(Nichols et al., 1989), reads and adheres to instructions (Johnson, 2005), and follows cognitive steps necessary to answer a survey question correctly. Four such steps were proposed by Tourangeau, Rips and Rasinski (2000): a) question comprehension, b) retrieving relevant information, c) integrating this information into a judgement, d) mapping this judgement onto a provided response scale. Skipping one of these steps or performing it superficially may result in nonoptimal survey responding - satisficing (Krosnick, 1991). The reasons for satisficing are multiple: lack of or inadequate cognitive abilities, insufficient attention due to low interest, fatigue, confusion, or distraction, externally motivated (e.g. money-oriented) goal to participate that facilitates shortcutting the survey (Maniaci & Rogge, 2014). The task difficulty also contributes to satisficing with factors such as linguistic complexity and survey length being most important and with participants' literacy or educational level playing a role of potential moderator here (Narayan & Krosnick, 1996). According to the Krosnick's theory, main behavioural mechanisms of satisficing are limited information search, in which participants shortcut the retrieval search and provide first response that comes to their minds, and reducing response categories to a smaller set, thus reducing cognitive effort in the mapping stage (Krosnick, 1991). Straightlining, understood as responding with the use of only one option, is an extreme form of this behaviour (Kim et al., 2019). Satisficing, in all its forms, is a major threat to reliability and validity of the survey data.

Careless/insufficient effort responding and its consequences

Satisficing in survey/self-report research often takes the form of careless, insufficient effort or inattentive responding (Huang et al., 2012; Meade & Craig, 2012). The term careless/insufficient effort responding (C/IER) is most often used to describe the process, and the term inattentive participants to denote the persons engaging in it. Various patterns of C/IER are identified, e.g.: endorsing just one or two response options in the whole survey (straightlining, non-differentiation), pattern responding (responses form a pattern on the survey page, but are not logically connected), pseudorandom/inconsistent responding, item non-responding, speeding (completing the survey in a very short time that precludes full understanding of the items and instructions), skipping instructions or entire parts of survey (Gummer et al., 2021; Huang et al., 2012; Meade & Craig, 2012; Nichols et al., 1989; Ulitzsch et al., 2022a, 2022b).

The proportion of inattentive respondents differs widely from study to study, depending on task characteristics and methods used to identify them (Credé, 2010; Meade & Craig, 2012; Kung et al., 2018; van Laar & Braeken, 2022). The estimates vary from just 1-3% to well over 60%, with the most typical value between 5 and

20% (Barends & de Vries, 2019; Brühlmann et al., 2020; DeSimone & Harms, 2018; Mancosu et al., 2019; Maniaci & Rogge, 2014; Meade & Craig, 2012). As evidenced by in-depth analyses, as little as 5% of the inattentive participants, if left uncorrected, is enough to distort studies' inferential results, e.g. decrease effect sizes (Maniaci & Rogge, 2014) or lead to deterioration of model fit and distortion of factorial structure (Arias et al., 2020).

Consequences of C/IER probably depend on the proportion of the sample that was inattentive and the degree and the type of inattentiveness, but the systematic evidence here is missing (DeSimone et al., 2018). It was evidenced that uncorrected C/IER in the data can change associations among constructs of the interest, **impacting correlation patterns or regression analyses** (DeSimone et al., 2018; Maniaci & Rogge, 2014), **size of experiment effects** (Alvarez et al., 2019), the **internal structure of measures** (Woods, 2006), deteriorating the model fit (Arias et al., 2020; Voss, 2023) or changing factor loadings (Arias et al., 2020; Steedle et al., 2019), can affect **reliability estimates** (Carden et al., 2018) and may lead to **over/underestimation** of a measured trait (Jones et al., 2015).

For example, Arias et al. (2020) showed that even a small proportion of careless responders (around 5% of the overall sample) is enough to drastically deteriorate factor model fit, to such an extent that fit statistics would recommend model rejection. However, when only a "clean" sample was analysed (inattentive participants screened out), model fit improved drastically, indicating a very good fit. Inattentive participants also had a negative impact on factor loadings, but only for negatively worded items. The value of such factor loadings was reduced by 20-30% when careless respondents were not filtered out. Removal of inattentive participants also led to the elimination of cross-loadings and the reduction of measurement error.

Methods to account for careless or insufficient effort responding

Researchers' preoccupation with participants' (in)attentiveness is followed by a growing body of research aimed at preventing or amending C/IER in web surveys. Despite the constant flow of emergent evidence, the struggle to forge efficient and widely accepted methods to account for C/IER is far from accomplishment (Baumgartner & Weijters, 2022; Bowling et al., 2016; Ward & Meade, 2023).

Two main approaches to account for C/IER can be distinguished; first, to prevent such behaviour, and second, to react to its presence after data are collected. Manipulated instructions, e.g., using warnings, threats, or pleas to respond attentively, are an example of preventive methods that can induce attentive responding (Huang et al., 2012; Pokropek et al., 2023). However, research to date signals their limited effectiveness (e.g. Toich et al., 2022). Reactive measures

can be divided into indicator-based and model-based approaches (Ulitzsch et al., 2022a, 2022b), with the former being much more researched (cf. Meade & Craig, 2012).

