Ulitzsch *et al.*
Large-scale Assessments in Education (2024) 12:18
https://doi.org/10.1186/s40536-024-00205-y

Large-scale Assessments in Education

RESEARCH Open Access

Using a novel multiple-source indicator to investigate the effect of scale format on careless and insufficient effort responding in a large-scale survey experiment

Esther Ulitzsch[1,2,3*], Janine Buchholz[4], Hyo Jeong Shin[5], Jonas Bertling[6] and Oliver Lüdtke[3,7]

*Correspondence:
esther.ulitzsch@cemo.uio.no

[1] Centre for Educational Measurement, University of Oslo, Gaustadalleen 21, 0349 Oslo, Norway
[2] Centre for Research on Equality in Education, University of Oslo, Oslo, Norway
[3] IPN-Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[4] Institute for Educational Quality Improvement (IQB), Berlin, Germany
[5] Sogang University, Seoul, South Korea
[6] Educational Testing Service (ETS), Princeton, USA
[7] Centre for International Student Assessment (ZIB), Munich, Germany

Abstract

Common indicator-based approaches to identifying careless and insufficient effort responding (C/IER) in survey data scan response vectors or timing data for aberrances, such as patterns signaling straight lining, multivariate outliers, or signals that respondents rushed through the administered items. Each of these approaches is susceptible to unique types of misidentifications. We developed a C/IER indicator that requires agreement on C/IER identification from multiple behavioral sources, thereby alleviating the effect of each source's standalone C/IER misidentifications and increasing the robustness of C/IER identification. To this end, we combined a response-pattern-based multiple-hurdle approach with a recently developed screen-time-based mixture decomposition approach. In an application of the proposed multiple-source indicator to PISA 2022 field trial data we (a) showcase how the indicator hedges against (presumed) C/IER overidentification of its constituting components, (b) replicate associations with commonly reported external correlates of C/IER, namely agreement with self-reported effort and C/IER position effects, and (c) employ the indicator to study the effects of changes of scale characteristics on C/IER occurrence. To this end, we leverage a large-scale survey experiment implemented in the PISA 2022 field trial and investigate the effects of using frequency instead of agreement scales as well as approximate instead of abstract frequency scale labels. We conclude that neither scale format manipulation has the potential to curb C/IER occurrence.

Keywords: careless responding, screen times, Multiple-hurdle, Scale format

Introduction

Careless and insufficient effort responding (C/IER)—occurring when respondents do not invest effort into carefully evaluating and responding to survey content—poses a well-known threat to the quality of survey data. C/IER spans a wide range of behavioral patterns—ranging from random responding (as in Fig. 1a) through marking distinct patterns such as straight (as in Fig. 1b) or diagonal lines (as in Fig. 1c) or alternating extreme pole responses (see Fig. 1d) to providing no response at all—all of which result in response patterns that are unreflective of the constructs to be

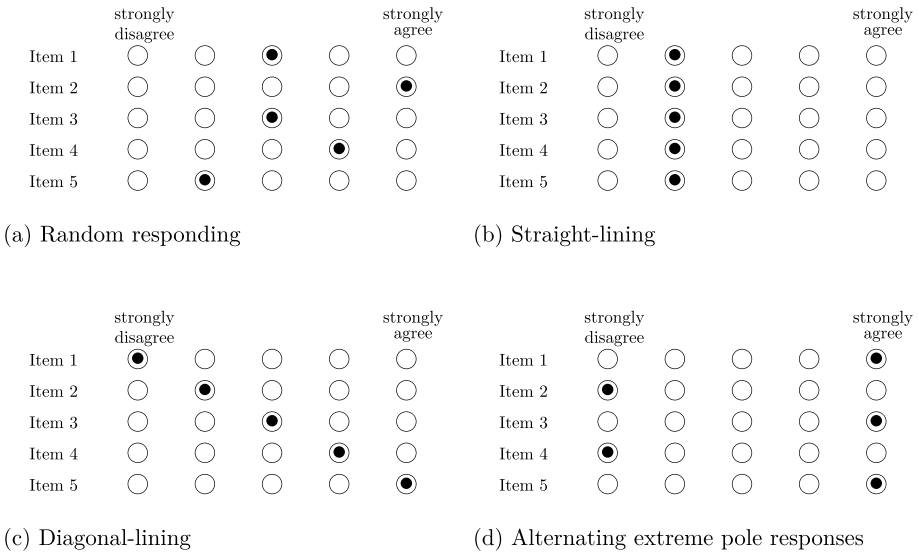


Fig. 1 Schematic illustration of different careless and insufficient effort response patterns. The illustration is adapted from Ulitzsch et al. (2021b)

measured and potentially introduce construct-irrelevant variability. As such, when left unconsidered, C/IER may heavily distort conclusions drawn from survey data (Huang et al., 2012; Woods, 2006; Schmitt & Stuits, 1985; DeSimone et al., 2018; Kam & Meyer, 2015).

So far, there is no consensus on how to best handle the occurrence of C/IER in survey data. Two general types of approaches can be distinguished. First, researchers may opt for survey designs that curb the occurrence of C/IER. For instance, in large-scale surveys, split questionnaire designs, also known as multiform designs, that administer only a fraction of the available items to each respondent (Graham et al., 1994) are rather common and aim to reduce respondent burden as one of the major drivers of C/IER (Krosnick, 1991). Second, researchers can employ post-hoc adjustments that scan available data for aberrances and subsequently either completely eliminate or downweigh response patterns presumably stemming from C/IER. Both types of approaches rely on valid and reliable C/IER measurement. For conducting investigations and accumulating knowledge on survey, scale, and item characteristics that foster or curb C/IER, valid C/IER measures pose a key prerequisite. Likewise, when adjusting for C/IER when drawing inferences on population-level parameters, the employed C/IER detection method should be as accurate as possible, neither leaving data contaminated with C/IER, nor discarding valid information stemming from attentive response behavior. This is because both types of misclassifications can potentially induce bias. Bias induced by leaving data contaminated is extensively documented (e.g., DeSimone et al., 2018; Woods, 2006; Schmitt & Stuits, 1985). The effects of excluding valid, attentive data are less well studied. However, one can easily imagine scenarios where the exclusion of presumed C/IE responses leads to the systematic exclusion of specific sub-groups from the data (e.g., those with very high trait levels when respondents selecting solely the upper response

option on a homogeneous scale are flagged). This, in turn, can result in bias of parameters of interest, e.g., correlation coefficients or group means.[1]

In the present study, we develop a C/IER indicator that requires agreement on C/IER identification from multiple behavioral sources, thereby balancing classification inaccuracies from each of its constituting components. In so doing, we aim to provide a tool that facilitates the improvement of both prevention and adjustment approaches to handling C/IER. We then apply the developed indicator to data from the Programme for International Student Assessment (PISA) 2022 field trial, administering a lengthy questionnaire in a large, cross-cultural sample. The purpose of our application is twofold. First, we study and aim to replicate associations with commonly reported external correlates of C/IER using data with high ecological validity. Second, we use the developed indicator to leverage the unique opportunities provided by two large-scale survey experiments implemented in the PISA 2022 field trial and evaluate whether different scale formats—namely (a) the use of frequency instead of agreement scales and (b) approximate versus abstract frequency scale labels—are associated with different risk of C/IER occurrence.

In what follows, we first briefly discuss the advantages and limitations of current C/IER identification approaches. We highlight that different behavioral C/IER indicators possess different opportunities and pitfalls, and that their combination in an ensemble approach may, therefore, facilitate more robust identification of C/IER. Second, we review research on survey and scale characteristics associated with the occurrence of C/IER and delineate the need for a deeper understanding of scale design features associated with the occurrence of C/IER. In the main body of this study, we then address both issues by (a) developing an ensemble indicator of C/IER, (b) replicating associations of the proposed indicator with commonly reported external correlates of C/IER, and (c) employing the developed indicator to study the effects of changes of scale characteristics on the occurrence of C/IER.

Indicator-based approaches for C/IER

C/IER identification approaches leverage response patterns and/or collateral behavioral data such as information on timing or mouse movements. C/IER identification approaches have in common that they scan the employed data for aberrances, but differ in how these aberrances are conceptualized and identified. In the present study, we focus on indicator-based approaches, i.e., approaches that compress specific types of aberrances in response patterns and/or collateral behavioral data into C/IER indicators. These are widely applied in practice and used for both studying survey design characteristics for informing C/IER prevention (e.g., Robie et al., 2022; Magraw-Mickelson et al., 2020; Huang et al., 2015; Nichols & Edlund, 2020; McKay et al., 2018; Bowling et al., 2016; Galesic & Bosnjak, 2009; Ward & Meade, 2018) and C/IER adjustment in substantive research, informing the exclusion or downweighting of presumed C/IER respondents (see Smith et al., 2020; Landers et al., 2017; Hamari et al., 2019; Curry et al., 2019;

[1] We acknowledge that in the context of individual diagnostics, there may be applications where researchers may want to increase sensitivity at the cost of lowered specificity or vice versa. For instance, practitioners may want to have an individual re-take an assessment due to data validity concerns only in the case that there is very strong evidence that the individual engaged in C/IER. In this case, researchers may opt for an identification method with high specificity, potentially at the cost of lowered sensitivity.

Mitchell et al., 2015, for recent examples from different social science disciplines).[2] In the following, we will briefly discuss opportunities and pitfalls of both established and recently developed C/IER indicators and will delineate their potential sources of C/IER classification error.

Attention check, bogus, and instructional manipulation check items

A common technique for identifying C/IER is to administer attention check, instructional manipulation check, or bogus items. These are items that researchers presume attentive respondents will answer in the same way (Curran, 2016; Meade & Craig, 2012, e.g., disagreement with "I have never brushed my teeth" or compliance with the prompt "Choose response option 5"). A response other than the expected one signals C/IER. When employed for drawing inferences on C/IER contamination, the proportion of respondents who fail on these items is used as an estimate. When used for adjustment, failed respondents are eliminated from further analyses.

Advantages and limitations

Failure to pass attention check, instructional manipulation check, or bogus items conveys straightforward interpretation. The same, however, does not hold true for passing such items, which is often possible with a high chance level. As such, obtained estimates of C/IER proportions can be perceived as a lower bound, and data cleaned from C/IER are likely to still contain contaminated response patterns. Further, administering items with the sole purpose of measuring C/IER imposes additional burden on respondents. Attention check, instructional manipulation check, and bogus items pose obtrusive measures of C/IER, and extensive use might confuse attentive respondents (Meade & Craig, 2012). Hence, in a questionnaire, such items should be used judiciously. When the aim is to get a coarse estimate of how C/IER evolved across the questionnaire (as in Bowling et al., 2009; Galesic & Bosnjak, 2020), this implies that C/IER contamination can only be inferred for the few positions of a questionnaire where attention check, instructional manipulation check, or bogus items have been administered.

