances to one. Item parameters are modeled as fixed effects. The proposed model's likelihood can be written as

$$\begin{aligned} \mathcal{L} = \prod_{i=1}^{N}\prod_{s=1}^{S}\Bigg(& p\left(\Delta_{is}=1|\psi_i,\iota_s\right)\prod_{j=1}^{J_s} p\left(x_{ijs}|\eta_{is},v_{js},\boldsymbol{b}_{js}\right)^{\left(1-d^{(x)}_{ijs}\right)} f\left(t_{ijs}|\tau_i,\eta_{is},\beta_{js},\gamma,v_{js},o_{js},\sigma^2_A\right)^{\left(1-d^{(t)}_{ijs}\right)} \\ & + \left(1-p\left(\Delta_{is}=1|\psi_i,\iota_s\right)\right)\prod_{j=1}^{J_s} p\left(x_{ijs}|\boldsymbol{\kappa}\right)^{\left(1-d^{(x)}_{ijs}\right)} f\left(t_{ijs}|\beta_C,\sigma^2_C\right)^{\left(1-d^{(t)}_{ijs}\right)}\Bigg) h\left(\boldsymbol{\psi},\boldsymbol{\tau},\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_S|\boldsymbol{\mu},\boldsymbol{\Sigma}\right), \end{aligned} \tag{8}$$

The first and second component represent the component models for attentive and C/IER behavior, respectively. Here, N, S, and J_s denote the number of respondents, screens (and, as such, traits to be measured), and number of items administered on screen s. The term $h(\boldsymbol{\tau}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_S|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density of the person parameters. The terms for responses and RTs incorporate the assumption of independence of response and RT indicators given the second-order variables of the model. The indicators $d^{(x)}_{ijs}$ and $d^{(t)}_{ijs}$ denote whether or not a response or RT of respondent i to the jth item measuring trait s is available, with $d^{(x)}_{ijs} = 0$ denoting an observed and $d^{(x)}_{ijs} = 1$ a missing response and $d^{(t)}_{ijs} = 0$ denoting an observed and $d^{(t)}_{ijs} = 1$ a missing RT.[5]

4.4. Model Modification for Screen-Level Timing Data

In computer-administered questionnaires, item-level RTs are not always available as these oftentimes require additional, sophisticated data processing (Kroehne et al., April 2019; Kroehne & Goldhammer, 2018). In most surveys (e.g., in PISA, OECD, 2017), only timing data on the screen-level (i.e. aggregated RTs) are available in public use files. To make the approach applicable for a broad audience, we present a model modification for more readily available screen-level timing data. The adapted model is a simplified version of the model for item-level RTs and assumes that respondents approach the assessment either with no or full attentiveness and that, as such, there are no response vectors in which both C/IE and attentive responses occur. Instead of employing a latent response approach, the adapted model assumes a respondent-specific attentiveness probability that is constant across screens and distinct from the traits to be measured,

[5]Item omissions and missing RT information are ignored in the presented model, that is, are assumed to be missing at random (MAR) given the observed data and parameters of the model. This is the current state-of-the-art procedure for dealing with omissions in survey data. Note that omissions tend to occur to a much lesser degree in non-cognitive as compared to cognitive assessments. In the data set considered for the empirical example, for instance, the omission rate was as low as 0.66%.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

that is $p(\Delta_{is} = 1) = \pi_i$. As such, as previous approaches for C/IER behavior, the model is well suited for scanning for C/IER behavior on the respondent level.

We make use of mean time spent on the items presented on screen s, defined as the total screen-level time divided by the number of items and denoted with $\bar{t}_{is}$, as a proxy for item-level RTs, and adapt the measurement model for aggregated RTs associated with attentive responses. Mean time spent on the items presented on screen s associated with inattentive responses is modeled according to Eq. 4, assuming a common mean and variance parameter. To adapt the measurement model for attentive aggregated RTs, we consider a screen- rather than an item-specific time intensity parameter β_s, determining the average time respondents require for providing attentive responses to the items presented on screen s.[6] For considering the distance–difficulty relationship between the respondents' trait levels and their RTs, we average the discrimination and middle step difficulty parameters of the items presented on screen s, taking the screen-level geometric mean of discriminations $v_{\cdot s}$ and the arithmetic mean of middle step difficulties $o_{\cdot s}$. The parameter γ thus gives the average distance–difficulty effect for all items presented on a screen. Hence, the average time respondent i spent on the items presented on screen s is modeled as

$$\ln\left(\bar{t}_{is}|\Delta_{is} = 1\right) \sim \mathcal{N}\left(\beta_s - \tau_i - \gamma|v_{\cdot s}\eta_{is} - o_{\cdot s}|, \sigma_A^2\right). \tag{9}$$

Screen-level time intensity parameters are subject to the constraint

$$\beta_s = \beta_C + \beta_s^* \qquad \text{where } \beta_s^* \geq 0. \tag{10}$$

For simplicity, attentiveness parameters are dropped from the multivariate normal distribution of person parameters. In the adapted model, the mean vector and covariance matrix of person parameters are given by

$$\boldsymbol{\mu} = \left(\mu_\tau, \mu_{\eta_1}, \ldots, \mu_{\eta_S}\right) \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\tau^2 & \sigma_{\tau\eta_1} & \ldots & \sigma_{\tau\eta_S} \\ \sigma_{\tau\eta_1} & \sigma_{\eta_1}^2 & \ldots & \sigma_{\eta_1\eta_S} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\tau\eta_S} & \sigma_{\eta_1\eta_S} & \ldots & \sigma_{\eta_S}^2 \end{pmatrix}. \tag{11}$$

This yields the following likelihood for the adapted model

$$\begin{aligned} \mathcal{L} = \prod_{i=1}^{N} \Bigg(& \pi_i \prod_{s=1}^{S} f\left(\bar{t}_{is}|\tau_i, \eta_{is}, \beta_s, \gamma, v_{\cdot s}, o_{\cdot s}, \sigma_A^2\right)^{\left(1-d_{is}^{(t)}\right)} \prod_{j=1}^{J_s} p\left(x_{ijs}|\eta_{is}, v_{js}, \boldsymbol{b}_{js}\right)^{\left(1-d_{ijs}^{(x)}\right)} \\ & + (1-\pi_i) \prod_{s=1}^{S} f\left(\bar{t}_{is}|\beta_C, \sigma_C^2\right)^{\left(1-d_{is}^{(t)}\right)} \prod_{j=1}^{J_s} p(x_{ijs}|\boldsymbol{\kappa})^{\left(1-d_{ijs}^{(x)}\right)} \Bigg) h\left(\boldsymbol{\tau}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_S|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \end{aligned} \tag{12}$$

[6] Note that due to the log transformations of item-level RTs and average RTs, β_s does not correspond to the mean of the item-level time intensity parameters β_{js}.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