Indicator-based approaches rely on creating indices that are aimed to identify certain types of inattentive responding and then screen out the sample based on preset cutoffs of the indicators values. Such indices can be divided into three main groups: a) non-differentiation indices that aim to capture respondents embracing only a limited set of response options (where various types of straightlining and intra-individual variability indices belong), b) inconsistent/random/aberrant responding indices that are to identify outliers (Mahalanobis distance and similar measures belong to this group), and c) speeding indices, based on response times and time spent on survey, where unrealistically short or long times are flagged as potentially indicating C/IER (Kroehne et al., 2019; Ulitzsch et al., 2022a, 2022b). The emerging indices based on other log-data such as clickstream (map of all actions and their order in a survey) or cursor/mouse moves complement these groups (Horwitz et al., 2017; Pokropek et al., 2023). A separate group consists of self-report measures, in which participants can identify themselves as inattentive respondents or report a level of their diligence (Meade & Craig, 2012).

All of these methods have serious drawbacks (Baumgartner & Weijters, 2022; Ward & Meade, 2023). Self-report measures do not possess sufficient sensitivity and thus result in too many inattentive participants mistakenly classified as attentive (these errors are called omissions or false negatives). The non-differentiation and inconsistency indices suffer from a reversed problem – the propensity to flag attentive participants as inattentive (the error of false alarms/false positives). Moreover, interpreting the values yielded by these indices is often challenging with the setting of valid cut-off values being somewhat arbitrary (Baumgartner & Weijters, 2022). Additionally, the values of these indicators depend on various unique characteristics of self-report scales used in a given study, making comparisons across various studies difficult, if not impossible (Bowling et al., 2016; DeSimone & Harms, 2018). While log-data indices are still scarcely validated; it already seems that they may suffer from similar limitations as other indices, such as arbitrariness in setting cut-off criteria and dependency on scale features (Pokropek et al., 2023).

Model-based approaches assume that different data-generating processes are responsible for responses associated with an attentive and inattentive responding (Ulitzsch et al., 2023). These distinct response behaviours are represented by separate models. Based on their observed response vectors (item response patterns) participants are then classified to one of the groups (van Laar & Braeken, 2022). There are two main advantages of model-based approaches over indicator-based ones. First, there is no need for arbitrary cut-offs, as the classification is done

internally in the model and the goodness of categorisation can be evaluated by the objective criteria. Second, the invalid responding is explicitly defined on the basis of theoretical concepts. For example, van Laar & Braeken (2022) defined C/IER as random responding modelled by equal probability of choosing response options, whereas Ulitzsch et al. (2022a, 2022b) defined C/IER as speeding. Supervised machine learning and latent class analysis can be also classified as certain types of model-based approaches, although no exact model is defined in these methods (Schroeders et al., 2022).

Model-based approaches are still in their infancy, as only a few studies proposing such remedies exist (Arias et al., 2020; Mansolf & Reise, 2018; Schroeders, et al., 2022; Ulitzsch et al., 2022a, 2022b). Their initial results are very encouraging and promise to soon become the main method to account for C/IER in self-reports. However, as to date, they are scarcely researched, insufficiently validated, too complicated and too demanding with regard to quantitative skills to be a mainstream method for non-methodological research.

In this context, attention checks are measures that can jointly offer ease-of-use of indicator-based approaches and higher objectivity of model-based solutions. Moreover, they are also believed to have preventive potential, when spotting attention checks, participants should increase their attention (Maniaci & Rogge, 2014; Shamon & Berning, 2020). The key words here are "can" and "believed", though - attention checks are simply not researched enough to pose more outright claims (Gummer et al., 2021).

2. ATTENTION CHECKS – TYPES AND DEFINITIONS

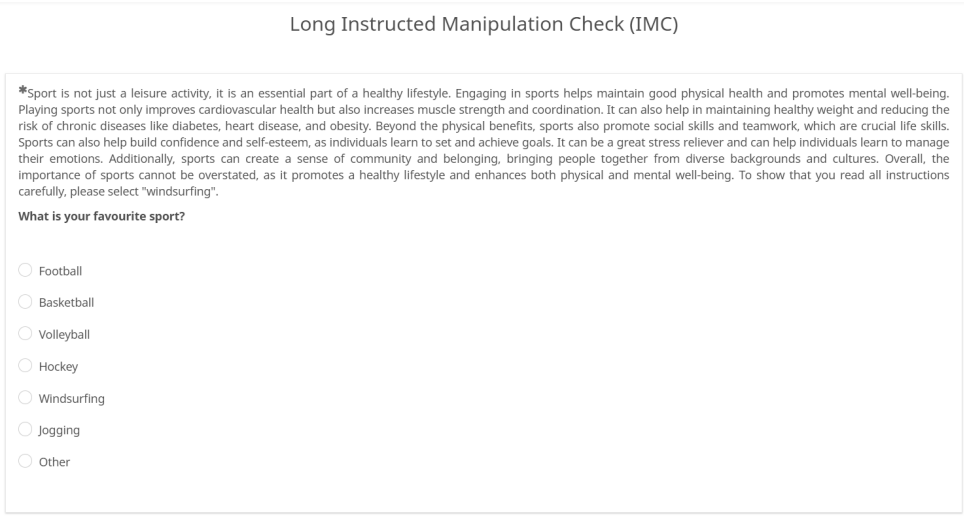
One of the most widely used approaches to solve the C/IER-problem are attention checks (Oppenheimer et al., 2009). This is a general term for a diversified family of instruments similar in one regard; they are based on adding to a survey an additional task or item aimed to identify inattentive respondents in an objective way. "Attention checks" is an umbrella term for instruments used to identify inattentive participants in various data collection occasions (tests, assessments, surveys, etc.). As defined by Silber et al. (2022), they "aim to capture whether respondents thoroughly read, comprehend, and answer questions". Synonymic names are "trap questions", "detection items", "red herring questions", "screeners", and others (Alvarez et al., 2019; Arthur et al., 2021). Here, the term "attention checks" is used consistently throughout the text. An example of such checks are items like: "If you read it carefully, please select «Strongly agree» in this question", "Orange is fruit" or "I have been to the moon" (Huang et al., 2012; Maniaci & Rogge, 2014). Obviously, only one type of response is logically valid for these items - if a respondent fails to give it, e.g., endorses any other answer than "Strongly