Response-pattern-based indicators

Response-pattern-based indicators scan respondents' response vectors for patterns arising from C/IER. Prominent examples are the long-string index as an indicator of response invariability, i.e., the longest sequence of subsequently occurring identical responses (Johnson, 2005), within-person correlations on psychometric synonym and/or antonym item pairs as indicators of response inconsistency (Curran, 2016; Jackson, 1976; Meade & Craig, 2012), Mahalanobis distance, aiming to detect outliers that deviate from typical response patterns (Curran, 2016; Huang et al., 2012), or high proportions of item omissions (Galesic & Bosnjak, 2009; Boe et al., 2002). Exhaustive overviews

[2] A second, just emerging class of approaches for identifying C/IER leverages item response theory (IRT) mixture modeling (Arias et al., 2020; Ulitzsch et al., 2021b, 2023, 2022; van Laar & Braeken, 2022) These approaches assume observed item responses to stem from mixtures of C/IER and attentive responding and formulate different measurement models for either type of item response. Response times can be employed to facilitate the separation of attentive and C/IER classes. Due to their complexity, large computational footprint, and heavy constraints on employable analysis models for presumed attentive responses, however, these approaches have yet not been employed in applied settings.

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

Typically, some threshold is set that encodes researchers' beliefs on values signaling C/IER. Note that the validity of any response-pattern-based indicator drastically hinges on the chosen threshold, i.e., depending on the employed threshold, the same response-pattern-based indicator can be an adequate tool for C/IER identification or essentially useless for this purpose. When employed for drawing inferences on the level of C/IER contamination, the proportion of respondents failing the set threshold is used as an estimate. For C/IER adjustment, failed respondents are eliminated from further analyses.

Advantages and limitations

Response-pattern-based indicators are the most universally applicable C/IER identification technique. They can be obtained unobtrusively and without changes to the survey design and do not require recording additional behavioral information such as timing data. They boast straightforward implementation and can easily be integrated with standard data pre-processing procedures.

A major limitation of response-pattern-based indicators is their sensitivity to threshold settings. There are no globally applicable rules-of-thumb 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. As such, any choice of threshold will ultimately remain arbitrary. What is more, for all indicators, distributions of attentive behavior and C/IER are likely to overlap; hence, misclassifications are inevitable. For instance, when items are homogeneous and worded in the same direction, a high value on the long-string index is plausible under attentive behavior and C/IER alike.

Further, response-pattern-based indicators are limited in that each indicator is sensitive to a different aspect of C/IER, but may be insensitive to others. The long-string index, for instance, performs well in detecting straight lining (as in Fig. 1b), but fails to detect other forms of C/IER such as random responding or diagonal lining (as in Fig. 1c). Conversely, consistency indicators such as the within-person correlation on psychometric synonyms are insensitive to straight lining since this results in consistency of response patterns (Curran, 2016). Accordingly, when applied to both empirical and simulated data, different methods may disagree in their C/IER identification (Meade & Craig, 2012; Niessen et al., 2016). As a remedy for this issue, Curran (2016) suggested a multiple-hurdle approach. In this ensemble approach, information from multiple indicators—ideally sensitive to different aspects of C/IER—is combined and respondents with extreme values on any of the considered indicators are flagged. Nevertheless, Ulitzsch et al. (2021b) illustrated that the multiple-hurdle approach, too, is highly contingent on the thresholds employed for each of its constituting components.

Timing-based indicators

With the widespread use of computer-administered questionnaires, the time respondents spent on each screen can easily be recorded. Approaches leveraging this additional source of behavioral information rest on the belief that C/IER is less time consuming

than attentive responding. Note that the majority of these approaches consider times spent on screens that jointly administer multiple items. Hence, as response-pattern-based indicators, timing-based indicators are used to draw conclusions on C/IER occurrence in a group of items. While it is theoretically possible to reconstruct item-level response times for items administered on a joint screen using the time stamps of provided item responses (e.g. Kroehne et al., 2019), this is hardly done in practice. First, standard data collection platforms typically do not provide time stamps of single item responses, and researchers need to write their own plug-ins to collect them. Second, and more importantly, even if time stamps are available, the reconstruction of item-level response times relies on strong assumptions of how respondents proceed through the items, e.g., that items are only read once a response to the precedingly answered item has been provided.

Threshold-based techniques

A common method for compressing screen time information into a binary C/IER indicator is to set a threshold based on an educated guess on the minimum amount of time required for providing attentive responses to the items of a given screen (e.g., 2 s per item, Huang et al., 2012; see Bowling et al., 2021, for investigations of the construct validity of this threshold) and classifying respondents falling below this threshold as careless.

Advantages and limitations One of the major advantages of timing-based over response-pattern-based indicators is that these do not entail presumptions on the specific C/IER patterns. They further allow to incorporate subject-matter considerations on the minimal amount of time it requires to respond to an item in an attentive manner. Nevertheless, as response-pattern-based indicators, traditional timing-based indicators are sensitive to the employed thresholds and prone to misclassifications whenever attentive and C/IER screen time distributions overlap.

Mixture modeling techniques

Recently, Ulitzsch et al. (2024) provided a data-driven screen time decomposition approach that circumvents the setting of time thresholds. This approach draws on mixture modeling to decompose log screen time distributions into several subcomponents, out of which the subcomponent with the lowest mean is assumed to stem from C/IER. The mixing proportion of the C/IER subcomponent is used as an estimate of the C/IER contamination rate. For adjustment, Ulitzsch et al. (2024) proposed to obtain posterior C/IER class probabilities for each respondent and use their negations as person weights in the subsequent analysis of response patterns. Effectively, this procedure downweighs response patterns according to their presumed probability of stemming from C/IER. Note that in contrast to other indicator-based approaches, the screen decomposition approach provides a probabilistic instead of a binary C/IER indicator.

Advantages and limitations The screen time decomposition approach overcomes major limitations of threshold-based techniques. This comes, however, at the price of strong distributional assumptions. The screen time decomposition approach rests on the assumption that log screen times factorize into normally distributed subcomponents.

Ulitzsch et al. (2024) and Ulitzsch et al. (2023) illustrated that violations of this assumption may heavily distort conclusions on C/IER and, as a consequence, C/IER adjustment. For instance, in the extreme case that there is no C/IER in the data, but the attentive distribution is heavy-tailed, the mixture decomposition will likely capture this heavy-tailedness in terms of two normal distributions with almost equal means but different variances (see Ulitzsch, Domingue, et al., 2023, for further illustrations and evaluations). Then, the component with the lower mean will be artefactually labeled as C/IER. Further, whenever there are at least two subcomponents, the screen time decomposition approach relies on the strong assumption that there is one and only one C/IER component. Both the absence of C/IER in the face of multiple attentive components as well as the presence of multiple C/IER subcomponents pose plausible violations of this assumption.

Of course, it can never be determined with certainty whether short screen times indeed stem from C/IER or are reflective of other phenomena. The distance-difficulty hypothesis, for instance, states that respondents being very sure of their answers may be able to provide their responses faster (Ferrando & Lorenzo-Seva, 2007). Neither threshold-based techniques nor the mixture decomposition approach is capable of disentangling these phenomena.[3]

Scale characteristics and prevalence of C/IER

Research on the association between survey design features and C/IER occurrence has primarily focused on characteristics of the survey as a whole—such as survey mode (Magraw-Mickelson et al., 2020; Bowling et al., 2020), questionnaire length (Gibson & Bowling, 2019; Galesic & Bosnjak, 2009; Eisele et al., 2022), or instructions (Marshall, 2019; Ward et al., 2018). Research on less aggregate features (i.e., on the item or scale level), in contrast, is scarce and has predominantly focused on position effects. The position effect describes the phenomenon that items and scales administered at later positions are more affected by C/IER and is one of the best-documented effects in the literature on C/IER occurrence. It has been reported for various types of surveys—ranging from paper-and-pencil questionnaires through online studies to educational large-scale assessment background questionnaires—and has been identified based on a broad array of C/IER detection methods—ranging from self reports and bogus items through classical pattern-based indicators to mixture item response theory (IRT) modeling approaches (Ulitzsch et al., 2022, 2024; Berry et al., 1992; Baer et al., 1997; Galesic & Bosnjak, 2009; Bowling et al., 2020).

A straightforward way to avoid C/IER due to item position effects is to administer shorter questionnaires. Shortening questionnaires to avoid C/IER occurrence, however, imposes heavy constraints on planned research designs. Hence, expanded knowledge on scale characteristics associated with the occurrence of C/IER that can be leveraged with more minimal interference with planned research designs is urgently needed. A prominent set of scale features that can easily be adapted without severe interference with planned research designs is the scale format as characterized, e.g., by the number of

[3] Ulitzsch et al. (2021b) developed a mixture IRT model incorporating item-level response times that takes such complexities in the identification of C/IER into account. Due to the model's complexity and the common unavailability of item-level response times, however, this model is of limited practicability for large-scale survey settings.

response categories, response category labels, or the use of frequency instead of agreement scales. It is well documented that different response formats provide data of different quality and exhibit different psychometric properties (see DeCastellarnau, 2018, for an overview). Further, from the response bias literature on mid-point, extreme, and acquiescent response styles, it is known that the extent to which scales and items are affected by response styles is related to rating scale format (e.g. Weijters et al., 2010; Deng & Bolt, 2016; Moors et al., 2014; Kieruj & Moors, 2013; Hui & Triandis, 1989; Henninger & Meiser, 2020). Little is known, in contrast, on whether manipulations of scale format can be leveraged to curb the occurrence of C/IER (see Robie et al., 2022, for a recent exception studying the effect of response option order.) It can, however, be speculated that different scale formats differ in the extent to which they scaffold and guide respondents' retrieval of relevant information and/or the process of mapping one's judgment onto an adequate response option (see Tourangeau et al., 2000; Krosnick, 1991, for discussions and cognitive theory of survey response processes), thus, in the imposed cognitive burden and, as a consequence in the extent to which C/IER is elicited.

To start filling this gap and provide practical guidance for questionnaire design for large-scale surveys, in the present study, we leverage a large-scale survey experiment implemented in the PISA 2022 field trial and investigate whether (a) the use of frequency instead of agreement scales and (b) the of use approximate instead of abstract frequency scale labels are associated with different risks of C/IER occurrence. It may be speculated that both manipulations hold the potential to scaffold the response process by making it easier for respondents to map the item's statements and response options to their daily experiences. This, in turn, could lower cognitive burden and, consequently, C/IER.

The present study

The objective of the present study comprises three parts. First, we aim to develop a robust C/IER indicator that requires agreement on C/IER identification from multiple behavioral sources, thereby alleviating the effects of each source's standalone C/IER misidentifications.