4.5. Prior Distributions

For model estimation, we employ Bayesian estimation techniques. Priors are set in accordance with Ulitzsch et al. (2020). We employ an LKJ prior (Lewandowski, Kurowicka, & Joe, 2009) with shape 1 for the correlation matrix of person parameters $\boldsymbol{\Omega}$, implying a uniform prior distribution for the correlation parameters. Half-Cauchy priors with location 0 and scale 5 are employed for all standard deviations, that is, the standard deviation of attentiveness σ_ψ and speed σ_τ, the residual standard deviation of log attentive RTs σ_A, and the common standard deviation of inattentive RTs σ_C, as well as item discriminations υ_{js}. Diffuse normal priors with mean 0 and standard deviation 10 are employed for each step difficulty b_{jsl}, time intensity offset parameter β^*_{js}, respectively β^*_s, the distance–difficulty parameter γ as well as the common mean β_C. For C/IER category probabilities, we suggest a diffuse Dirichlet prior with $\boldsymbol{\kappa} \sim \text{Dir}(\mathbf{1})$. For attentiveness probabilities π_i in the model with screen-level timing data, we employ a Dirichlet prior, parameterized as $(\pi_i, 1-\pi_i) \sim \text{Dir}(\lambda\pi_{\mathcal{P}}, \lambda(1-\pi_{\mathcal{P}}))$, where $\pi_{\mathcal{P}}$ gives the population-level proportion of attentive respondents and λ is a concentration parameter (see Kemp, Perfors, & Tenenbaum, 2007; Salakhutdinov, Tenenbaum, & Torralba, 2012). Population-level proportions of attentive and C/IE respondents are equipped with a diffuse Dirichlet prior, with $(\pi_{\mathcal{P}}, 1-\pi_{\mathcal{P}}) \sim \text{Dir}(1, 1)$. The concentration parameter λ is equipped with a half-Cauchy prior with location 0 and scale 5.

5. Parameter Recovery

For investigating parameter recovery under realistic conditions, we generated data according to the model for item-level RTs. Data-generating values were chosen to resemble parameter estimates reported in the empirical example below. We considered a scenario with different C/IER patterns—uniform random responding, random responding around the endpoints, straight lining, and diagonal lining (see also Curran & Denison, 2019). This allows illustrating that the proposed approach indeed can deal with various patterns arising from C/IER, as long as C/IE responses do not reflect the trait to be measured and, on average, are not slower than attentive responses. We further investigated the potential loss in accuracy resulting from model simplification and aggregating RT information.

Under the investigated conditions, the data-generating model with item-level RTs yielded good parameter recovery and could deal well with the different simulated C/IER patterns as well as low C/IER rates. The model with screen-level timing data could well recover person parameter variances and correlations, step difficulties, marginal C/IER category probabilities as well as the distance–difficulty parameter. However, population-level C/IER rates were overestimated (median estimated C/IER rate: 9.39%, true C/IER rate: 5%). The misclassification of attentive as C/IE responses when using aggregated RTs was also mirrored in biased estimates of parameters related to the RT measurement model. Further, the loss of information on item-level RT variability resulted in estimates of RT residual variances close to zero. Detailed results of the simulation study are given in the supplementary material.

6. Empirical Example

The empirical example serves a) to illustrate the insights that can be gained on the basis of the presented approach, b) to investigate differences between different measures of aggregated RTs available in large-scale assessment data as well as c) to compare the proposed approach to customary indicator-based procedures.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

We took responses, screen-level timing data, and raw log data from the background questionnaire from PISA 2015 (PISA OECD, 2017). The PISA 2015 assessment focused on science as the major domain. For illustrating the proposed approach, we focused on the constructs "environmental awareness" and "enjoyment of science", measured with 7 and 5 four-point Likert scale items, respectively. Items for either scale were presented on a single screen. For measuring environmental awareness, respondents were asked to gauge how informed they are on different environmental issues, e.g. nuclear waste or water shortage. Enjoyment of science was measured by asking respondents to express their level of agreement with statements such as "I generally have fun when I am learning science topics". We analyzed data from the German sample, comprising $N = 2847$ respondents. All analyses were performed using R version 3.6.3 (R Development Core Team, 2017).

6.1. Implementation of Model-Based Approaches

To investigate differences in conclusions based on different RT measures, we conducted four separate analyses. We considered three measures for aggregated RTs, each aggregating different information of the response process, and one measure for item-level RTs. We used the LogFSM package (Kroehne, 2019) to extract item-level RTs from raw log events. The package implements the finite state machine (FSM) framework for log data presented by Kroehne and Goldhammer (2018). In the FSM framework, the RT for an item is defined as the difference between the time stamp associated with choosing a response option on that item and the time stamp associated with providing the preceding response. Note that in this framework the RT for the first item cannot be reconstructed as it is confounded with the time taken for reading the question stem. The FSM framework does not require items to be answered in a linear order (see Kroehne et al., April 2019, for details).

We considered aggregated RT measures that pose proxies for item-level RTs. As the most coarse proxy for item-level RTs, we considered the total time spent on screen (denoted with TT) divided by the number of items presented on the screen J_s. TT is oftentimes publicly available in computer-administered questionnaires. It, however, poses an aggregate of the time required for both reading and evaluating the question stem and the time required for reading, evaluating, and generating responses to the items presented on the screen. To separate these aspects, Kroehne et al. (April 2019) suggested to subtract the time elapsed until the first response (FRT) from TT. Note that FRT contains the time required for reading the question stem and answering the first item (see Kroehne & Goldhammer, 2018) . Hence, to eliminate reading time from the aggregated RT measure, we also considered TT−FRT divided by $J_s - 1$ (denoted with TTFRT). Note that TT as given in the PISA 2015 data set is already cleansed. Since aggregated RT data provided in public use files are often cleansed, we aimed at investigating whether this preliminary data cleansing impacts conclusions and compared TTFRT against the average of reconstructed item-level RTs as a further measure for the average answering time (denoted with AAT). If the data cleansing procedure performed on aggregated RT data available in the PISA public use file does not impact conclusions, results for TTFRT and AAT should be similar. Note that both TTFRT and AAT are measures containing information on the average amount of time respondents required to generate responses to all but the first item answered.

Bayesian estimation was conducted using Stan version 2.19 (Carpenter et al., 2017) employing the rstan package version 2.19.3 (Guo, Gabry, & Goodrich, 2018). Stan code for both model types is provided in Appendix. For all models, we ran four Markov chain Monte Carlo (MCMC) chains with 4,000 iterations each, with the first half being employed as warm-up. The sampling procedure was assessed on the basis of potential scale reduction factor (PSRF) values, with PSRF values below 1.10 for all parameters being considered as satisfactory (Gelman & Rubin, 1992; Gelman & Shirley, 2011). In our first analyses, the model with item-level RTs did not converge.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

We therefore trimmed RTs, removing item-level RTs exceeding the 99.9th percentile of 90 seconds. Given the median of 2.51 seconds and the middle 50% range of [1.61; 3.79], RTs above 90 seconds are aberrantly large and occurred very rarely. In total, this led to the exclusion of 29 item-level RTs (i.e., 0.08% of the RT data points). We further excluded 13 AAT and 20 TTFRT values exceeding 90 seconds. No TT values were excluded as these did not show such aberrances. Note that none of the respondents exceeded 90 seconds on all available timing measures, such that all analyses were based on the same set of respondents. With this setup, for all models, the chains mixed well and no PSRF values below 1.10 were encountered.