agree", disagrees that orange is a fruit or claims to visit the moon, it is a clear and objective sign of inattentiveness (Abbey & Meloy, 2017). Or at least researchers want to believe that attention checks of this kind are valid and objective signs of C/IER enabling to screen out inattentive participants from the analysed sample - as described in detail in the following sections of the article - attention checks are often misinterpreted by respondents (Curran & Hauser, 2019; Silber et al., 2022) and fail to validly identify inattentive participants (e.g. Gummer et al., 2021).

Many types of attention checks have been proposed until now. The most researched one is probably the **instructional manipulation check** (IMC, Oppenheimer et al., 2009). This attention check aims to assess whether participants have read the instructions and acted accordingly. The idea behind this check is to ask participants to do something less conventional than usual to prove that they have read and comprehended instructions - e.g., to skip a question, select a given response or click on the indicated element of the assessment screen. The instructional manipulation check can be stylised as an instruction screen or a regular screen with questions/tasks but with manipulated instructions added. Typically, there is a distinction between long (see Figure 1) and short IMCs (Figure 2), the former being related to long stems/instructions (at least a few longer sentences or one whole paragraph of text), the latter consisting of just one or two sentences with instructions (Morren & Paas, 2020).

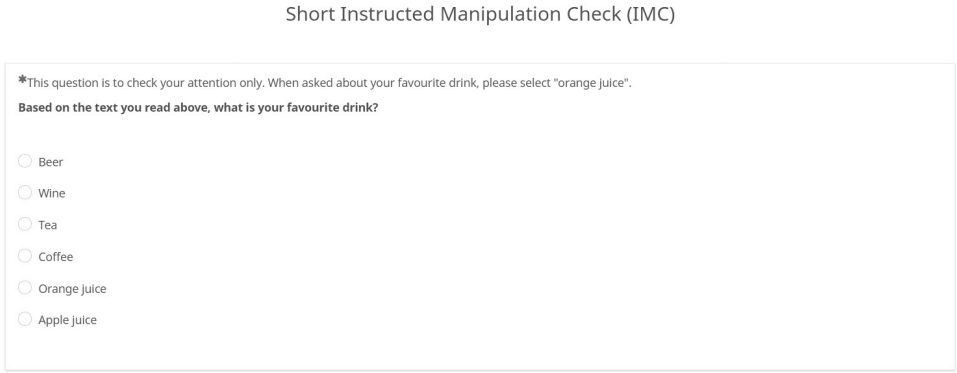
Previous research brought many caveats **against the use of instructional manipulation checks (IMCs)** to filter respondents' attentiveness in web surveys (Ladini, 2022; Liu & Wronski, 2018; Maniaci & Rogge, 2014; Morren & Paas, 2020). It seems they produce an unreasonably high proportion of participants flagged as inattentive due to a **large number of false positives** (false alarms) - attentive participants mistakenly classified as inattentive. Participants failing instructional manipulation checks resemble more attentive participants than participants failing other attention checks, pointing to limited validity of screens based on this method (Alvarez et al., 2019; Anduiza & Galais, 2017; Babakhani et al., 2022; Berinsky et al., 2014; Morren & Paas, 2020). Moreover, long IMCs seem to be correlated with participants' traits other than attentiveness, e.g. reading ability or memory recall. They can also be deemed as tricking participants into wrong answers, as they often require atypical actions that can even be seen as unfair or breaking human behaviour norms.

Figure 1. Long Instructed Manipulation Check



Note: Text was generated by the author using ChatGPT free version.

Figure 2. Short Instructed Manipulation Check



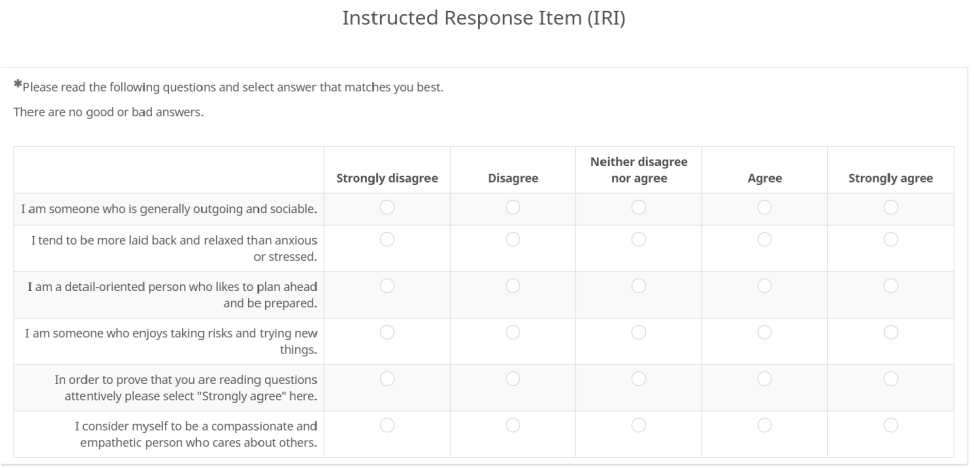
Note: Text was generated by the author.