Second, we study associations between the developed indicator and commonly reported external correlates of C/IER and investigate whether we can replicate (a) convergence between C/IER identified based on behavioral data and self-reported effort (Meade & Craig, 2012; Douglas et al., 2023) and (b) the position effect of C/IER (Ulitzsch et al., 2022, 2024; Berry et al., 1992; Baer et al., 1997; Galesic & Bosnjak, 2009; Bowling et al., 2020). Findings aligning with these expectations will not indicate that the proposed indicator provides more valid C/IER detection than previously developed techniques. Failure to replicate these effects, however, may give rise to the suspicion that the proposed combination of previously developed indicators results in less trustworthy C/IER detection.

Third, we use the developed indicator to investigate the effect of changes in scale format characteristics on the occurrence of C/IER. We focus on two aspects of scale format—frequency instead of agreement scales and approximate instead of abstract labeling of frequency-type scales—and investigate their effects on C/IER contamination risk through the following explorations

*E*1 (a) Does the occurrence of C/IER differ for scales with agreement and frequency formats and (b) do potential effects appear consistently across country and economy groups?

*E*2 (a) Does the occurrence of C/IER on scales with frequency formats differ for abstract and approximate frequency labels and (b) do potential effects appear consistently across country and economy groups?

Data

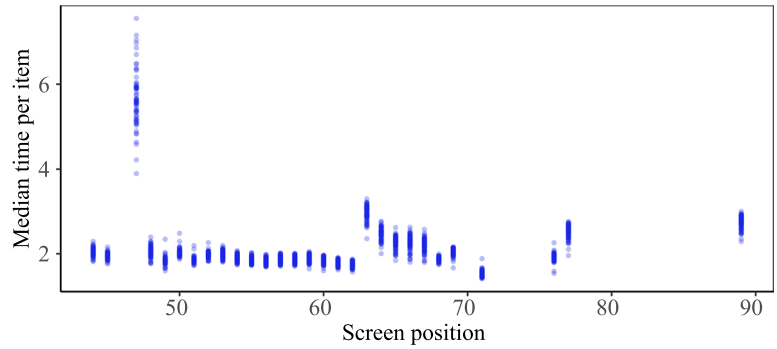
We based our analyses on data from the computer-administered PISA 2022 field trial student questionnaire. In total, the data set comprised item responses and screen times from 206,153 students from 75 country and economy groups, with group-level sample size ranging from 309 to 7,140. Students were administered one of two booklets comprising different sets of scales. To further reduce respondent burden, the PISA 2022 field trial student questionnaire implemented an incomplete block design for longer scales (i.e., each student received only a fraction of the scale's items). Although many scales contained items with both negative and positive wording, due to the incomplete block design, not all respondents were administered items with different wording for the same scale. Each scale was administered on a separate screen. We focused our analyses on scales with a closed response format, at least two response options, and at least three items, and analyzed only scales that were common to all groups. This resulted in 31 and 33 scales from booklet 1 and 2, respectively. For booklet 1, students were administered 5 items for 72% of the considered scales; 5% of the scales contained 6 to 8 items, and the remaining scales comprised 3 or 4 items. For booklet 2, students were administered 5 items for 73% of the considered scales. The remaining scales comprised 3 or 4 items.

For booklet 1, the median time per item was 1.92 s with an interquartile range of [1.82; 2.10]. For booklet 2, the median time per item was 1.95 s with an interquartile range of [1.83; 2.10]. However, as evidenced in Fig. 2, median time per item considerably varied across groups and, even more so, across scales.[4]

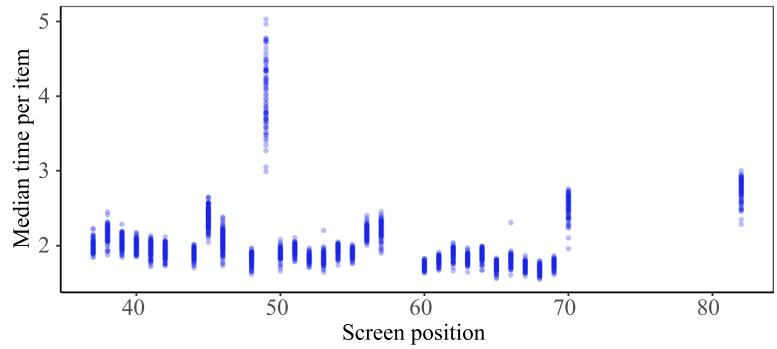
As a measure of test-taking effort, we considered the third item of the PISA scale on effort on the achievement test and questionnaire, which asks students to gauge the amount of effort they put into giving accurate answers on the questionnaire on a 10-point scale.

The PISA 2022 field trial implemented two experimental manipulations of scale formats. The original purpose of the manipulations was to investigate whether these could improve cross-country comparability. In booklet 1, 10 scales concerning general social and emotional characteristics (such as assertiveness, cooperation, perseverance, and trust) were administered either with a standard agreement-type scale or asked students to recall the frequency with which they felt, thought, or acted in specific ways. Both scale formats were implemented with five response options but different labels. That is, for the same statement (e.g., "I believe most people are kind" for measuring trust) response options were either labeled using agreement ("strongly disagree", "disagree",

[4] The two screens with markedly increased median times per item administered scales measuring bullying (booklet 1; stem: "During the past 12 months, how often have you had the following experiences in school?") and familiarity with mathematical concepts (booklet 2; stem: "Thinking about mathematical concepts: how familiar are you with the following terms?").



(a) Booklet 1



(b) Booklet 2

Fig. 2 Screen-by-group-level median time per item plotted against screen position. Note that only scales with a closed response format, at least two response options, and at least three items were analyzed and that each scale was administered on a separate screen

"neither agree nor disagree", "agree", "strongly agree") or frequency ("never or almost never", "less than half of the time", "about half of the time", "more than half of the time", "all or almost all of the time") labels. In booklet 2, six frequency-based scales concerning exposure to mathematics content and mathematics teacher behavior (e.g., "The teacher pointed out mistakes in my mathematics work" for measuring teacher feedback) were either administered with abstract response anchors used in previous PISA cycles (e.g., "never", "rarely", "sometimes", "frequently") or concrete response anchors (e.g., "never or almost never", "about once or twice a year", "about once or twice a month", "about once or twice a week", "every day or almost every day").

Analysis strategy

Part I: Developing a multiple-source indicator of C/IER

The proposed multiple-source indicator is rooted in the rationale that behavioral indicators of C/IER stemming from multiple sources—more specifically, response patterns and screen times—should agree in the identification of C/IER, thereby increasing robustness against misclassifications based on each of the behavioral sources alone. To this end, we combine the response-pattern-based multiple-hurdle approach suggested by Curran (2016) with the screen-time-based decomposition approach proposed in Ulitzsch et al. (2024).

Response-pattern-based component

The multiple-hurdle approach by Curran (2016) can be formalized by associating with each response vector $\mathbf{x}_{is}$ of person $i \in \{1, \ldots, N\}$, on scale $s \in \{1, \ldots, S\}$, a set of indicator functions $\mathbf{I}(\mathbf{x}_{is})$. Each indicator function $I_j(\mathbf{x}_{is})$, $j \in \{1, \ldots, J\}$, stores information on whether or not response pattern $\mathbf{x}_{is}$ is classified as C/IER using a specific response-pattern-based indicator. In the present study, we consider $J = 3$ response-pattern-based indicators of C/IER. The first indicator function $I_1(\mathbf{x}_{is})$ encodes C/IER classification based on the proportion of item omissions and takes the value 1 if person i omitted all K_s items of scale s and is 0 otherwise. The second indicator function $I_2(\mathbf{x}_{is})$ is based on the long-string index and takes the value 1 if this index equals K_s and is 0 otherwise (i.e., in the case respondent i chose the same response option on all K_s items of scale s). The third indicator function $I_3(\mathbf{x}_{is})$ is based on Mahalanobis distance and takes the value 1 if respondent i's squared Mahalanobis distance for scale s exceeds the 99th quantile of the χ^2 distribution with K_s degrees of freedom and is 0 otherwise. We combine these indicator functions into a response-pattern-based multiple-hurdle indicator of C/IER d_{is}^{MH} as follows

$$d_{is}^{\text{MH}} = \begin{cases} 1 \text{ if } \sum_{j=1}^{J} I_j(\mathbf{x}_{is}) \geq 1 \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

That is, given a response pattern $\mathbf{x}_{is}$, d_{is}^{MH} evaluates to 1 if at least one of the J indicator functions evaluates to 1 and is 0 otherwise.[5] The hope in combining different response-pattern-based indicators is that false negatives due to single indicators' lack of sensitivity to some aspects of C/IER are avoided by balancing these sensitivities to different behavioral aspects in a carefully chosen ensemble of indicators.

Screen-time-based component

In the screen time decomposition approach, for each screen s, screen times are decomposed into C_s subcomponents by means of Gaussian mixture models. The number of components is determined using the Bayesian information criterion (BIC). To ease the interpretation of C/IER screen time means and accommodate the fact that due to the incomplete block design respondents may have been administered different numbers of items, following Ulitzsch et al. (2024), we considered the geometric mean time averaged across the number of items presented on the given screen. That is, the screen time st_{is} respondent i spent on screen s is transformed as $st_{is}^{\frac{1}{K_s}}$. We denote the geometric mean time per item of respondent i for screen s with t_{is}. Transformed screen times from subcomponent $c_s \in \{1, \ldots, C_s\}$ are assumed to be distributed as

$$\ln(t_{is}|z_{is} = c_s) \sim \mathcal{N}(\mu_{c_s}, \sigma_{c_s}^2), \tag{2}$$

with $z_{is} \in \{1, \ldots, C_s\}$ denoting respondent i's unobserved component membership, and μ_{c_s} and $\sigma_{c_s}^2$ giving the subcomponent's mean and variance. Then, the marginal distribution of $\ln(t_{is})$ is

[5] Note that the construction of this indicator slightly differs from the multiple-hurdle approach by Curran (2016) where data are screened sequentially with different indicators.

$$f(\ln(t_{is})) = \sum_{c_s=1}^{C_s} \pi_{c_s} \frac{1}{\sigma_{c_s}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(t_{is}) - \mu_{c_s}}{\sigma_{c_s}}\right)^2\right), \tag{3}$$

where $\pi_{c_s} = p(z_{is} = c_s)$ denotes the mixture proportion for component c_s.

In the case that at least two components are obtained (i.e., $C_s \geq 2$), the component with the lowest mean is assumed to comprise of screen times associated with C/IER and labeled accordingly, that is

$$c_s^{\text{C/IER}} = \underset{c_s \in \{1,\ldots,C_s\}}{\arg\min}\ \mu_{c_s}. \tag{4}$$

The mixture proportion associated with the lowest mean $\pi_s^{\text{C/IER}}$ is assumed to give the respondent-level proportion of C/IER on the considered screen. The remaining components are assumed to stem from attentive responding. Posterior C/IER class probabilities $\pi_{is}^{\text{C/IER}} = p(z_{is} = c_s^{\text{C/IER}} | \ln(t_{is}))$ are determined for each respondent and employed as a probabilistic C/IER indicator. Following Ulitzsch et al. (2024), to avoid screen time decompositions that reflect differences in language-specific time requirements rather than differences in response processes, we analyzed screen times separately for each group.