6.2. Implementation of Indicator-based Procedures

We compared and contrasted the proposed approach with the performance of customary indicator-based procedures. For fair comparisons with the presented approach, that can deal with different forms of C/IER, we implemented a multiple-hurdle approach (Curran, 2016). This approach uses multiple indicators that are sensitive to different aspects of C/IER. We focused on three commonly used indicators. Following Meade and Craig (2012), we first filtered respondents with extremely low RTs. For doing so, we employed AAT. Next, response vectors were scanned for C/IER employing the long string index. Finally, to balance off the long string index's insensitivity to C/IER deviating from straight lining, Mahalanobis distance was employed to search the remaining response vectors for C/IER. In the multiple-hurdle approach, thresholds have to be set for each of its components. There are no globally applicable values for these thresholds, as the distributions of the indicators for careless and attentive respondents are scale-specific (Curran, 2016), depending, for instance, on the similarity of the administered items in the case of the long string index or the degree of normality in attentive and careless response distributions in the case of Mahalanobis distance. In order to evaluate the range of possible results and the impact of threshold settings, we implemented two sets of thresholds, choosing either a liberal or a conservative cut-off for all three indicators employed. Under the conservative threshold settings, mean time spent per item below 1 second was set to indicate C/IER. This value corresponds to the halved "educated guess" of 2 seconds for the time required for generating an attentive response by Huang et al. (2012), and is thus aimed at filtering out only the very extreme cases. We required the long string index to correspond to the total number of investigated items (i.e., 13) to be seen as indicating C/IER, which is the most conservative approach possible. Recall that squared Mahalanobis distance can be approximated by a χ^2 distribution with degrees of freedom corresponding to the number of variables (Rousseeuw & Van Zomeren, 1990). Respondents with squared Mahalanobis distances exceeding the 99th quantile of the χ^2 distribution with 13 degrees of freedom were classified as multivariate outliers, indicating C/IER. Under the liberal threshold settings, for the RT threshold, we employed the original "educated guess" by Huang et al. (2012), i.e., set the RT threshold to 2 seconds. For the long string index under liberal threshold settings, we classified respondents as careless when they chose the same response option on at least 5 out of 7 items on the environmental awareness scale and at least 4 out of 5 items on the enjoyment of science scale. Further, squared Mahalanobis distances exceeding the 95th quantile of the χ^2 distribution were seen as indicating C/IER. The long string index and Mahalanobis distance were calculated using the package careless (Yentes & Wilhelm, 2021). In contrasting the multiple-hurdle procedure against the proposed approach, we focused on differences in C/IER classifications.

6.3. Results

6.3.1. Model-Based Approaches Table 1 gives C/IER rates retrieved from all considered approaches. An overview over the remaining parameters from the model-based approaches with different RT measures is displayed in Table 2. By and large, differences between parameter estimates retrieved from models using aggregated and item-level RT information corroborated those

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

TABLE 1.
Rates of careless and insufficient effort responses of threshold-based multiple-hurdle and model-based approaches

Threshold-based		Model-based			
MH_c	MH_l	RT	TT	AAT	TTFRT
9.91%	22.83%	6.29%	8.28%	11.10%	13.02%

Notes: MH_c and MH_l denote the multiple-hurdle approach with conservative and liberal threshold settings, respectively; RT: item-level response times reconstructed from raw log events; TT: total time spent on screen divided by the number of items J_s; AAT: average item-level response time; TTFRT: difference between total time spent on screen and time to the first response divided by $J_s - 1$.

observed in the simulation study. Further, TTFRT and AAT did not yield the same but comparable results, indicating that the data cleansing procedure performed on aggregated RT data available in the PISA public use file does not heavily impact conclusions.

We retrieved attentiveness difficulties of $\iota_1 = -2.74$ and $\iota_2 = -3.47$ for the environmental awareness and enjoyment of science screen, corresponding to screen-level C/IER rates of 7.97% and 3.93%, respectively. In the model with item-level RTs, in total 6.29% of responses were classified as C/IER. Implementing the proposed approach with different RT measures resulted in different conclusions concerning the prevalence of C/IER behavior; there was a twofold difference in C/IER rates between the measure yielding the lowest (item-level RTs) and highest (TTFRT) C/IER rate (see Table 1).

The models yielded rather different common mean and variance estimates for the distribution of C/IE RTs. While the model employing item-level RTs identified the common mean of inattentive RTs to be 0.74, corresponding to 2.10 seconds, TTFRT, for instance, yielded a much higher common mean of 1.11, corresponding to 3.00 seconds. While inattentive RTs as classified by the model employing item-level RTs strongly varied, the distribution of inattentive times in the AAT and TTFRT models showed very low variability. Note that the RT parameters for the model employing TT are not directly comparable with the other models as TT also contains information on reading time.

Consistent across RT measures, results suggest that, marginally, respondents tended to favor middle response categories. This is in line with cognitive theories on edge aversion in decision making processes when items do not need to be (or, as in the present case, are not) processed (Bar-Hillel, 2015).

Besides differences in the variability of speed, mirroring the variability of inattentive RTs, all models yielded comparable conclusions on the relationship between speed and the two traits. Both traits assessed were only weakly related to speed, indicating that respondents with different levels of environmental awareness and enjoyment of science did not considerably differ in the speed with which they generated attentive responses. Environmental awareness and enjoyment of science showed a medium positive correlation. The model with item-level RTs yielded small positive correlations of attentiveness with both traits, indicating that respondents with higher environmental awareness and enjoyment of science levels tended to approach the questionnaire with higher attentiveness. Such conclusions are not possible to draw from the models with aggregated RTs.

Parameters of the measurement model of attentive responses showed very high agreement, with correlations between parameters being above .95 between all models considered. Category probabilities are displayed in Fig. 2.