A similar idea as to IMCs stands behind **instructed response items** (IRIs), some papers even classify them as short IMCs (e.g. Morren & Paas, 2020). The main differentiation between the two types is that IRIs are embedded into a list of other items of the same format, whereas short IMCs are most typically standalone

questions (Gummer et al., 2021). IRIs instruct respondents to take a specific action, e.g. mark a given response category or skip the item (see Figure 3; Liu & Wronski, 2018).

Instructed response items yield mixed results - there is an ample evidence to confirm they successfully identify inattentive participants (e.g. Alvarez et al., 2019; Gummer et al., 2021; Meade & Craig, 2012), but they are also related to significant problems. Most important of them are **misinterpretations** of such items and overt **noncompliance** with instructions (Curran & Hauser, 2019; Silber et al., 2022). Providing adequate explanations for using such attention checks seems to lower participants' negative opinions (Shamon & Berning, 2020; Silber et al., 2022), but exact effects of employing such rather unconcealed attention checks are not sufficiently researched (Gummer et al., 2021). Moreover, participants' attitudes towards instructed response items in particular and attention checks in general, are still elusive (Silber et al., 2022).

Figure 3. Instructed Response Item



Note: Exemplary items were generated by the author using ChatGPT free version.

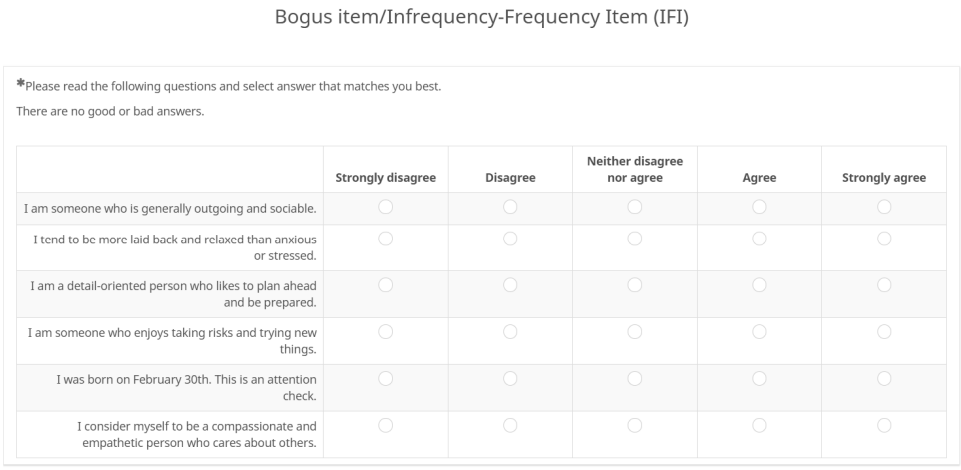
Bogus items or **infrequency/frequency items** (IFIs) are attention checks similar to IRIs, but with a slightly different idea - here, the items do not instruct the respondent what to do, but are framed in such a way that any attentive respondent should endorse a predicted response (Beach, 1989; Bowling et al., 2016; Huang et al., 2015; Kay & Saucier, 2023). For example, the item *"I have never brushed my teeth"* should elicit only disagreeing responses, whereas *"I have used a computer*

before" should be endorsed by the entire sample (especially in case of online survey). Several subtypes of IFIs exist. Most notable are absurd ("*I eat cement occasionally*") or impossible ("*I was born on February 30th*") statements, items conveying obvious truths ("*Water is wet*"), and more subtle infrequency/frequency items ("I have never cried").

Some types of the bogus items also met heavy critique based on empirical evidence: absurd and unclear statements (e.g. "*All my friends are aliens*", *"I can teleport across time and space")* produce large rates of false positive and also false negative rates, while **less conspicuous infrequency/frequency items yield much higher accuracy** (Curran & Hauser, 2019). Evidence points out that participants often try to make sense of absurd bogus items and interpret them non-literally (e.g. *"We are all aliens", "I can teleport in my dreams"*). Same caveat applies to less absurd, but impossible items (e.g. *"I work 14 months a year", "I can eat as much as a horse"*), as some respondents treat them more as a metaphor *("I work really a lot", "I can eat a lot of food"*) than a literal, straight-up sentence. It is also important to avoid double-barreled items, in which two parts can be judged separately (e.g. *"I am paid biweekly by leprechauns", "I am paid biweekly, just not by leprechauns"*). Finally, also bogus items containing non-existent words (e.g. "parabanjology") or any abbreviations/acronyms should be avoided as these terms can potentially confuse participants or cause item misinterpretations, e.g. due to wrong interpretation of the presented abbreviation/acronym (Curran & Hauser, 2019). The research shows that there is no other way than to pilot bogus/IFIs before their use in a survey to check their quality (Kay & Saucier, 2023).

See Figure 4 for a classic example. Please note that the addition of "This is an attention check" is not a standard procedure - some studies inform participants about the use of attention checks, some do not. There is evidence suggesting that such informing statements reduce noncompliance and other negative reactions (Silber et al., 2022). It seems that providing participants a reason to include attention checks in a survey is relevant for the majority of participants. Some respondents may use these explanations to scan surveys for key expressions pointing to the use of attention checks; other studies showed that indeed more experienced participants failed attention checks less often than inexperienced (Gummer et al., 2021; Peer et al., 2014). It is not known, however, whether this is due to higher attentiveness of more experienced participants or due to adaptation to the use of attention checks (Hauser & Schwarz, 2016).

Figure 4. Bogus item/Infrequency-Frequency Item (IFI)



Note: Exemplary items were generated by the author using ChatGPT free version.