Multiple-source indicator

For obtaining the proposed multiple-source indicator of C/IER q_{is}, we enrich the information on aberrances in response patterns contained in d_{is}^{MH} with information on aberrances in screen times contained in posterior C/IER class probabilities $\pi_{is}^{\text{C/IER}}$ as

$$q_{is} = d_{is}^{\text{MH}} \pi_{is}^{\text{C/IER}}. \tag{5}$$

The indicator combines two common assumptions that are tacitly made in most of the literature on C/IER identification, namely that (a) C/IER results in response patterns exhibiting "peculiarities" that can be picked up by different response-pattern-based indicators and that (b) on average, producing these patterns requires less time than generating attentive responses. An indicator that is sensitive to both constituting aspects of this conceptualization of C/IER requires taking information from different behavioral sources into account, namely timing and response-pattern data. The indicator q_{is} signals C/IER to the extent of evidence that both assumptions are met.

The multiple-source indicator q_{is} can be understood as the probability that respondent i exhibited C/IER on screen s—given the information and classification decisions encoded in $\pi_{is}^{\text{C/IER}}$ and d_{is}^{MH}. Multiplying $\pi_{is}^{\text{C/IER}}$ with d_{is}^{MH} ensures that q_{is} stores information on the agreement in C/IER classification of the multiple-hurdle and the screen time decomposition approach. As such, q_{is} will take high values only in the case that both response patterns and screen times exhibit aberrances pointing towards C/IER. If $\pi_{is}^{\text{C/IER}}$ and d_{is}^{MH} disagree on whether or not respondent i displayed C/IER on scale s, q_{is} will take values equal or close to 0. The hope in requiring agreement between $\pi_{is}^{\text{C/IER}}$ and d_{is}^{MH} is that this will dampen the impact of false positives of each of q_{is}'s constituting components, i.e. that it will buffer the impact of threshold settings required for implementing the multiple-hurdle approach and alleviate the consequences of false C/IER labels of the screen time decomposition approach. For instance, if the screen time decomposition

approach attributes a respondent's screen time to C/IER with a high probability but his or her response pattern does not indicate any aberrances (i.e., $d_{is}^{\text{MH}} = 0$), this may indicate a false C/IER label obtained from the screen time decompostion approach and, as such, untrustworthiness of $\pi_{is}^{\text{C/IER}}$. In this case, q_{is} will correspond to 0. Likewise, a respondent choosing the same response option on all items on a given screen will always be classified as C/IER by d_{is}^{MH} (given that the long-string index is considered in its construction). If this response pattern occurred due to behavior other than C/IER (which is a plausible alternative explanation in the case of, say, homogeneous items worded in the same direction), the respondent's time spent on screen should align with what would typically be expected for attentive respondents. This information is encoded in $\pi_{is}^{\text{C/IER}}$ close to zero (given that attentive components are correctly labeled), which, as a consequence, will shrink the impact of d_{is}^{MH} toward zero in the construction of q_{is}.

In Part I of our analyses, we provide intuition for the proposed multiple-source indicator by (a) contrasting conclusions on the occurrence of C/IER drawn from q and each of its constituting components and (b) inspecting correlations of q with each of its components to understand the contribution of each of these to the multi-source detection of C/IER. Further, we (c) illustrate its robustness against too liberal thresholds by exemplary implementing a more liberal threshold for Mahalanobis distance in the construction of $I_3(\mathbf{x}_{is})$, setting the cut-off for the squared Mahalanobis distance at the 95th instead of the 99th quantile of the χ^2 distribution with K_s degrees of freedom. Finally, we (d) explore presumed prevention of false C/IER labels by investigating selected cases of disagreement between q's constituting components and (e) explore potential sources of false C/IER labels by investigating selected counter-intuitive spikes in C/IER trajectories across the PISA field trial questionnaire.

Part II: Investigating associations with external correlates

Relating C/IER to self-reported effort

We investigated within-group correlations between respondents' q values averaged across all considered scales and their self-reported effort. Analyses were conducted separately by booklet.

Relating C/IER to scale position

Following Ulitzsch et al. (2024), we analyzed the relationship between scale position and the (presumed) occurrence of C/IER on the scale-by-group level using Beta regression, which is well suited for proportion data.[6] Analyses were conducted separately by booklet. To accommodate the nesting of scales within groups and account for group-specific baseline propensities to show C/IER, we ran a hierarchical random intercept model with the average q_{sg} on screen s in group $g \in \{1, \ldots, G\}$ being modeled as

$$q_{sg} \sim \text{beta}(\mu_{sg}\phi, (1-\mu_{sg})\phi) \qquad \text{where} \qquad \mu_{sg} = \frac{\exp(\beta_{0g} + \beta_1 x_s)}{1 + \exp(\beta_{0g} + \beta_1 x_s)}. \tag{6}$$

[6] We employed scale position in the full questionnaire (all routes included) as a proxy, but note that due to the routing, data from each scale comprised data from respondents who have been administered the respective scale at different positions.

The parameter β_{0g} gives the group-specific intercept, β_1 denotes the fixed regression weight for the scale's position x_s, and ϕ is a precision parameter. Group-specific intercepts are assumed to be normally distributed with $\mathcal{N}(\mu_{\beta_0}, \sigma_{\beta_0})$. For diffuse prior settings, we adhered to the set-up employed in Ulitzsch et al. (2024). We employed diffuse normal priors with mean 0 and standard deviation 10 for the average intercept μ_{β_0} and the fixed regression weight β_1. The standard deviation of the random intercept σ_{β_0} and the precision parameter ϕ were equipped with half-Cauchy priors with location 0 and scale 5.

Bayesian estimation of the Beta regression was conducted using Stan version 2.19 (Carpenter et al., 2017) employing the rstan package version 2.19.3 (Guo et al., 2018). We ran two Markov chain Monte Carlo (MCMC) chains with 10,000 iterations each, using the first half as warm-up. The sampling procedure was assessed on the basis of potential scale reduction factor (PSRF) values, with PSRF values below 1.05 for all parameters being considered as satisfactory (Gelman & Rubin, 1992; Gelman & Shirley, 2011). We further inspected the effective sample size (ESS) for all parameters, considering an ESS above 400 sufficient to accurately summarize the posterior distribution (Hoff, 2009). We employed the posterior mean (EAP) as a Bayesian point estimate.

Part III: Investigating the effect of scale format on C/IER

For the experimentally manipulated scales, we compared pairs of multiple-source-indicator-implied scale-by-group C/IER proportions across different scale formats. To this end, we first computed the difference between group-level C/IER proportions on scale s administered with different scale formats as $\pi_{s1g}^{\text{C/IER}} - \pi_{s2g}^{\text{C/IER}}$. For the experimental manipulation implemented in booklet 1, pairs comprised agreement ($\pi_{s1g}^{\text{C/IER}}$) and frequency ($\pi_{s2g}^{\text{C/IER}}$) scale formats. For the experimental manipulation implemented in booklet 2, pairs comprised frequency scales with abstract ($\pi_{s1g}^{\text{C/IER}}$) and approximate ($\pi_{s2g}^{\text{C/IER}}$) frequency labels. To test the null that $\pi_{s1g}^{\text{C/IER}} - \pi_{s2g}^{\text{C/IER}} = 0$, we obtained standard errors for differences in C/IER proportions as

$$SE_{sg} = \sqrt{\frac{\pi_{sg}^{\text{C/IER}}\left(1 - \pi_{sg}^{\text{C/IER}}\right)}{N_{s1g}} + \frac{\pi_{sg}^{\text{C/IER}}\left(1 - \pi_{sg}^{\text{C/IER}}\right)}{N_{s2g}}}, \tag{7}$$

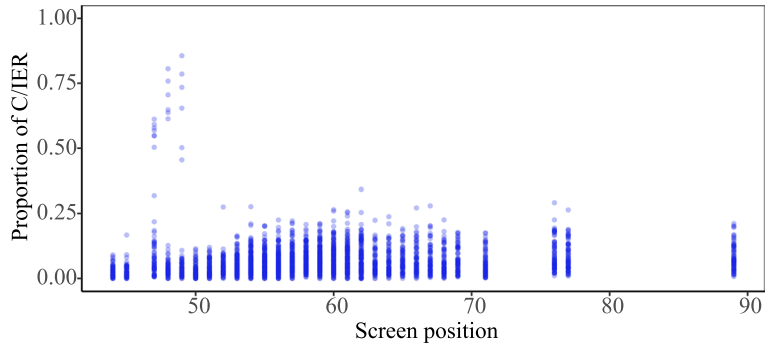
where N_{s1g} and N_{s2g} denote the number of respondents administered scale format 1 and 2, respectively, and

$$\pi_{sg}^{\text{C/IER}} = \frac{\pi_{s1g}^{\text{C/IER}} N_{s1g} + \pi_{s2g}^{\text{C/IER}} N_{s2g}}{N_{s1g} + N_{s2g}} \tag{8}$$

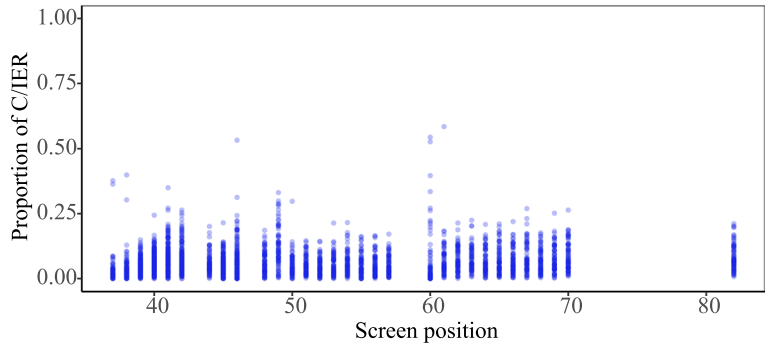
gives the pooled estimate of the sample C/IER proportion. Using the test statistic

$$z = \frac{\pi_{s1g}^{\text{C/IER}} - \pi_{s2g}^{\text{C/IER}}}{SE_{sg}} \tag{9}$$

we tested each group-by-scale difference for significance with a .05 confidence level.



(a) Booklet 1



(b) Booklet 2

Fig. 3 Multiple-source-indicator-implied group-by-screen-level C/IER proportions plotted against screen position. Note that only scales with a closed response format, at least two response options, and at least three items were analyzed and that each scale was administered on a separate screen

Results

Part I: Developing a multiple-source indicator of C/IER

Figure 3 displays group-by-screen-level C/IER proportions plotted against screen position for both booklets. For both booklets, median C/IER proportions oscillated around .05, with a slight increase with later scale positions. Scale position effects are explored in greater detail in Part II.