Time intensity offset parameters in the model with item-level RTs tended to decrease across the seven items of the environmental awareness screen (first two: $\beta^*_{11} = 0.46$ and $\beta^*_{21} = 0.68$; last two: $\beta^*_{61} = 0.13$ and $\beta^*_{71} = 0.00$), indicating that, on average, respondents increased their pace towards the end of this rather long screen. This was not the case for the five items of the enjoyment

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

TABLE 2.
Results for different response time measures

	RT				TT				AAT				TTFRT			
	$\beta_C = 0.74, \sigma_C^2 = 0.78$				$\beta_C = 1.11, \sigma_C^2 = 0.45$				$\beta_C = 1.15, \sigma_C^2 = 0.03$				$\beta_C = 1.11, \sigma_C^2 = 0.05$			
Person parameter variances and correlations																
	ψ	τ	η_1	η_2	τ	η_1	η_2		τ	η_1	η_2		τ	η_1	η_2	
ψ	1.98															
τ	.05	0.11			0.04				0.10				0.15			
η_1	.24	− .14	1.00		− .03	1.00			− .17	1.00			− .07	1.00		
η_2	.17	− .06	.43	1.00	− .04	.43	1.00		− .09	.43	1.00		− .04	.43	1.00	
C/IER category probabilities																
	κ_0	κ_1	κ_2	κ_3	κ_0	κ_1	κ_2	κ_3	κ_0	κ_1	κ_2	κ_3	κ_0	κ_1	κ_2	κ_3
	.14	.34	.41	.11	.15	.46	.31	.07	.10	.40	.41	.09	.09	.39	.44	.08

Notes: RT: item-level response times reconstructed from raw log events; TT: total time spent on screen divided by the number of items J_s; AAT: average item-level response time; TTFRT: difference between total time spent on screen and time to the first response divided by $J_s - 1$; ψ: attentiveness; τ: speed; η_1: environmental awareness; η_2: enjoyment of science; β_C and σ_C^2 give the mean and variance of the inattentive response time distribution.

of science screen (first two: $\beta_{12}^* = 0.04$ and $\beta_{22}^* = 0.24$; last two: $\beta_{42}^* = 0.51$ and $\beta_{52}^* = 0.29$). For the model with TT, screen-specific time intensity offset parameters were 0.32 and 0.24. AAT and TTFRT yielded very low screen-specific time intensity offset parameters (0.05 and 0.00 for AAT and 0.11 and 0.02 for TTFRT), leading to the conclusion that, on average, log aggregated RTs associated with attentive and C/IER behavior in those models did not considerably differ.

With $\gamma = 0.04$ in the model with item-level RTs, we found evidence for the distance–difficulty hypothesis in the selected two scales. That is, when the absolute difference between the respondent's trait level and the item's middle step difficulty increases by one standard deviation, attentive RTs are expected to decrease by the factor $\exp(-0.04) = 0.96$. Comparable conclusions can be drawn on the basis of the models with aggregated RTs, with γ ranging from 0.03 to 0.05.

6.3.2. Comparison with indicator-based procedures In total, the conservative and liberal multiple-hurdle approaches classified 9.91% and 22.83% respondents as careless, respectively (see Table 1). That is, the liberal threshold settings yielded the highest C/IER rate out of all approaches considered and by far exceeded even those obtained from the model-based procedure drawing on TTFRT. The C/IER rate under the conservative threshold settings were similar to those of the model-based procedure drawing on TT and AAT. Under the conservative threshold settings, the C/IER rate goes back to 64 respondents not passing the RT hurdle, 147 respondents failing to pass the long string hurdle, and 135 respondents not passing the Mahalanobis distance hurdle. Under the liberal threshold settings, 401, 46, and 244 respondents did not pass the RT, long string, and Mahalanobis distance hurdle, respectively.

Figure 3 investigates agreement in the classification of respondents under the conservative and liberal threshold settings. The two threshold settings agreed in classifying respondents as attentive and careless in 2153 and 238 cases, respectively. The liberal threshold settings marked 412 respondents as careless that were classified as attentive under the conservative threshold settings, while the opposite was true for only 44 respondents. That is, employing more liberal thresholds lead to adding new respondents to the group of careless respondents rather than identifying different respondents as careless. To investigate agreement between the multiple-hurdle and model-based approaches, Fig. 3 displays median attentiveness parameters from the model-based approaches

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

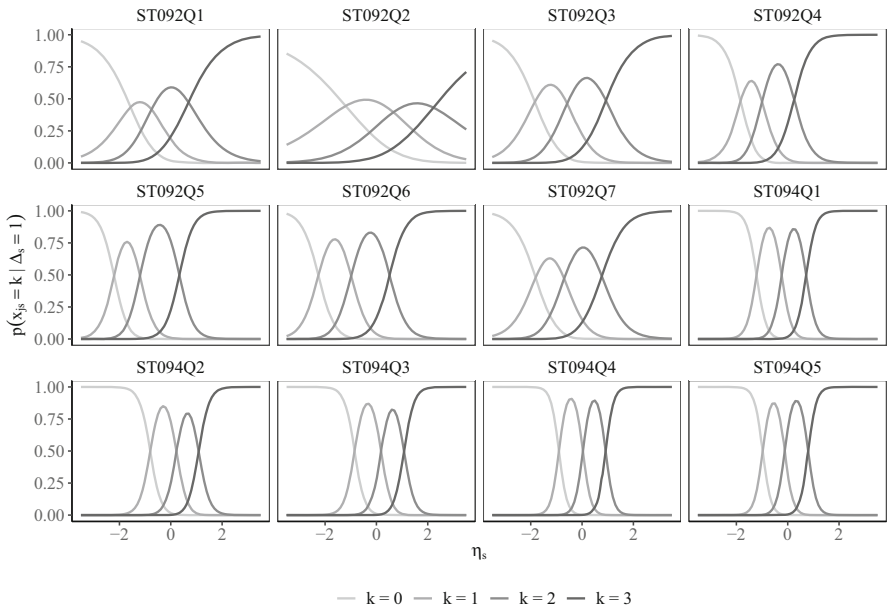


FIGURE 2.
Category probabilities for attentive responses in the model with item-level response times. ST092 and ST094 denote items measuring environmental awareness and enjoyment of science

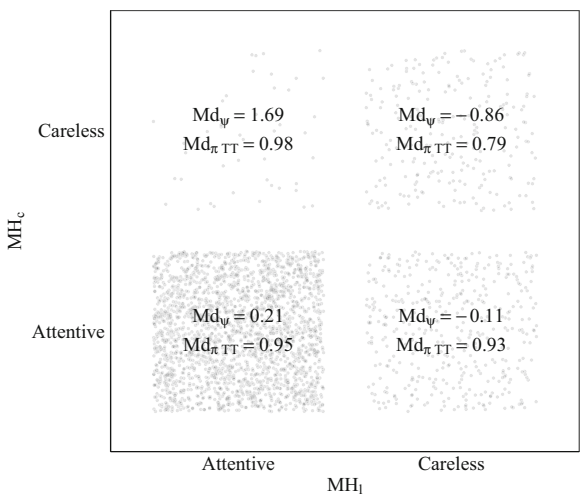


FIGURE 3.
Agreement between the different approaches. Each dot represents a respondent. MH_c and MH_l denote the multiple-hurdle approach with conservative and liberal threshold settings, respectively; Md_{ψ}: median attentiveness parameters from the model-based approach using item-level RTs; $Md_{\pi TT}$: median attentiveness probabilities from the model-based approach using total time spent on screen

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

using item-level RTs and TT, yielding the lowest and highest C/IER rates out of all model-based approaches, for each of these four groups of respondents. Despite differences in overall C/IER rates, we observed agreement between the model-based and multiple-hurdle approaches in that respondents identified under both threshold settings as careless had markedly lower median attentiveness parameters than those identified as attentive under both threshold settings. Median attentiveness parameters in the group of respondents classified as careless under the liberal but attentive under the conservative threshold settings were between those obtained for the two groups where the different threshold settings agreed in their classifications. Interestingly, respondents classified as careless under the conservative but attentive under the liberal threshold settings yielded the highest median attentiveness parameters. Due to the small group size, however, this result needs to be interpreted with caution.