Two other types of attention checks, which have been poorly studied, but quite often used in the applied research, are **maths quiz** and **captcha-like items** (Abbey & Meloy, 2017; Liu & Wronski, 2018; Lundmark et al., 2023). The former combines a short manipulated instruction and an easy mathematical expression to solve, e.g. *"To show that you read attentively, please solve 3 + 2 = ?"*. The latter consists of captcha-like tasks like selecting pictures or typing text into a box. Another idea for maths-based attention checks was proposed by Saad (2021). In this approach, participants are required solve a basic mathematical equation presented in words and provide their answer verbally as well (*"Please solve the following problem. Enter your answer in words. What is four divided by two?"*). Saad (2021) argues that this type of check has the capability not only to identify inattentive respondents but also to detect bots and participants with insufficient linguistic proficiency. An interesting hybrid of IRI and maths quiz was proposed by Peer et al. (2022) who constructed a following task:

> *"participants were presented with a paragraph of instructions followed by two 7-point-scale questions, and were asked to answer "2" on the first question, add 3 to that number, and use that value for the second question (any different response for either question indicated a failure to pass this attention check" with two survey items that asked participants "to answer "2" on the first question, add 3 to that number, and use that value for the second question"*. (citation from Peer et al. (2022)

Despite their face validity, such tasks remain scarcely studied and their validity in identifying inattentive participants is currently unknown (Liu & Wronski, 2018). It is also worthy to notice that these new tasks can be correlated with a number of respondents' traits (e.g. maths abilities, maths anxiety, working memory capacity, etc.), thus lowering the validity of such checks.

Figure 5. Math Quiz/Numerical Task

Math Quiz (Numerical Task)

*Please read the following questions and select answer that matches you best.

There are no good or bad answers.

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly agree
I am someone who is generally outgoing and sociable.	○	○	○	○	○
I tend to be more laid back and relaxed than anxious or stressed.	○	○	○	○	○
I am a detail-oriented person who likes to plan ahead and be prepared.	○	○	○	○	○
I am someone who enjoys taking risks and trying new things.	○	○	○	○	○
Is 1 + 2 = 3? This is an attention check.	○	○	○	○	○
I consider myself to be a compassionate and empathetic person who cares about others.	○	○	○	○	○

Note: Exemplary items were generated by the author using ChatGPT free version. Math Quiz/Numerical Task attention checks are also often presented as standalone survey items.

Other tasks used to assess participants' attention, e.g. counting pronouns in a text (Maniaci & Rogge, 2014) or evaluating quality of open-ended answers to a question (e.g. "*What is your favourite city to visit?*"), have been used as attention checks, but knowledge of their validity is limited. Moreover, such tasks can consume a lot of survey time and present little resemblance to other survey items, which can cause problems with participants' motivation and compliance. Also, some of their types lack one, objective answer that is correct. Hence these items should be classified as attention tasks (for a review, see Göritz et al., 2021), not attention checks.

A somewhat distinct type of attention checks are manipulation checks and comprehension checks, but as they are employed to test whether participants paid attention to experimental manipulation or whether the manipulation yielded predicted results, they serve different purposes than other attention checks and are out of the scope of this proposal (see Freitas-Lemos et al., 2022; Hauser et al., 2018; Prolific Team, 2023).

All attention checks reside on having only one, objectively correct answer and differ in the level of explicitness and typicality of instructions and actions demanded from respondents (Kay & Saucier, 2023). As shown in the next part of this paper, they also differ in their ability and validity to identify inattentive participants and in response mechanisms (and problems) related to a particular attention check type.

3. ATTENTION CHECKS - WHAT WE KNOW, WHAT WE DO NOT KNOW

This section aims to provide a review of vital topics related to attention checks validity. Their ability to identify inattentive participants and correlates of attention check passing rates are discussed here. Summary of problems and research gaps with regard to attention checks concludes the section.

Consequences of screening sample on the basis of attention checks passing

The evidence on attention checks' ability to validly identify inattentive participants is debatable: some studies point to enhanced data quality after screening out participants who failed attention checks (Abbey & Meloy, 2017; Alvarez et al., 2019; Kam & Chan, 2018; Maniaci & Rogge, 2014; Oppenheimer et al., 2009), while others point to negligible, null or even adverse effects of such screening (Berinsky, et al., 2016; DeSimone & Harms, 2018; Gummer et al., 2021; Ladini, 2022; Shamon & Berning, 2020; Silber et al., 2022). Main indicators of higher data quality after screening participants who failed attention checks are: a) higher internal consistency, b) better model fit and c) stronger relations with validity indicators, e.g. higher regression coefficients and R^2 in regression equations (Abbey & Meloy, 2017; Maniaci & Rogge, 2014). **However, there is no systematic evidence on when such screening can improve data quality.** For example, it was not systematically shown which attention checks and in which conditions enhance data quality. We also do not know, how many attention checks should be used, and how many should be failed to validly identify inattentive participants.

Large number of false positives as well as false negatives produced by some attention checks may stem not only from excessive difficulty of these checks, but also **negative attitudes towards them** that may result in **spillover effects**, negative survey behaviour, such as socially desirable responding, negative survey attitudes, noncompliance with instructions, item-nonresponse or even drop-outs (prematurely terminated surveys). There is evidence that such effects indeed take place in survey practice (Berinsky et al., 2016; Clifford & Jerit, 2015; Gummer et al., 2021; Silber et al., 2022). For example, Silber et al. (2022) and Liu and

Wronski (2018) showed that, respectively, 61% or 44% of respondents who failed an attention check reported noticing it. According to Silber et al. (2022) this result shows that a large proportion of participants who failed the check did it for some other reasons than inattentiveness. Such reasons may be negative attitudes towards attention checks, Silber et al. (2022) reported that some participants see them as confusing or manipulating (ca. 10%), controlling or do not want to be instructed (around 25%). These negative attitudes are twice or thrice more frequent among non-complaint and inattentive groups than the attentive group (Silber et al., 2022).