Contrasting C/IER proportions implied by multiple-source and single-source indicators

We first contrast conclusions on the overall occurrence of C/IER implied by the multiple-source indicator q against C/IER occurrence implied by q's constituting components. In the diagonal, Table 1 displays medians and interquartile ranges of scale-by-group C/IER proportions separately for each booklet implied by different C/IER indicators. Median C/IER proportions implied by q's constituting components ranged from .04 to .11 in booklet 1 and from .04 to .20 in booklet 2. Low correlations between the majority of d^{MH}'s constituting components suggest that different response-pattern-based indicators flagged different respondents. This is not surprising, given that each indicator is sensitive to different behavioral aspects of C/IER. Note that both I_2 and I_3 evaluate to 0 when all item responses are missing, and I_1 evaluates to 0 when I_2 and I_3 can meaningfully be obtained.

Table 1 Median C/IER proportions across scale-by-group combinations implied by and correlations among different C/IER indicators

Booklet 1					
	I_1	I_2	I_3	$\pi^{C/IER}$	q
I_1	.05				
	[.02; .10]				
I_2	-.07	.09			
	[-.13; -.04]	[.05; .14]			
I_3	-.05	.00	.04		
	[-.07; -.03]	[-.05; .09]	[.02; .07]		
$\pi^{C/IER}$	.79	-.04	.00	.07	
	[.65; .87]	[-.11; .15]	[-.01; .08]	[.04; .12]	
q	.83	-.02	.08	.99	.05
	[.72; .89]	[-.10; .21]	[.00; .19]	[.88; 1.00]	[.03; .09]
Booklet 2					
	I_1	I_2	I_3	$\pi^{C/IER}$	q
I_1	.05				
	[.02; .09]				
I_2	-.11	.20			
	[-.18; -.07]	[.14; .29]			
I_3	-.05	-.03	.04		
	[-.07; -.03]	[-.10; .04]	[.02; .07]		
$\pi^{C/IER}$	.76	-.06	-.01	.07	
	[.59; .86]	[-.08; .29]	[-.02; .05]	[.03; .12]	
q	.80	-.04	.07	.98	.05
	[.67; .88]	[-.11; .13]	[.00; .17]	[.87; 1.00]	[.03; .09]

•*Notes:* Median scale-by-group-level C/IER proportions are displayed in the diagonal and median scale-by-group-level correlations between different indicators are displayed in the off-diagonal. Interquartile ranges are given in squared brackets. I_1 indicates whether or not a respondent omitted all items of a given scale; I_2 encodes C/IER classification based on the long-string index and indicates whether or not respondents chose the same response option on all items of a given scale; I_3 encodes C/IER classification based on Mahalanobis distance and indicates whether or not respondents' response vectors on a given scale were classified as an outlier on a .01 confidence level; $\pi^{C/IER}$ gives respondents' posterior C/IER class probabilities obtained from the mixture decomposition approach; q denotes the proposed multiple-source C/IER indicator

Investigating the contribution of single indicators to multiple-source detection of C/IER

Inspecting the correlations between q and each of its constituting components displayed in Table 1 supports understanding the contribution of each of these to the multiple-source detection of C/IER. For both booklets, I_1 (omitting all items) and posterior C/IER class probabilities $\pi^{C/IER}$ obtained from the screen time decomposition approach exhibited the by far strongest correlations with q, indicating (a) that respondents identified as displaying C/IER oftentimes tended to omit all items administered, (b) that in the case that d^{MH} evaluated to 1 because of omission behavior, agreement between d^{MH} and $\pi^{C/IER}$ tended to be high (i.e., respondents omitting all items tended to have high posterior C/IER class probabilities), and (c) that aberrant screen times as identified by the screen time decomposition oftentimes were accompanied by aberrances in response patterns as detected by d^{MH}.

The low correlations of q with I_2 (choosing the same response option on all items administered) and I_3 (outlier detection based on Mahalanobis distance) indicate that the detected aberrances in response patterns were oftentimes not accompanied by

aberrantly short screen times as identified by the screen time decomposition. Therefore, I_2 and I_3 did not fully contribute to the multiple-source detection of C/IER. At the same time, however, their correlation with q exhibited strong variation across scale-by-group combinations, as indicated by rather broad interquartile ranges. This suggests that the extent to which the considered indicators contributed to q strongly varied across different scale-by-group combinations, and highlights that the informativeness of single behavioral indicators regarding C/IER may be scale-dependent. For instance, the long-string index's informativeness on C/IER can be assumed to vary as a function of, among others, scale length, item homogeneity, and whether or not all items are worded in the same direction. On a long scale comprising items of different polarity, the long-string index has high face validity as in indicator of C/IER. On a rather short scale with homogeneous items, however, interpretation of the long-string index is less clear, as response vectors comprising one response option only are also plausible to occur under attentive responding. In the present application, for most scales, respondents were administered at most five items and negatively worded items were rarely employed. Hence, the long-string index can be assumed to be an ambiguous indicator on the majority of the considered scales.

To provide further intuition for the multiple-source indicator q, Fig. 4 displays group-level C/IER proportions implied by different indicators for an exemplary selected scale (ST311; measuring empathy with an agreement scale). Groups are sorted by the C/IER proportion implied by the multiple-source indicator q. Figure 4 illustrates how q integrates information from multiple behavioral sources. C/IER proportions implied by q's constituting components falling below the proportions implied by q suggest that the respective indicator may have underidentified C/IER in a given group, while C/IER proportions falling above the proportions implied by q indicate potential overidentification. Recall that both under- and overidentification are indicated by disagreement between the different components involved in the construction of q. As can be seen, directions of (presumed) misclassifications for all response-pattern-based indicators varied across groups. This was different for posterior C/IER class probabilities $\pi^{\text{C/IER}}$ obtained from the screen time decomposition approach, which either agreed with q or exhibited (presumed) C/IER overidentification. This is because the response-pattern-based indicators are first combined in the multiple-hurdle indicator d^{MH} and aberrances flagged by one of the involved indicators may be left undetected by the others. Whenever d^{MH} evaluates to 0, however, q will evaluate to 0 regardless of the value obtained for $\pi^{\text{C/IER}}$.

Illustrating robustness against too liberal threshold settings

Figure 4 also illustrates q's robustness against (presumably) too liberal threshold settings employed in the construction of the multiple-hurdle indicator d^{MH}, which we see as one of the major advantages of combining information from different behavioral sources for C/IER detection. This robustness can be studied by comparing C/IER proportions implied by components of d^{MH} implemented with different thresholds—in the present case, outlier detection based on Mahalanobis distance with a .01 (I_3) and .05 confidence level (I_3^{L})—with those implied by the multiple-source indicator q considering these indicators with either threshold (denoted with q when considering I_3 and with q^{L} when considering I_3^{L}). FigURE 4 indicates how minor differences

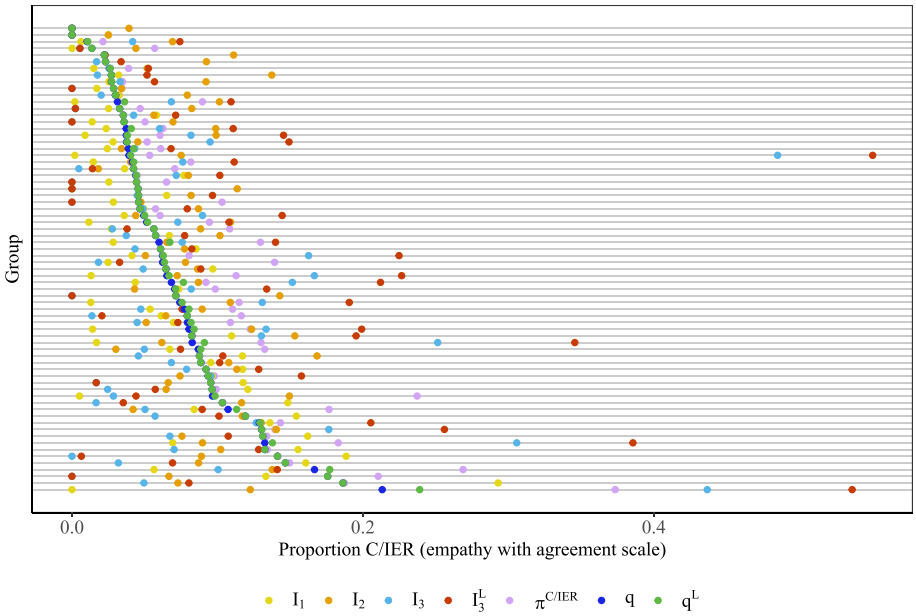
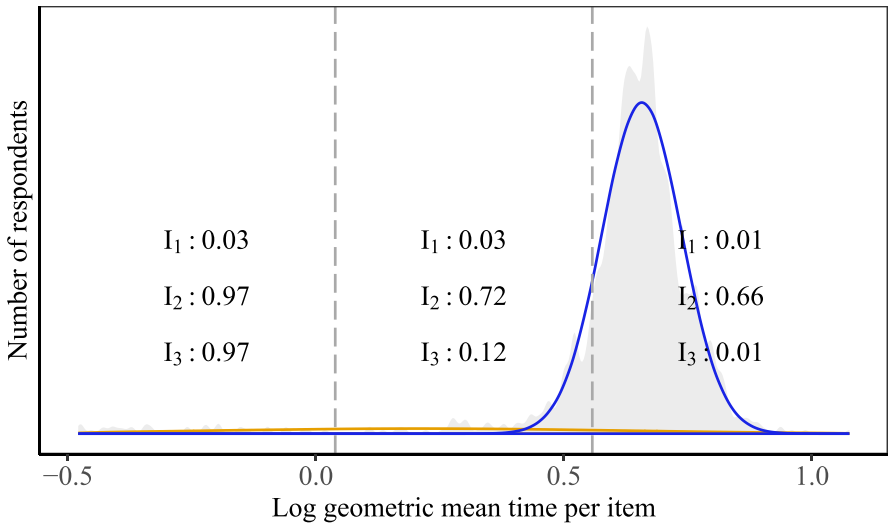


Fig. 4 Group-level careless and insufficient effort responding (C/IER) proportions for an exemplary selected scale (ST311; measuring empathy with an agreement scale) implied by different indicators. Groups are sorted by the C/IER proportions implied by the multiple-source indicator q. I_1 indicates whether or not a respondent omitted all items; I_2 encodes C/IER classification based on the long-string index and indicates whether or not respondents chose the same response option on all items; I_3 encodes C/IER classification based on Mahalanobis distance and indicates whether or not respondents' response vectors were classified as an outlier on a .01 confidence level; I_3^L encodes C/IER classification based on Mahalanobis distance with a more liberal threshold and indicates whether or not respondents' response vectors were classified as an outlier on a .05 confidence level; $\pi^{C/IER}$ gives respondents' posterior C/IER class probabilities obtained from the mixture decomposition approach; q denotes the proposed multiple-source C/IER indicator; q^L denotes the proposed multiple-source C/IER indicator considering I_3^L instead of I_3