7. Discussion

We presented a model-based approach that leverages response time (RT) information for identifying careless and insufficient effort responding (C/IER). This was achieved by integrating theoretical considerations on C/IER in non-cognitive assessments with recent model developments for identifying non-effortful behavior in cognitive assessment data (Ulitzsch et al., 2020). In doing so, the presented approach overcomes major limitations of previous methods for detecting and dealing with C/IER.

First, as a purely model-based approach, the presented approach does not require setting cut-off values for classifying C/IER. Rather, C/IER is identified employing mixture modeling techniques, assuming different data-generating processes for responses and RTs associated with C/IER and attentive behavior. Second, the approach can detect and deal with multiple response patterns arising from C/IER. This is done in a single step and does not require making assumptions on specific C/IER patterns that may be present in the data at hand. Third, by employing a latent response approach with attentiveness probabilities modeled as a function of person and item parameters, the approach allows for C/IER behavior to vary on the screen-by-respondent level as well as to assess screen and respondent characteristics associated with such behavior. Allowing for C/IER behavior to vary across the computer-administered questionnaire also allows keeping information from attentive responses of respondents who showed C/IER on some but not all screens. Further, the employed latent response approach supports considering differences in attentiveness when estimating the traits to be measured.

The approach comprises model classes for both item- and screen-level timing data. Item-level RTs potentially allow for a more precise depiction of response processes, however, are oftentimes not readily available. Conversely, screen-level timing data can easily be recorded with common tools for computer-administered questionnaires, making the approach readily applicable for typical data sets and do not require the collection of raw log events. The model for screen-level timing data poses a simplified version that—similar to previous approaches for C/IER—allows to identify C/IER at the respondent level. As such, it gives up some of the advantages of employing a latent response approach. Investigating parameter recovery, we found the model drawing on item-level RTs to yield unbiased estimates even under conditions with sparse information on C/IER behavior. Simplifying the model and using aggregated RTs led to overestimating the extent of C/IER behavior. Nevertheless, estimates of the correlations between traits were still unbiased under the investigated conditions. Correlations between traits commonly pose parameters of interest for applied researchers. Hence, in the case that no item-level RT information is available, the model for aggregated RTs can be used for screening for C/IER behavior and retrieving valid conclusions concerning the traits to be measured. When doing so, researchers should keep in mind that proportions of C/IER may be biased. We further note that, based on results from the empirical

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

example, we simulated attentiveness to be only weakly related to the traits to be measured. When this is not the case and C/IER prevalences are higher, falsely assuming attentiveness to be unrelated to the traits to be measured, as done in both the model for aggregated RTs as well as customary indicator-based procedures, may yield biased conclusions (see Ulitzsch et al., 2020).

The approach was illustrated on data from the German PISA 2015 background questionnaire, employing different RT measures. We could show that the presented approach yields meaningful results for all RT measures and that the models employing different RT measures did not impact conclusions on trait correlations. Differences in conclusions concerning the prevalence of C/IER behavior corroborated those observed in the simulation study. Further, different RT aggregates that differed in whether they contained reading time yielded different conclusions on C/IER prevalence. We further illustrated the advantages of the proposed approach over previous indicator-based procedures. To this end, we showed how conclusions drawn on the basis of indicator-based procedures are heavily dependent on threshold settings, with vast differences being observable even for small differences in the exemplarily employed thresholds. Note that the difference between the different threshold settings was much larger than those between different RT aggregates, suggesting that decisions on thresholds are much more critical than decisions on the RT aggregate employed when no item-level RTs are available.

7.1. Limitations and Future Research

The proposed approach's component models for attentive and C/IE responses and timing data are formulated based on theoretical considerations on response behavior. Further, agreement with previously validated indicator-based procedures in that respondents with lower attentiveness parameters were at greater risk of being identified as careless in multiple-hurdle approaches provided first supporting validity evidence for the proposed approach. Nevertheless, further research on the approach's validity for identifying C/IER is needed. Here, both studies conducted in the context of non-cognitive assessments (e.g., Meade & Craig, 2012; Niessen et al., 2016) as well as in the context of cognitive assessments may serve as blueprints (see Ulitzsch, Penk, von Davier, & Pohl, 2021, for a validation of the SA+E model for rapid guessing behavior)

Further, investigating to which extent reading time carries valid information on C/IER behavior is a pertinent topic for future research. This is a question that can only be addressed by a combination of theory and empirical research. In the case that C/IE respondents can be assumed to evaluate the question stem in a manner comparable to attentive respondents, reading time would pose a nuisance that confounds speed with which respondents read the question stem with differences in attentiveness and therefore should be left out when identifying C/IER on the basis of RT data. Conversely, in the case that respondents are assumed to skip reading the question stem, reading times should be considerably shorter and would, as such, pose a valid source of additional information on inattentiveness. Combinations of the two mechanisms may also be present in empirical data. Results from studies investigating these issues could then be integrated with the presented approach for an even finer-grained depiction of response behavior.

It should also be noted that the FSM used to reconstruct item-level RTs rests on assumptions on how respondents evaluate and respond to items that may be violated in practice. Examples for violations may be respondents that do not start by reading the question stem (Kroehne et al., April 2019) or respondents that first cognitively evaluate all items and then "bundle" the technical processes of choosing their answers, resulting in long times until their first response and only short times elapsing between subsequent responses. If that is the case, aggregated RT information that does not rely on these assumptions might pose a more stable and reliable source of information on respondents' answering behavior. This issue could be addressed by simulating data that differ in whether or not assumptions of the FSM used to reconstruct item-level RTs from raw log events

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

hold. The performance of different RT measures could then be compared to identify conditions under which each measure gives the most accurate estimate of the prevalence of C/IER behavior.

Regardless of the specific type of RT information employed, the presented approach heavily relies on this information for identifying C/IER. Hence, violations of assumptions on data-generating processes underlying RTs associated with attentive and C/IE responses may potentially result in misclassifications (see Molenaar, Bolsinova, & Vermunt, 2018). For instance, when the distance–difficulty effect as incorporated in the component model for attentive RTs does not adequately capture the relationship between attentive RTs and trait levels, assumptions on data-generating processes underlying attentive RTs are violated. To address this, a better understanding of the cognitive processes underlying attentive RTs in questionnaire data is urgently needed. A further possible violation of assumptions are changes in speed due to, for instance, warming up effects (Weitensfelder, 2017). By allowing for item-specific time intensity offsets, the model for item-level RTs can capture changes in speed that are shared by all respondents. The model cannot deal, however, with changes in speed that vary across respondents. To accommodate this, the presented approach may be extended by a growth curve model for speed (Fox & Marianti, 2016). To make mixture modeling approaches more robust to violations of distributional assumptions, Molenaar et al. (2018) suggested to employ a semi-parametric approach by categorizing RTs that could also be integrated with the presented approach.