Systematic research on **participants' attitudes** towards attention checks is one of the keys to eliminating false positives and false negatives, simultaneously reducing spillover effects, but until now only **very few studies have scrutinized the topics of attitudes,** and **behavioural processes used to answer attention checks**. It seems that above-mentioned problems with attention checks', **interpretation** may be responsible for false participants classifications as failing attention checks can arise not only from inattentiveness but also from noncompliance, confusion, misinterpretation, taught survey behaviour (e.g. skipping instructions) or just local distraction (Curran & Hauser, 2019; Silber et al., 2022).

There is also evidence that participants can **learn how to avoid or pass attention checks without paying much attention to other survey questions** (Hauser & Schwarz, 2016). The evidence of experienced participants' familiarity with attention checks is robust – in a study of Lovett et al. (2018) 92% of surveyed MTurk workers said they were "on the lookout" for attention checks and similar measures in tasks they perform. Qualitative data from this study also confirms that experienced survey takers are well familiar with the attention checks.

When confronted with attention checks, participants may feel observed or even controlled, which may result in adverse survey behaviour that appears on noticing an attention check. Using this control method may be also seen as breaking the social contract that both parties (researcher and survey participant) act honestly and diligently in a survey (Przybyłowska & Kistelski, 1986; Schwarz, 1995). Thus, attention checks may be seen by some participants as introducing negative interaction environment and discourage participants from participating diligently (Liu & Wronski, 2018), or even evoke a tit-for-tat mentality with purposeful inattentive responding or at least ignoring/purposeful failing of attention checks (Anduiza & Galais, 2017; Jones et al., 2015).

Some participants may also treat adding attention checks as breaking Gricean conversation norms (Grice, 1975), e.g., as including irrelevant information, thus breaking the maxim of relation (cf. Gummer et al., 2021). Participants actively interpret everything they see in a survey, from the questions' content, through instructions to formal questionnaire characteristics, such as the number of response categories and their numerical and verbal labels (Schwartz, 1995). It is worth

emphasising that attention checks are also prone to be subjectively interpreted by participants themselves. It is the researchers' responsibility to prevent misinterpretations and opaque inferences about the aims of attention checks. The research shows that informing participants about presence and purpose of attention checks increases passing rate but only slightly, e.g. Silber et al. (2022) reported an increase of a few points, from around 92% to 96%. There is possibility that providing reasons may also lower negative reactions and increase attention checks validity (Gummer et al., 2021; Silber et al., 2022) but such evidence is still to be provided.

It seems that **"innocuous"** attention checks should be used so that they are not noticed by participants, and responded as regular items (Kay, 2023; Kay & Saucier, 2023). Item "*I like to spend my time doing things I enjoy*" is an example of this type of attention check - it seems very "regular", but at the same time it is hard to not agree with. Such checks can provide more valid identification of inattentive participants and also reduced spillover effects. Kung et al. (2018) argued that the presence of attention checks did not bring any validity threats, as shown by the unchanged scale means, and evidenced by measurement invariance. However, more studies alike are needed, with more measurement instruments, more types of attention checks (Kung et al. tested only IRI and IMC), and more elaborated validity checks.

Precise guidelines on how to implement attention checks are also missing (for some evidence-based guidelines please consult section 4 of this paper (practice and recommendation)). For example, it is unknown, **how many attention checks and of which type should be used** in a survey to balance sensitivity and specificity, in order to identify inattentive participants, time efficiency and prevention of unwanted participants' reactions (spillover effects). Previous studies often used just one attention check per study (Gummer et al., 2021; Silber et al., 2022) or a whole battery of checks (Maniaci & Rogge, 2014). The former approach seems to offer insufficient power to detect most of inattentive participants (Berinsky et al., 2014; Keith et al., 2017; Paas & Morren, 2018), while the latter is unfeasible in everyday survey practice due to time and motivation restrictions (Arthur et al., 2021). Beach (1989) proposed a formula to calculate the probability of identifying an inattentive responder, where the said probability increases with the number of items dedicated to identify C/IER. It is also worth remembering that, especially with the use of IRIs and bogus/IFIs, even a completely inattentive respondent can pass these attention checks by chance. If an IRI has five response options, then the probability of choosing the correct one by random is a non-negligible 0.2. However, if two five-point IRIs are used, then the probability of answering them both correctly and by random is only 0.04 (provided that we treat both answers as independent events, which they probably are not). Exact probabilities depend on

many factors, including treating inattentiveness dichotomously or as a continuum, the number of response categories in the IRIs/IFIs used, the number of attention checks, etc. This problem probably calls for a separate simulation study.

Having more than one attention check in a survey should also help to measure the ebbs and flows of participants' attention in the course of responding (Anduiza & Galais, 2017; Berinsky et al., 2014). It is well-evidenced that attention in a survey is not constant, with a probability of decrease at the end of a lengthy survey (Galesic & Bosnjak, 2009). Hence, there is a need for more attention checks to enhance the chances of capturing possible moments of inattention (Jones et al., 2015). Comparing attention checks with the newly appeared model-based methods to dynamically model participants', responding process in a survey (Merhof & Meiser, 2023; Welz & Alfons, 2023) could also bring interesting information on number and placement of attention checks in a survey.