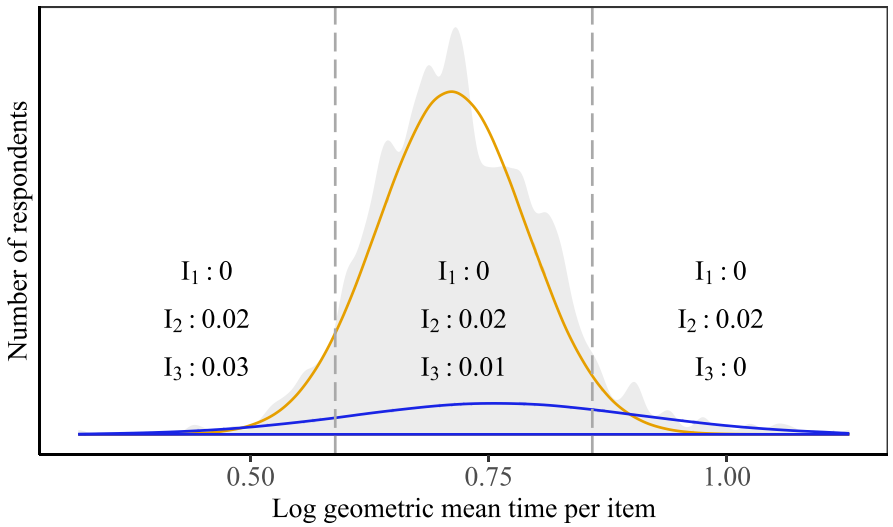
in threshold settings for outlier detection through Mahalanobis distance may heavily impact conclusions on C/IER. In the present example, I_3 and I_3^L identified median C/IER proportions across all groups of .05 (interquartile range: [.02; .08]) and .08 [.04; .14], respectively. q and q^L, in contrast, were not heavily impacted by these different choices of threshold settings in one of their constituting components, both yielding a median C/IER proportion of. 04 [.08; .11] and exhibiting a correlation of C/IER proportions of .99.

Exploring cases of presumed prevention of false C/IER labels

To further illustrate the added value of combining different data sources and requiring agreement in C/IER detection, Fig. 5 provides examples of selected scale-by-group combinations from booklet 1 where screen-time-based and response-pattern-based C/IER identification disagreed. In both figures, distributions of log geometric mean time per item of the presumed C/IER (orange) and attentive (blue) classes implied by the mixture decomposition are superimposed on the observed distributions. Proportions of respondents being classified as inattentive using item omissions (I_1), the long-string index (I_2), and Mahalanobis distance (I_3) are given for three time bins, marked with dashed vertical lines.



(a) Screen ST266 with low agreement between screen-time-based C/IER identification and the long-string index.



(b) Screen ST269 with low agreement between screen-time-based and response-pattern-based C/IER identification.

Fig. 5 Observed distributions of log geometric mean time per item for three selected scale-by-group combinations. Distributions of the presumed C/IER (orange) and attentive (blue) classes implied by the mixture decomposition approach are superimposed. Proportions of respondents failing response-pattern-based indicates are given for three time bins, marked with dashed vertical lines. I_1: item omissions; I_2: long-string index; I_3: Mahalanobis distance

In Fig. 5a, depicting an example of strong agreement of screen-time-based C/IER identification with Mahalanobis distance but low agreement with the long-string index, aberrantly short screen times as identified by the mixture decomposition tended to

co-occur with response patterns that were flagged by either the long-string index or Mahalanobis distance. In contrast, only very few response patterns associated with regular screen times contained irregularities detectable by Mahalanobis distance. The long-string index, however, still flagged a considerable proportion of respondents with regular screen times. It can be speculated that these were misclassified by the long-string index. Based on the long-string index only, 64% of the respondents were flagged. The proposed indicator, however, considered information from the long-string index only for respondents with aberrantly short screen times and indicated a much lower C/IER rate of 6%.

FigURE 5b depicts an example where a unimodal but heavy-tailed screen time distribution was decomposed into two normal distributions and the component with the slightly lower mean was labeled as C/IER. The response patterns of this scale-by-group combination, however, did hardly exhibit any aberrances. The mixture decomposition indicated a C/IER rate of 85%. The proposed indicator was capable to alleviate the presumably artefactual conclusions on C/IER, yielding a much lower C/IER rate of 4%.

Exploring potential sources of false C/IER labels

Following Ulitzsch et al. (2024), we investigated exemplary selected scale-by-group combinations with counter-intuitively high C/IER proportions. We observed the highest C/IER proportion of .86 on screen ST266 administered at position 49 in booklet 1 (see Fig. 3a). ST266 comprises five items with a yes/no format, asking respondents to state whether or not during the past four weeks they encountered phenomena related to low school safety, e.g., gangs in school. The high C/IER proportion was driven by agreement of the mixture decomposition component, yielding a C/IER proportion of .86, and the long-string index component, yielding a C/IER proportion of .82, i.e., 82% of respondents chose the same response option on all five items. Closer inspection of response patterns failing on the long-string index revealed that 99% of these went back to respondents stating that they did not encounter any of the safety-related phenomena. We observed similar patterns for other counter-intuitively high C/ER proportions on this screen. Thus, a potential, speculative explanation for the counter-intuitively high C/IER proportion is that the samples of the affected scale-by-group combinations comprised different groups of attentive respondents: Attentive respondents who were exposed to some aspects of an unsafe school environment (thus providing response vectors comprising both response options) took longer to generate their responses. Attentive respondents disagreeing with the presented statements required a shorter amount of time to do so, and were falsely labeled as careless.

To further illustrate this, Fig. 6 shows the observed distribution of log geometric mean time per item on screen ST266 for the group with the highest C/IER proportion of .86. Distributions of the presumed C/IER (orange) and attentive (blue) classes implied by the mixture decomposition are superimposed. Further, Fig. 6 provides proportions of respondents failing the long-string index hurdle for three time bins, marked with dashed vertical lines. As can be seen, observed times exhibited a clear bimodal shape, with the faster mode dominating the distribution. Proportions of respondents failing the long-string hurdle were particularly high for times falling into the faster mode and rapidly decreased with an increasing time per item. The mixture decomposition correctly identified the bimodal shape, labeling the faster mode as C/

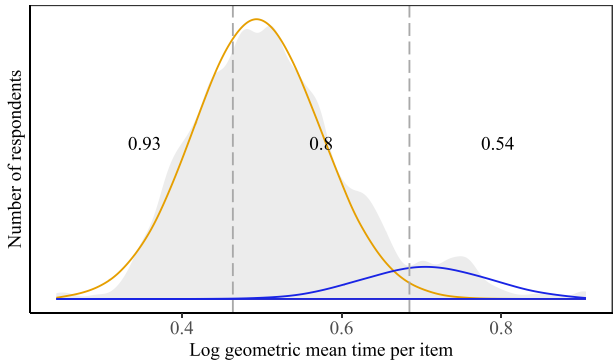


Fig. 6 Observed distribution of log geometric mean time per item on screen ST266 for the group with a multiple-source-indicator-implied C/IER proportion of .86. Distributions of the presumed C/IER (orange) and attentive (blue) classes implied by the mixture decomposition approach are superimposed. Proportions of respondents failing the long-string index hurdle are given for three time bins, marked with dashed vertical lines

IER. Note that the mean (corresponding to 1.64 s) of the presumed C/IER component was still relatively high compared to the means (corresponding, on average, to 0.98 s; interquartile range: [0.72; 1.33]) for all mixture-decomposition-implied C/IER components in booklet 1. This provides further indication that the faster component on the investigated scale-by-group combination may capture behavioral aspects different from C/IER.

Part II: Investigating associations with external correlates

Relating C/IER to self-reported effort

We observed small negative correlations between self-reported effort and average values on q. Effects were comparable in size across booklets. For booklet 1, the median within-group correlation was -.14 (interquartile range: [-.17; -.09]). For booklet 2, the median within-group correlation was -.13 (interquartile range: [-.17; -.09]). That is, in line with previous research using other behavioral C/IER indicators (Meade & Craig, 2012; Douglas et al., 2023), persons with higher C/IER levels as indicated by q tended to report lower effort on the questionnaire.

Relating C/IER to scale position

We observed no PSRF values above 1.05 and no ESS below 400 in the Bayesian hierarchical Beta regressions conducted to investigate the relationship between scale position and the occurrence of C/IER. As evidenced in Table 2, for both booklets, screen position was positively related to scale-by-group-level C/IER proportions. Inserting the regression coefficients for booklet 1 displayed in Table 2 into Eq. 6 yields expected average C/IER proportions of .05 and .10 for the first (44) and last (89) scale position considered. For booklet 2, the expected average C/IER proportions for the first (37) and last (82) scale position considered are .05 and .08, respectively. At the same time, in line with previous research (Ulitzsch et al., 2024), there was considerable variation in baseline C/IER proportions across groups, as indicated by large standard deviations of random intercepts.

Table 2 Results for the Bayesian hierarchical Beta regression

	Booklet 1		Booklet 2	
	Estimate	**95% CI**	**Estimate**	**95% CI**
Fixed effects				
Intercept	-3.85	[-4.03; -3.65]	-3.42	[-3.59; -3.23]
Screen position	0.02	[0.02; 0.02]	0.01	[0.01; 0.02]
Random effects				
Intercept (SD)	0.59	[0.50; 0.70]	0.52	[0.44; 0.62]
Precision parameter	30.88	[29.16; 32.62]	25.87	[24.42; 27.40]

•*Note:* SD: standard deviation; CI: credibility interval

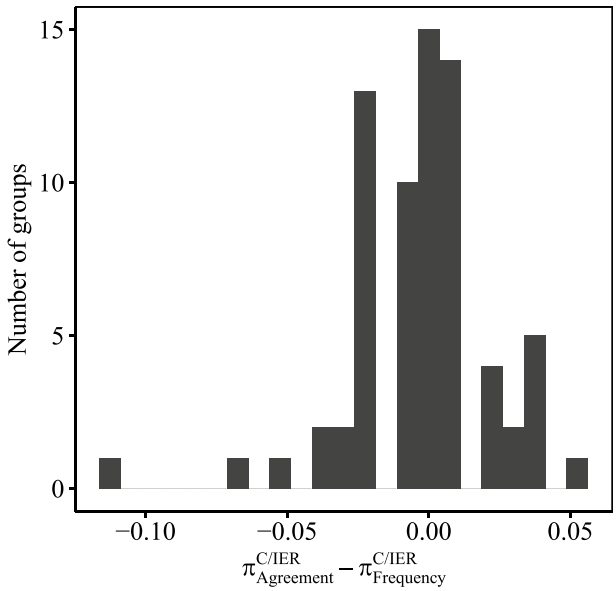


Fig. 7 Differences in group-level C/IER proportions across scales with agreement and frequency formats

Part III: Investigating the effect of scale format on C/IER

E1: Comparing scales with agreement and frequency formats

Overall, the multiple-source indicator q indicated no effect of scale format on the occurrence of C/IER; q implied C/IER proportions of .08 [.04; .11] for both agreement and frequency scales. In total, 25% of the scale-by-group pairs exhibited significant differences in C/IER proportions. Out of these, approximately half (45% of all scale-by-group pairs with significant differences; i.e., 11% of all scale-by-group pairs) indicated higher C/IER levels for the agreement format. That is, scale-by-group pairs with significant differences in C/IER proportions did not indicate a systematic advantage for either scale format.