As it is the case with previous behavioral measures of C/IER in non-cognitive assessment data (Huang et al., 2012; Meade & Craig, 2012; Niessen et al., 2016) and (rapid) guessing in cognitive assessment data (Nagy & Ulitzsch, 2021; Ulitzsch et al., 2020; Wang & Xu, 2015; Wise, 2017), the presented approach assumes inattentiveness to manifest itself in responses that do not reflect the construct to be measured. It does not consider C/IER that reflects the construct to be measured to some degree, which may occur when respondents skip lengthy instructions (Maniaci & Rogge, 2014) or read the item but do not put effort in retrieving the relevant information (see Ulitzsch et al., 2021, for a discussion of non-effortful responding in cognitive assessment). Further, the empirical application indicated that the model can not deal well with outrageously long RTs. These may stem from both attentive and C/IE response processes. Long attentive RTs may stem from respondents having problems understanding the questions or being indecisive between different response options. In online-administered questionnaires, long inattentive RTs may stem, for instance, from switching to another browser tab and subsequently providing a C/IE response. Using the proposed approach, we can already model very short RTs, assuming that these are likely to stem from C/IER behavior. Better understanding other types of C/IER behavior as well as the behavior underlying the occurrence of very long RTs and subsequently integrating these behaviors with the proposed approach remains an important task for future research.

A strength of the approach is that it can detect and deal with various types of C/IER at once. The price for this is that it does not allow for inferences on the specific types of C/IER. We can identify respondents with C/IER behavior, but do not know which type of C/IER behavior is shown. Note, however, that researchers are usually only interested in unbiased estimation of trait levels (i.e., accounting fo C/IER behavior), but not necessarily in the specifics of C/IER behavior. In case these are of interest, one may investigate response patterns of respondents with low attentiveness estimates and scan for specific patterns (e.g., straight or diagonal lining). If the goal is to model specific types of response styles, other approaches might be more appropriate.

It should also be noted that the scalability of the presented approach to the analysis of data with large samples and a high number of investigated constructs is limited, as, due to model complexity, infeasible running times may be encountered. Although we expect this issue to resolve with algorithmic and computational advances, for now, under such data constellations, researchers may find heuristic indicator-based approaches to be more practical for gauging the extent of C/IER in the data at hand.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

The proposed approach allows for identifying and modeling C/IER in data retrieved from computer-administered questionnaires and, thereby, increasing the validity of inference drawn from such data. Implementing the proposed approach for real-time estimation of attentiveness poses a highly promising extension. Doing so would allow to monitor C/IER during the assessment procedure, issue warnings once pre-defined thresholds of acceptable aberrances are exceeded, and nudge respondents to provide more valid responses, thus increasing data quality. In experimental settings, Huang et al. (2012) as well as Wise, Bhola, and Yang (2006) already demonstrated the positive effects of warnings on attentive responding in both cognitive and non-cognitive assessments.

The approach is aimed at improving the validity of conclusions drawn on C/IER prevalences in the data at hand as well as relationships among constructs of interest by identifying and modeling C/IE responses. When doing so, the approach does not only provide parameter values for the employed IRT model that are based on attentive responses only, but also provides information on the attentiveness of respondents. Future research may investigate what further information the attentiveness variable provides on respondents, e.g., whether it provides a behavioral measure of respondent's personality (see Pohl, Ulitzsch, & von Davier, 2021, for a neighboring discussion on behavioral aspects impacting test results). In support of this, Bowling et al. (2016) could show that individual differences in C/IER behavior as reflected in response–pattern-based indicators are consistent across time and study situations, and that C/IER is related to acquaintance-reported personality as well as to college grade point average and class absences. The proposed approach provides a sophisticated tool for furthering investigations of the additional information contained in response behavior and its relevance for real-life outcomes.

Note that, although not as widely available as RTs from cognitive assessments, in the context of non-cognitive assessments, item-level RT data become increasingly available (see Henninger & Plieninger, 2020, for recent studies; and Tunguz, November 2018, for a publicly available large-scale personality inventory data set with item-level RTs) or can be reconstructed using FSMs (Kroehne, 2019; Kroehne & Goldhammer, 2018). The present study showcased the utility of item-level RTs to gain a finer-grained understanding of respondents' behavior in general and identifying C/IER behavior in particular and can as such be understood as a call for recording this rich source of information in non-cognitive assessments.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix

Stan Code

See Figs. 4, 5, 6 and 7.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

```
functions {
  real pcm(int y, real eta, vector b) {
    vector[rows(b) + 1] unsummed;
    vector[rows(b) + 1] probs;
    unsummed = append_row(rep_vector(0.0, 1), eta - b);
    probs = softmax(cumulative_sum(unsummed));
    return categorical_lpmf(y| probs);
  }
  vector subset(vector x, int[] is, int target) {
   int count;
   count = 0;
   for (n in 1:size(is)){
       if (is[n] == target){
        count = count + 1;
       }
   }
   {
   vector[count] result;
   int pos;
   pos = 1;
   for (n in 1:size(is)) {
     if (is[n] == target) {
       result[pos] = x[n];
       pos = pos + 1;
     }
   }
   return result;
  }
  }
}
```