Covariates of attention checks passing rates

Failing attention checks is related to a number of indices traditionally linked with C/IER, like: straightlining, low reliability, speeding, item-nonresponse, inconsistent responses, e.g. logical errors, skipping instructions, googling answers, and self-reporting low effort and inattention (Alvarez et al., 2019; Babakhani et al., 2022; Gummer et al., 2021; Jones et al., 2015; Kam & Chan, 2018; Maniaci & Rogge, 2014). Interestingly, **attention checks seem to be never used in model-based approaches** to account for C/IER (at least to the author's best knowledge), and it is unclear whether they would offer incremental validity over and above other indices used in these models (e.g. response times).

The review of previous research also points to various important covariates of attention check passing rate, e.g., the type of sample matters, with volunteer samples (supposedly motivated internally, e.g. by topic interest) yielding higher success rates in attention checks in comparison to participants motivated only by monetary incentives (e.g. crowdworking platforms or opt-in panels; Maniaci & Rogge, 2014; Shamon & Berning, 2020). It was suggested that high levels of attention checks passing are characteristic of high-quality panels, especially probability-based panels (Lundmark et al., 2023). Large differences in detected careless levels were also identified between various types of panels and, interestingly, between different crowdworking platforms, with MTurk participants identified as yielding the highest levels of C/IER and lowest attention checks passing rates (Douglas et al., 2023; Peer et al., 2017). This may be related to socio-demographic composition of this platform, as MTurk participants are much younger than general American population, has lower earnings and is more educated (Huff & Tingley, 2015), as well as spend more time online and has higher Internet skills (Hargittai & Shaw,

2020). MTurk samples are also characterised by the monetary motivation that predominantly guides its workers (Chandler et al., 2019; Peer et al., 2022).

Platforms and panels that contain a large proportion of participants that treat web surveys as an important part of their income seem to be more at risk with large C/IER levels, and will typically result in substantial percent of participants flagged as inattentive and failing attention checks (Keith et al., 2017). Moreover, "blocked" MTurk workers, participants flagged as yielding data of lower quality, failed attention checks more often than Standard or Approved MTurk workers (unverified and high data quality groups, respectively) in a study presented by Hauser et al. (2022). Hence, it seems that higher topic interest, larger survey experience, intrinsic motivation, and positive attitudes towards attention checks all predict a higher success rate (Anduiza & Galais, 2017; Gummer et al., 2021; Mancuso et al., 2019; Paas & Morren, 2018; Silber et al., 2022).

The relationship between passing attention checks and other variables, like gender, age or cognitive ability, remains unclear (Anduiza & Galais, 2017; Maniaci & Rogge, 2014; Paas & Morren, 2018). Morren and Paas (2020) suggest that younger and more educated respondents are typical "weak" satisficers, whereas older and less educated respondents are more often "strong" satisficers. "Weak" satisficers process survey questions superficially, but yield responses similar to attentive respondents, while "strong" satisficers provide more disordered response vectors. Other studies found that younger and less educated participants fail attention checks more often (Alvarez et al., 2019; Berinsky et al., 2014; Gummer et al., 2021; Mancuso et al., 2019; Silber et al., 2022). Regarding gender differences, some studies identified none (Alvarez et al., 2019; Anduiza & Galais, 2017; Silber et al., 2022), while others identified males, as failing attention checks more often (Berinsky et al., 2014; Maniaci & Rogge, 2014; Olamijuwon, 2021). Most probably, these socio-demographic differences are moderated by participants' interest and motivation to participate in the survey, with intrinsic motivations predicting lower C/IER and attention checks failing (Anduiza & Galais, 2017).

Interestingly, no differences in attention checks' passing rates were found between compared devices: computers, tablets, and smartphones (Gummer et al., 2021; Silber et al., 2022). Further studies in this vein are needed, though, as there is evidence for different behavioural patterns on various devices, e.g. smartphone users tend to multitask more and take surveys in a more distractive environment (Antoun et al., 2017). However, measurement error differences across devices seem to stem from individual differences of their users, not devices per se (provided that a survey is well-designed for all types of devices used in a study; Lugtig & Toepoel, 2016). Hence, respondents choose devices according to their preferences, motivations and attitudes towards survey participation. These traits can affect data quality, including C/IER and attention checks passing rates. Further

studies comparing cross-device response processes is clearly needed to understand mechanisms of these differences.

Failing/passing attention checks is also stable in time, as most of the participants who failed attention checks in one part or wave of a survey will probably fail them too in another one (Ladini, 2022; Gummer et al., 2021; Olamijuwon, 2021; Paas & Morren, 2018). That means that attention checks are capable of grasping the temporarily stable part of the C/IER variance (Bowling et al., 2016). Indeed, they may offer a more time stable within-person measure of inattentiveness, as they are more comparable across studies than classic C/IER indices; correlation of bogus items from two different surveys amounted to 0.50, while between-survey correlations of other C/IER indices did not exceed 0.40 (Camus, 2015).