This effect held true across groups. Figure 7 displays the distribution of differences in median group-level C/IER proportions between agreement and frequency scale formats averaged across the 10 experimentally manipulated scales. As can be seen, differences were distributed around zero and, by and large, there was only small variation in differences across groups (interquartile range: [−0.02; 0.01]). Nevertheless, for some few groups, we observed markedly different proportions of average presumed C/IER on

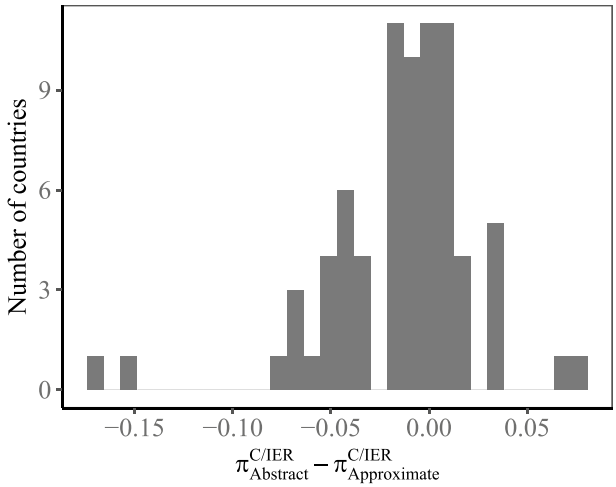


Fig. 8 Differences in group-level C/IER proportions across scales with abstract and approximate frequency labels

frequency as opposed to agreement scales (range of differences in median group-level C/IER proportions: [−0.11; 0.05]).

E2: Comparing abstract and approximate frequency labels

Likewise, q did not indicate marked differences in C/IER proportions between scales with abstract and approximate frequency labels. We observed scale-by-group-level C/IER proportions of .05 (interquartile range: [.03; .08]) for abstract frequency labels and of .06 [.03; .10] for approximate frequency labels. In total, 38% of the 450 scale-by-group pairs exhibited significant differences in C/IER proportions. Out of all pairs with significant differences, 65% (corresponding to 112 scale-by-group pairs) indicated higher C/IER levels for approximate frequency labels.

figure 8 displays the distribution of differences in median group-level C/IER proportions between abstract and approximate frequency labels averaged across the 6 experimentally manipulated scales. Differences were distributed around −0.01, and variation was somewhat larger than for our comparisons of agreement and frequency formats (interquartile range: [−0.03; 0.01]). Again, there were some few groups for which there were markedly pronounced differences in both directions (range: [−0.17; 0.08]).

Discussion

The present study introduced a multiple-source indicator for C/IER and subsequently applied this indicator to explore whether or not subtle changes in scale format have the potential to curb C/IER. To this end, we leveraged the rich data and opportunities provided by the PISA 2022 field trial background questionnaire.

The rationale of the presented indicator q is to provide more robust C/IER identification by requiring agreement in C/IER identification from indicators derived from both response patterns and timing data. To this end, we integrated Curran's (2016) multiple-hurdle approach with the screen time decomposition approach by Ulitzsch et al. (2024). The requirement of agreement in aberrance detection is rooted in a well-established

stream of psychometric literature combining item response detect aberrances in both low-stakes (e.g., rapid guessing, Ulitzsch et al., 2020; Wang & Xu, 2015; Nagy & Ulitzsch, 2021, 2022) and high-stakes testing (e.g., item compromise, van der Linden & Belov, 2023)[7] and is grounded in the conceptualization of C/IER as a behavior that requires markedly less time than attentive responding and results in aberrant response patterns.

In Part I of this study, we developed the multiple-source indicator q and provided intuition for its working. We illustrated (a) how q integrates information from multiple behavioral sources and (b) how doing so safeguards against the scale-dependency of single behavioral indicators' informative value and too liberal threshold settings in its multiple-hurdle component. Hence, considering information from multiple behavioral sources not only allows for C/IER detection that closely aligns with common conceptualization of C/IER behavior but also provides a means for preventing false C/IER labels that would have been given on the basis of single indicators.

Further, our investigations of counter-intuitively high C/IER proportions illustrated that q does not entirely hedge the risk of false positives. More specifically, there will be false positives whenever the mixture decomposition and at least one multiple-hurdle component provide the same false C/IER label.

In Part II, we investigated associations of the proposed indicator with commonly reported external correlates of C/IER and replicated (a) the well-documented position effect of C/IER and (b) agreement with self reports on effort expended on the questionnaire. In line with our expectations on C/IER trajectories across lengthy questionnaires and previous findings (Ulitzsch et al., 2022, 2024; Berry et al., 1992; Baer et al., 1997; Galesic & Bosnjak, 2009; Bowling et al., 2020), the multiple-source indicator signaled an increase of C/IER with increasing screen position. The small negative relationship between the proposed indicator and self-reported effort corroborates previous results regarding the relationship between behavioral and self-report measures of disengagement in cognitive assessments (Ulitzsch et al., 2021a; Wise & Kong, 2005). Nevertheless, although commonly investigated in C/IER research (e.g., Douglas et al., 2023; Meade & Craig, 2012), it may be questioned to which extent the relationship between self-reported effort and C/IER identified based on behavioral information is informative. The validity of this correlation rests on the assumption that careless respondents provided valid self reports on effort, which is plausible to be violated. Our results exhibited robustness and effects were comparable in direction and size across both investigated booklets.

In Part III, we employed the developed indicator to study the effects of scale format on C/IER occurrence, leveraging data from a large-scale survey experiment implemented in the PISA 2022 field trial. Here, two aspects of scale format were experimentally varied—(a) frequency instead of agreement formats and (b) frequency scales with approximate instead of abstract labels. We found negligible effects for both scale format manipulations. Therefore, for now, we recommend that researchers' decision on scale format should predominantly be guided by substantive consideration—i.e., by considerations on

[7] If researchers consider this requirement too strong for the application at hand, they can combine response pattern and timing information in a sequential procedure as suggested by Curran (2016).

which format is most suitable to capture the targeted construct—because overall, the effects of this decision on C/IER occurrence are negligible.

Recommendations for using the multiple-source indicator

We see multiple use cases for the employed multiple-source indicator q. First and foremost, q can be employed to further investigate antecedents and correlates of C/IER and inform a better understanding of C/IER occurrence. Researchers may relate q to person and scale characteristics to identify potential person- and scale-level drivers of C/IER contamination (as done using other indicators in Bowling et al., 2016; Bowling et al., 2020) or, as in the present study, analyze survey experiments to evaluate the utility of changes in survey characteristics for curbing C/IER. Recent examples for such survey experiments comprise both experiments evaluating interventions targeted to reduce C/IER (e.g., promising rewards for attentive behavior or issuing warnings that C/IER will be punished as in Gibson & Bowling, 2019) and studies aiming to investigate potential unintended consequences of changes in survey designs (Sischka et al., 2022).

Second, and equally important, as its $\pi^{\text{C/IER}}$ component, q can be used to construct person weights $1 - q_{is}$ that downweigh response patterns according to their (presumed) probability of stemming from C/IER. As a sensitivity check, researchers could complement their analyses with an attentiveness-weighted counterpart and investigate the robustness of results when presumed C/IER response patterns are given less weight (see Ulitzsch et al., 2024, for an example). In simulations and real-data analysis, Ulitzsch et al. (2023) illustrated the potential consequences of weighting response patterns according to their (presumed) probability of stemming from C/IER in addressing substantive research questions. They showed that this procedure may result in an adjustment of both point estimates and standard errors, both due to a reduction of the effective sample and taking C/IER identification uncertainty into account. Such comparisons may support researchers in deciding whether or not their results are likely to be distorted by C/IER. Third, q may be added as a person-by-scale level quality indicator in public use files of (large-scale) surveys. Researchers may then investigate the distribution of q to gauge data quality of the provided data set as well as for different sub-groups. For instance, when researchers want to conduct analyses by groups differing in, say, socio-economic status, they may first investigate whether the groups can be assumed to provide data of comparable quality.

We point out that although we developed and evaluated the proposed indicator in the context of background questionnaires of large-scale educational studies, its usage is not restricted to these settings. In fact, we see its major use case as a tool for gauging data quality and probing robustness of conclusions in the context of online surveys that, in contrast to large-scale educational studies, have markedly less rigorous quality control and commonly lack proctoring, and may, therefore, have a much higher risk to be contaminated with C/IER responses. For instance, using a combination of different indicators, Douglas et al. (2023) found that the quality of data collected through commonly used platforms for online data collection may be alarmingly low, with rates of respondents providing high-quality data ranging from 68% to as low as 26%, depending on the platform used for data collection. Hence, as the employment of online surveys for educational and psychological research is rising, so does the need for adequate tools to

evaluate the quality of the data obtained and probe the robustness of conclusions against potential C/IER contamination.

Note that as the multiple-hurdle approach by Curran (2016), the proposed indicator is sensitive to the response-pattern-based indicators employed to construct d^{MH}. Consider the case where d^{MH} evaluates to zero, i.e., none of the employed response-pattern-based indicators signals C/IER. In this case, q will also evaluate to zero, regardless of whether or not a respondent showed suspiciously short screen times. When the considered indicator-threshold combinations indeed capture all relevant peculiarities in the response patterns and short screen times occurred for a reason other than C/IER (e.g., because someone is a very fast reader), this property hedges against wrong conclusions on C/IER occurrence. However, if the short screen time indeed goes back to C/IER, but the employed indicator-threshold combinations used for d^{MH} simply do not detect the resultant aberrant response pattern, requiring agreement between response patterns and timing data may be overly cautious. For now, we therefore recommend employing many different response-pattern-based indicators that are sensitive to different aspects of C/IER for d^{MH} and opt for somewhat more liberal threshold settings. As illustrated in Part I of this study, q exhibits some degree of robustness against too liberal threshold settings in the construction of d^{MH}. The validity of this recommendation should, however, be scrutinized in future research.