FIGURE 4.
Function for **a** the partial credit model as well as **b** for subsetting vectors, used for identifying middle step difficulties

```
data{
  int<lower = 1> N; // number of respondents
  int<lower = 1> K; // number of items (in total)
  int<lower = 2> C; // number of categories
  int<lower = 1> Ntot; // number of data points (responses, RTs)
  int<lower = 1> jj[Ntot]; // item id
  int<lower = 1> ii[Ntot]; // person id
  int<lower = 1> ss[Ntot]; // screen id
  int<lower=1,upper=C> y[Ntot]; // responses
  real logtAgg[Ntot]; // log RTs
  int<lower=1,upper=C> d[Ntot]; // missingness indicators RTs (missingness due to FSM)
  int<lower = 1>S; // number of screens
  int<lower = 0>midAgg[K*(C-1)]; // mid point indicator for each item
  // (s for midpoint position and 0 otherwise)
}
transformed data {
  int pos[K];  // first position in beta vector for item
  int m = C-1;  // number of  step difficulties per item
  pos[1] = 1;
  for(i in 2:(K)) pos[i] = m + pos[i-1];
}
parameters{
  corr_matrix[S+2] correlP; // correlation person variables
  vector<lower=0>[2] sigmaS; // standard deviation attentiveness and speed
  vector[m*K] b; // step difficulties
  vector<lower=0>[K] v; // discriminations
  vector<lower=0>[K] diffbeta; // time intensity offset parameters
  row_vector[S+2] PersPar[N]; // person parameter, 1: attentivess, 2: speed, 3:(S:2): traits
  real<lower = 0> sigmaA; // residual variance attentive RT
  real<lower = 0> sigmaC; // residual variance C/IE RT
  real betaC; // mean log C/IE RT
  simplex[C] kappa; //  C/IER category probabilities
  real gamma; // regression parameter distance-difficulty parameter
  vector[S] iota; // attentiveness difficulties
}
transformed parameters{
  vector[S+1] sigmaP; // person parameter standard deviations
  cov_matrix[S+1] SigmaP; // person parameter covariance matrix
  vector[K] beta; // time intensity parameters
  vector[K] o; // middpoint step difficulties
  real<lower=0,upper=1> pDelta[Ntot]; // attentivess probability
  for(n in 1:Ntot) pDelta[n] = 1/(1+exp(-PersPar[ii[n],1] + iota[ss[n]]));
  sigmaP[1:2] =sigmaS;
  sigmaP[3:(S+2)] =rep_vector(1,S);
  SigmaP=quad_form_diag(correlP, sigmaP);
  for(j in 1:K) beta[j]=muC+diffbeta[j];
  for(j in 1:K) o[j]=mean(subset(b,mid,j));
}
model{
  sigmaS ~  cauchy(0,5);
  correlP ~ lkj_corr(1);
  PersPar~ multi_normal(rep_vector(0,(S+1)),SigmaP);
  iota ~ normal(0, 10);
  b ~ normal(0, 10);
  v ~  cauchy(0,5);
  diffbeta  ~ normal(0, 10);
  gamma~ normal(0, 10);
  kappa~ dirichlet(rep_vector(1, C));
  betaC ~ normal(0, 10);
  sigmaC ~  cauchy(0, 5);
  sigmaA ~  cauchy(0, 5);
  for(n in 1:Ntot) {
  target += log_mix(pDelta[n],
    // attentive
    pcm(y[n],v[jj[n]]*PersPar[ii[n],ss[n]+2], segment(b, pos[jj[n]], m)) +
    (1-d[n])*normal_lpdf(logt[n]|beta[jj[n]]-
    gamma*fabs(v[jj[n]]*PersPar[ii[n],ss[n]+2]-o[jj[n]]) - PersPar[ii[n],2], sigmaA),
    // inattentive
    categorical_lpmf(y[n]|kappa) +
    (1-d[n])*normal_lpdf(logt[n]|betaC, sigmaC));
  }
}
```

FIGURE 5.
Stan code for the model with item-level RTs

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

```
data{
  int<lower = 1> N; // number of respondents
  int<lower = 1> K; // number of items (in total)
  int<lower = 2> C; // number of categories
  int<lower = 1> Ntot; // number of data points (responses)
  int<lower = 1> NtotS; // number of data points (RTs, aggregated across screens)
  int<lower = 1> jj[Ntot]; // item id
  int<lower = 1> ii[Ntot]; // person id
  int<lower = 1> ss[Ntot]; // screen id
  int<lower=1,upper=C> y[Ntot]; // responses
  real logtAgg[NtotS]; // log screen-level timing data
  int<lower = 1> iiS[NtotS]; // person id for timing data
  int<lower = 1> ssS[NtotS]; // screen id for timing data
  int<lower = 1>S; // number of screens
  int<lower = 0>midAgg[K*(C-1)]; // mid point indicator for each item
  // (s for midpoint position and 0 otherwise)
  int<lower = 1>vAgg[K]; // screen-indicator for discrimination parameter
  real<lower = 1>kS[S]; // number of items per screen
}
transformed data {
  int pos[K];   // first position in beta vector for item
  int m = C-1;  // number of  step difficulties per item
  pos[1] = 1;
  for(i in 2:(K)) pos[i] = m + pos[i-1];
}
parameters{
  corr_matrix[S+1] correlP; // correlation person variables
  real<lower=0> sigmaS; // standard deviation speed
  vector[m*K] b; // step difficulties
  vector<lower=0>[K] v; // discriminations
  vector<lower=0>[S] diffbeta; // screen-specific time intensity offset parameters
  row_vector[S+1] PersPar[N]; // person parameter, 1: speed, 2:(S+1): traits
  real<lower = 0> sigmaA; // residual variance attentive RT
  real<lower = 0> sigmaC; // residual variance C/IE RT
  real betaC; // mean log C/IE RT
  simplex[C] kappa; //  C/IER category probabilities
  simplex[2] piAtt[N]; // person-specific attentiveness probability
  real gamma; // regression parameter distance-difficulty parameter
  simplex[2] piPop; // population-level class probabilities
  real<lower=0> lambda; // for hierarchical dirichlet prior
}
```

FIGURE 6.
Stan code for the model with aggregated RTs (part I)

```
transformed parameters{
  vector[S+1] sigmaP; // person parameter standard deviations
  cov_matrix[S+1] SigmaP; // person parameter covariance matrix
  vector[S] beta; // screen-specific time intensity parameters
  vector[S] meano; // mean middle step difficulty
  vector[S] meanv; // mean discrimination
  sigmaP[1] = sigmaS;
  sigmaP[2:(S+1)] = rep_vector(1,S);
  SigmaP=quad_form_diag(correlP, sigmaP);
  for(s in 1:S) {
    beta[s] = muC+diffbeta[s];
    meano[s] = mean(subset(b,midAgg,s));
    meanv[s] = prod(subset(v,vAgg,s))^(1/kS[s]);
  }
}
model{
  vector[2] contributionsY[Ntot]; // responses
  vector[2] contributionsR[NtotS]; // response times
  sigmaS ~  cauchy(0,5);
  correlP ~ lkj_corr(1);
  PersPar ~ multi_normal(rep_vector(0,(S+1)),SigmaP);
  b ~ normal(0, 10);
  v ~  cauchy(0,5);
  diffbeta  ~ normal(0, 10);
  gamma ~ normal(0, 10);
  kappa ~ dirichlet(rep_vector(1, C));
  muC ~ normal(0, 10);
  sigmaC ~  cauchy(0, 5);
  sigmaA ~  cauchy(0, 5);
  for(i in 1:N) piAtt[i] ~ dirichlet(piPop*lambda);
  piPop  ~ dirichlet(rep_vector(1,G));
  lambda  ~ cauchy(0,5);
  for(n in 1:Ntot) {
    // class 1: attentive
     contributionsY[n,1] = log(piAtt[ii[n],1]) + pcm(y[n],v[jj[n]]*PersPar[ii[n],ss[n]+1],
       segment(b, pos[jj[n]], m));
    // class 2: inattentive
    contributionsY[n,2] = log(piAtt[ii[n],2]) + categorical_lpmf(y[n]|kappa);
    target += log_sum_exp(contributionsY[n]);
  }
  for(n in 1:NtotS){
    // class 1: attentive
    contributionsR[n,1] = log(piAtt[iiS[n],1]) + normal_lpdf(logtAgg[n]|beta[ssS[n]]-
      gamma*fabs(meanv[ssS[n]]*PersPar[iiS[n],ssS[n]+1]-meano[ssS[n]])-
      PersPar[iiS[n],1], sigmaA);
    // class 2: inattentive
    contributionsR[n,2] = log(piAtt[iiS[n],2]) + normal_lpdf(logtAgg[n]|betaC, sigmaC);
    target += log_sum_exp(contributionsR[n]);
  }
}
```

FIGURE 7.
Stan code for the model with aggregated RTs (part II)

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology, 72*(3), 466–485. https://doi.org/10.1111/bmsp.12169

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*(1), 139–151. https://doi.org/10.1207/s15327752jpa6801_11

Bar-Hillel, M. (2015). Position effects in choice from simultaneous displays: A conundrum solved. *Perspectives on Psychological Science, 10*(4), 419–433. https://doi.org/10.1177/1745691615588092

Baumgartner, H., & Steenkamp, J.-B.E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4*(3), 340. https://doi.org/10.1037/1040-3590.4.3.340

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology, 70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528–541. https://doi.org/10.1037/met0000016

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*(2), 218.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, https://doi.org/10.18637/jss.v076.i01

Curran, P. G., & Hauser, K. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, https://doi.org/10.1016/j.jrp.2019.103849

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Curran, P. G., & Denison, A. J. (2019). Creating carelessness: A comparative analysis of common techniques for the simulation of careless responder data. https://doi.org/10.31234/osf.io/ge6fa

Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement*, 58 (2), 281–303. https://doi.org/10.1111/jedm.12290

DeSimone, J. A., DeSimone, A. J., Harms, P., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology, 67*(2), 309–338. https://doi.org/10.1111/apps.12117

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Ehlers, C., Greene-Shortridge, T., Weekley, J., & Zajack, M. (2009). The exploration of statistical methods in detecting random responding. *Paper presented at the Annual Meeting of the Society for Industrial/Organizational Psychology*. Atlanta, GA.

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224–247. https://doi.org/10.1177/0146621607302479

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*(6), 525–543. https://doi.org/10.1177/0146621606295197

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 163–174). Boca Raton, FL: Chapman Hall.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC (OECD Education Working Papers No. 133)*. OECD Publishing. https://doi.org/10.1787/19939019

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766

Guo, J., Gabry, J., & Goodrich, B. (2018). Rstan: R interface to Stan. R package version 2.18.2. Retrieved from https://CRAN.R-project.org/package=rstan

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Henninger, M., & Plieninger, H. (2020). Different styles, different times: How response times can inform our knowledge about the response process in rating scale measurement. *Assessment*, https://doi.org/0.1177/1073191119900003

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114. https://doi.org/10.1007/s10869-

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

011-9231-8
Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828–845. https://doi.org/10.1037/a0038510
Jackson, D. (1976). The appraisal of personal reliability. *Paper presented at the Meetings of the Society of Multivariate Experimental Psychology*. University Park, PA.
Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009
Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*(3), 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x
Khorramdel, L., Jeon, M., & Leigh Wang, L. (2019). Advances in modelling response styles and related phenomena. *British Journal of Mathematical and Statistical Psychology, 72*(3), 393–400. https://doi.org/10.1111/bmsp.12190
Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2018). Straightlining: Overview of measurement, comparison of indicators, and effects in mail-web mixed-mode surveys. *Social Science Computer Review, 37*(2), 214–233. https://doi.org/10.1177/0894439317752406
Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement, 54*(4), 397–419. https://doi.org/10.1111/jedm.12154
Kroehne, U. (2019). LogFSM: Analysis of log data using finite-state machines. Retrieved from https://github.com/kroehne/LogFSM
Kroehne, U., Buchholz, J., & Goldhammer, F. (2019). Detecting carelessly invalid responses in item sets using item-level response times. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*. Toronto, Canada.
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. https://doi.org/10.1007/s41237-018-0063-y
Kuncel, R. B., & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement, 34*(4), 743–755. https://doi.org/10.1177/001316447403400401
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*, https://doi.org/10.1186/s40536-014-0008-1
Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008
Mahalanobis, P. C. (1936). *On the generalized distance in statistics*. National Institute of Science of India.
Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008
Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*(4), 523–547. https://doi.org/10.1007/BF02294327
McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*(3), 450–470. https://doi.org/10.1037/a0019216
Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085
Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4), 311–314. https://doi.org/10.1177/014662169401800402
Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 205–228. https://doi.org/10.1111/bmsp.12117
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684
Muraki, E. (1997). A generalized partial credit model. In *Handbook of modern item response theory* (pp. 153–164). Springer.
Nagy, G., & Ulitzsch, E. (2021). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211045351.
Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use. *Journal of Research in Personality, 63,* 1–11. https://doi.org/10.1016/j.jrp.2016.04.010.
OECD. (2017). PISA 2015 technical report. OECD Publishing. Paris, France. Retrieved from https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf
Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423–452. https://doi.org/10.1177/0013164413504926
Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340. https://doi.org/10.1126/science.abd3300
R Development Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org
Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling, 55*(4), 361–382.

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press

Rose, N. (2013). Item nonresponses in educational and psychological measurement (Doctoral dissertation, Friedrich-Schiller-Universität Jena). Retrieved from https://d-nb.info/1036873145/34
Rose, N., von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT) (ETS Research Report No. RR-10-11). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x
Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association, 85*(411), 633–639. https://doi.org/10.1080/01621459.1990.10474920
Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2012). Learning with hierarchical-deep models. IEEE transactions on pattern analysis and machine intelligence, 35 (8), 1958-1971. https://doi.org/0.1109/TPAMI.2012.269
Samejima, F. (2016). Graded response models. In Handbook of item response theory (pp. 123-136). Chapman and Hall/CRC.
Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*(4), 367–373. https://doi.org/10.1177/014662168500900405
Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x
Schroeders, U., Schmidt, C., & Gnambs, T. (2020). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211004708
Takagishi, M., van de Velden, M., & Yadohisa, H. (2019). Clustering preference data in the presence of response-style bias. *British Journal of Mathematical and Statistical Psychology, 72*(3), 401–425. https://doi.org/10.1111/bmsp.12170
Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77*, 184–197. https://doi.org/10.1016/j.chb.2017.08.038
Tunguz, B. (November 2018). Big Five personality test. Retrieved from https://www.kaggle.com/tunguz/big-five-personality-test
Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (2021). Modell meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment*. https://doi.org/10.1080/10627197.2020.1858786
Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level nonresponse. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12188
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z
Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477. https://doi.org/10.1111/bmsp.12054
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. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61. https://doi.org/10.1111/emip.12165
Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21–30. https://doi.org/10.1111/j.1745-3992.2006.00054.x
Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x
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. https://doi.org/10.1080/08957347.2017.1353992
Wise, S. L., Kingsbury, G., Thomason, J., & Kong, X. (April 2004). An investigation of motivation filtering in a statewide achievement testing program. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Diego, CA.
Wise, S. L., & Ma, L. (April 2012). Setting response time thresholds for a CAT item pool: The normative threshold method. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Vancouver, Canada.
Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185–205. https://doi.org/10.1080/08957340902754650
Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 189–94. https://doi.org/10.1007/s10862-005-9004-7
Yentes, R. D., & Wilhelm, F. (2021). Careless: Procedures for computing indices of careless responding. *R package version, 1*(2), 1.

Manuscript Received: 25 AUG 2020
Final Version Received: 11 OCT 2021
Published Online Date: 2 DEC 2021

https://doi.org/10.1007/s11336-021-09817-7 Published online by Cambridge University Press