Even though we believe that the proposed indicator has the potential to provide more robust C/IER identification, we also acknowledge that there are still multiple subjective, somewhat arbitrary decisions involved in the construction of q. We, therefore, strongly recommend probing conclusions on C/IER for sensitivities to these decisions, e.g., through specification curve analysis (Simonsohn et al., 2020) varying the involved indicators and thresholds.

Limitations and future directions

We point out that the proposed indicator does not provide error-free C/IER identification. Due to the complexity of response behavior, it is questionable whether a perfect C/IER indicator can ever be developed. Nevertheless, future research may refine the proposed indicator (or its constituting components) to further improve its C/IER identification precision. A potential starting point may be further in-depth investigations of potential sources of error. Subsequently, researchers may incorporate decision rules with the indicator that aim to avoid typical decision errors of the current version of the multiple-source indicator.

In our illustrations, we could identify examples where the proposed indicator prevented presumably erroneous conclusions on C/IER that would have been made based on standalone indicators. Nevertheless, from these illustrations it cannot be concluded that the proposed indicator generally provides a more valid measure of C/IER than its constituting components. To support such conclusions, experimental evidence is needed. Only experimental manipulation of C/IER occurrence can provide researchers with "ground truth" as to whether respondents were attentive or engaged in C/IER, allowing to evaluate the sensitivity and specificity of different detection techniques. Prior research has implemented experimental C/IER manipulations and data have been made publicly available (e.g. Schroeders et al., 2020; Pokropek

et al., 2023). Typically, however, C/IER is evoked by the instruction to respond as fast as possible (as in Schroeders et al., 2020) or to behave inattentively (as in Pokropek et al., 2023; Niessen et al., 2016). It can be questioned whether participants receiving these instructions exhibit behavior resembling real-life C/IER, threatening the validity of these data sets. We, therefore, belief that the field would strongly profit from the development of ecologically valid C/IER manipulations. For instance, manipulations could be targeted at mechanisms underlying C/IER such as fatigue or lack of interest rather than instructing participants to exhibit some behavior.

We did not find that using frequency instead of agreement formats or approximate instead of abstract frequency labels impacts C/IER occurrence. From these results, however, it cannot be concluded that C/IER occurrence cannot be curbed by careful scale design, and the effect of a broad array of scale characteristics still remains to be investigated. The investigated changes in scale format were relatively subtle and only concerned wording, while the overall mode of responding set by the Likert-type scales was left unchanged. Future research may investigate more severe changes, such as changes in the number of response options or even completely different modes of responding, e.g., the use of forced choice or recently proposed drag-and-drop formats (see Böckenholt, 2017; Henninger et al., 2022, for further evaluations) instead of Likert-type scales. Further, for some few groups, there occurred pronounced differences in both directions. One explanation may be culture- or language-specific effects of scale formats on C/IER occurrence. The investigation of potential culture- or language-specific effects remains an interesting topic for future research.

Finally, the minor effects of the investigated scale characteristics on C/IER do not imply that the investigated scale formats provide data of comparable quality. C/IER is by far not the only threat to data quality, and it may well be that other aspects (e.g., construct validity or the occurrence of other biases such as response styles) are impacted by the investigated changes in scale characteristics.

Acknowledgements
Not applicable.

Author Contributions
Esther Ulitzsch: Conceptualization; methodology; formal analysis; visualization; writing - original draft. Janine Buchholz: Conceptualization; methodology; data curation; writing - review & editing. Hyo-Jeong Shin: Conceptualization; methodology; writing - review & editing. Jonas Bertling: Conceptualization; methodology; writing - review & editing. Oliver Lüdtke: Conceptualization; methodology; writing - review & editing.

Funding
This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 33160, and the Organisation for Economic Co-operation and Development (OECD).

Availibility of data and materials
The data that support the findings of this study are available from the Organisation for Economic Co-operation and Development (OECD) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the OECD upon reasonable request.

Declarations

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 August 2023 Accepted: 6 May 2024
Published online: 10 June 2024

References

Arias, V. B., Garrido, L., Jenaro, C., Martinez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*, 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

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

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. (2017). Measuring response styles in likert items. *Psychological Methods, 22*(1), 69–83. https://doi.org/10.1037/met0000106

Boe, E.E., May, H., & Boruch, R.F. (2002). Student task persistence in the third international mathematics and science study: A major source of acheievement differences at the national, classroom, and student levels. (tech. rep. No., CRESP-RR-2002-TIMSS1). Pennsylvania Univ., Philadelphia. Center for Research and Evaluation in Social Policy.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods, 24*(4), 718–738. https://doi.org/10.1177/1094428120947794

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.

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2021). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods, 26*(2), 323–352. https://doi.org/10.1177/10944281211056520

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1). https://doi.org/10.18637/jss.v076.i01

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

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality, 78*, 106–124. https://doi.org/10.1016/j.jrp.2018.10.008

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity, 52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

Deng, S., & Bolt, D.M. (2016). Rating scale format and item sensitivity to response style in large-scale assessments. In L. Van der Ark, L. Wiberg, S. Culpepper, J. Douglas, & W. Wang (Eds.), Quantitative psychology: The 81st annual meeting of the psychometric society (pp., 347–356).

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

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, Cloud Research, Qualtrics, and SONA. *Plos ONE, 18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment, 29*(2), 136–151. https://doi.org/10.1177/1073191120957102

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

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly, 73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

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). Chapman Hall.

Gibson, A. M., & Bowling, N. A. (2019). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment, 36*(2). https://doi.org/10.1027/1015-5759/a000526

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research (tech. rep.). *National Institute on Drug Abuse*. https://doi.org/10.1037/e495862006-003

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

Hamari, J., Malik, A., Koski, J., & Johri, A. (2019). Uses and gratifications of Pokémon Go: Why do people play mobile location-based augmented reality games? *International Journal of Human-Computer Interaction, 35*(9), 804–819.

Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 2): Applications and novel extensions. *Psychological Methods, 25*(5), 577–595.

Henninger, M., Plieninger, H., & Meiser, T. (2022). The effect of response formats on response style strength: An experimental comparison. *European Journal of Psychological Assessment*. https://doi.org/10.31234/osf.io/5jxg7

Hoff, P.D. (2009). *A first course in Bayesian statistical methods*. Berlin: Springer.

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/s10869-011-9231-8

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

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296–309. https://doi.org/10.1177/0022022189203004

Jackson, D. (1976). The appraisal of personal reliability (tech. rep.) (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
Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*(3), 512–541. https://doi.org/10.1177/1094428115571894
Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style? *Quality and Quantity, 47*(1), 193–211. https://doi.org/10.1007/s11135-011-9511-4
Kroehne, U., Buchholz, J., & Goldhammer, F. (2019). Detecting carelessly invalid responses in item sets using item-level response times (tech. rep.) (Paper presented at the Annual Meeting of the National Council on Measurement in Education). Toronto, Canada.
Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236.
Landers, R. N., Bauer, K. N., & Callan, R. C. (2017). Gamification of task performance with leaderboards: A goal setting experiment. *Computers in Human Behavior, 71*, 508–515. https://doi.org/10.1016/j.chb.2015.08.008
Magraw-Mickelson, Z., Wang, H., & Gollwitzer, M. (2020). Survey mode and data quality: Careless responding across three modes in cross-cultural contexts. *International Journal of Testing, 22*(2), 121–53.
Marshall, A.D. (2019). Caring more about careless responding: Applying the theory of planned behavior to reduce careless responding on online surveys [Doctoral dissertation, Colorado State University].
McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior, 84*, 295–303. https://doi.org/10.1016/j.chb.2018.03.007
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
Mitchell, M. S., Vogel, R. M., & Folger, R. (2015). Third parties' reactions to the abusive supervision of coworkers. *Journal of Applied Psychology, 100*(4), 1040–1055. https://doi.org/10.1037/apl0000002
Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology, 44*(1), 369–399. https://doi.org/10.1177/0081175013516114
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, 82*(5), 845–879. https://doi.org/10.1177/00131644211045351
Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning, 39(3)*, 751-766. https://doi.org/10.1111/jcal.12719
Nichols, A. L., & Edlund, J. E. (2020). Why don't we care more about carelessness? understanding the causes and consequences of careless participants. *International Journal of Social Research Methodology, 23*(6), 625–638. https://doi.org/10.1080/13645579.2020.1719618
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
Pokropek, A., Zólstroktak, T., & Muszyński, M. (2023). Mouse chase: Detecting careless and unmotivated responders using cursor movements in web-based surveys. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000758
Robie, C., Meade, A. W., Risavy, S. D., & Rasheed, S. (2022). Effects of response option order on like rt-type psychometric properties and reactions. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211069406
Sischka, P. E., Décieux, J. P., Mergener, A., Neufang, K. M., & Schmidt, A. F. (2022). The impact of forced answering and reactance on answering behavior in online surveys. *Social Science Computer Review*, *40*(2), 405–425. https://doi.org/10.1177/0894439320907067.
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
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
Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human. *Behaviour, 4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z
Smith, B. M., Twohy, A. J., & Smith, G. S. (2020). Psychological inflexibility and intolerance of uncertainty moderate the relationship between social isolation and mental health outcomes during COVID-19. *Journal of Contextual Behavioral Science, 18*, 162–174.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
Ulitzsch, E., Domingue, B. W., Kapoor, R., Kanopka, K., & Rios, J. (2023). *A probabilistic filtering approach to non-effortful responding*. Educational Measurement: Issues and Practice. https://doi.org/10.1111/emip.12567
Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (2021). Model meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment, 26*(2), 104–124. https://doi.org/10.1080/10627197.2020.1858786
Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2021). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika, 87*, 593–619. https://doi.org/10.1007/s11336-021-09817-7
Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2023). Using response times for joint modeling of careless responding and attentive response styles. *Journal of Educational and Behavioral Statistics, 49*(2), 173–206. https://doi.org/10.3102/10769986231173607
Ulitzsch, E., Shin, H.-J., & Lüdtke, O. (2024). Accounting for careless and insufficient effort responding in large-scale survey data–Development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods, 56*, 804–825. https://doi.org/10.3758/s13428-022-02053-6

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, 73*(1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in survey data. *British Journal of Mathematical and Statistical Psychology, 75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

van der Linden, W. J., & Belov, D. I. (2023). A statistical test for the detection of item compromise combining responses and response times. *Journal of Educational Measurement, 60*(2), 235–254. https://doi.org/10.1111/jedm.12346

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement, 59*(4), 470–501. https://doi.org/10.1111/jedm.12317

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

Ward, M., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology, 67*(2), 231–263. https://doi.org/10.1111/apps.12118

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

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

Publisher's Note