The type of attention check is also relevant, more difficult checks yield higher failure rates (Morren & Paas, 2020). Long instructional manipulation checks, which contain a large portion of text that has to be read and understood before answering, yield the highest level of failing rate (Anduiza & Galais, 2017; Liu & Wronski, 2018; Mancuso et al., 2019; Oppenheimer et al., 2009). IRIs and bogus items, but also short IMCs, normally result in a lower percent of sample flagged as inattentive (Brühlmann et al., 2020; Maniaci & Rogge, 2014). It is worth to note that a higher proportion of respondents flagged does not mean that long IMCs are anyhow better as they do not capture "all" or "more" inattentive respondents, but rather generate large numbers of false positives (Anduiza & Galais, 2017). Long IMCs also have lower within-person time stability (Anduiza & Galais, 2017) and are probably more related to cognitive abilities than other types of checks (Paas & Morren, 2018).

Attentive checks do indeed pick up negative survey behaviour (Gummer et al., 2021), but the effect of screening out participants who failed them on data quality, may depend mainly on percent of sample screened out (related to participants' motivation, sample type, topic interest and survey length) and type of attention check used (Anduiza & Galais, 2017; Maniaci & Rogge, 2014). So, the exact consequences of screening the sample out on the basis of attention checks passing are not yet known. Systematic evidence that would enable to assess consequences of such screening for substantial analyses and psychometric quality of the data is missing. Moreover, the evidence on which attention checks bring what consequences is also scarce. Similarly, participants' attitudes towards attention checks are not sufficiently researched and, as a consequence, we lack evidence on potential mechanisms of false classification as well as possible side effects of attention checks' presence on participants' survey behaviour.

Despite many studies, exact guidelines on implementing attention checks in surveys are still rather vague e.g., we do not know which attention check types are best to account for C/IER and how many checks should be used in a survey to

guarantee the lowest level of false- positives and negatives. Knowledge of whether attention checks offer any incremental validity over and above other indicators in model-based approaches is also absent.

Problems with attention checks

The popularity of attention checks seems to suggest that using attention checks is a valid and well-researched method to identify inattentive participants. The reality is not that rose-coloured, though. The proportion of respondents failing attention checks differs widely, depending on the specific task/item used, from just a few percent up to appallingly high numbers of up to 50, 60 or even 87% (Curran & Hauser, 2019; Liu & Wronski, 2018; Mancuso et al., 2019; Maniaci & Rogge, 2014; Oppenheimer et al., 2009).

These high numbers do not mean that most survey participants are inattentive and attention checks do an excellent job in identifying them. On the contrary, there is evidence that even half of participants failing the checks may do so due to other reasons than inattentiveness, e.g. purposeful noncompliance, misinterpretation or learnt survey behaviours (Anduiza & Galais, 2017; Babakhani et al., 2022; Brosnan et al., 2019; Curran & Hauser, 2019; Jones et al., 2015; Liu & Wronski, 2018; Silber et al., 2022). It has been shown that some participants may react adversely towards attention checks, e.g. by failing them on purpose, starting to respond in a desirable or noncompliant way or even starting to respond inattentively in a survey (Gummer et al., 2021). Hence, large numbers of participants flagged by attention checks come from false positives (participants wrongly identified as inattentive) and purposefully failed checks in addition to correctly identified careless responses. This constatation seriously undermines using attention checks in surveys.

Even worse, there is no firm consensus that screening out participants, who failed attention checks, can meaningfully enhance data quality in any respect (Gummer et al., 2021; Liu & Wronski, 2018; Shamon & Berning, 2020; but see Maniaci & Rogge, 2014). Higher data quality is expected after screening out inattentive participants, but this is demanding to achieve when attention checks generate such a high number of falsely classified participants. Anduiza and Galais (2017) point out another problem: screening out respondents may result in sample bias, especially when attention check used is correlated with traits other than attentiveness, such as cognitive ability or socio-demographic characteristics. Such screening may result in sample bias making it "oversophisticated" (Berinsky et al., 2014) and harder to generalise on the general population. The induced bias may also inflate effect sizes resulting even in false positive findings (Varaine, 2022). However, some researchers rebut the sample bias threat and argue that it is an

overestimated threat (e.g. Thomas & Clifford, 2017). Another danger, especially related with IMCs, is that some respondents are non-naive towards attention checks and can efficiently browse surveys searching for them (Hauser & Schwarz, 2016). Attention checks that are easy to spot are also easy to trick and such checks may result in **negative discrimination**.

Despite being ubiquitously used as screens against inattentive participants, attention checks are not researched enough to justify such universal and uncritical use. The research review presented above reveals main problems with attention checks summed in Table 1:

Table 1. Main Problems with Attention Checks

Problem	Description
Unknown consequences of using attention checks to screen out samples	Eliminating participants who failed attention checks seems obvious. Yet, more empirical evidence needs to be gathered to justify this method. So far positive results of such screenings have not been sufficiently proved.
Unsettling levels of false positives (false alarms) and false negatives (omissions)	There is evidence that at least some types of attention checks falsely identify participants as inattentive. Of course, all attention checks to some extent fail to identify some of the inattentive participants. Good attention checks should minimise both types of classification errors.
Spillover effects	Participants, spotting an attention check, may react in an unwanted and unpredicted way, e.g. purposefully respond inattentively, respond in a socially desirable or self-enhancing way, or even lose motivation to respond and drop out. Eliminating these side effects is of paramount importance.
Sample bias	Screening the sample out on the basis of attention checks may result in biasing the sample if passing attention checks is correlated with participants' characteristics, such as age, gender, education, cognitive abilities or personality.
Insufficient evidence on how to implement attention checks in practice	Which type of attention checks is best? How many should be used? How to use them to screen out the sample? How to implement them in statistical models to account for C/IER? These and many other important questions are still difficult to answer in an evidence-based way.
Unknown participants' attitudes and mechanisms of responding	Participants' knowledge and attitudes towards attention checks are largely unknown, as are mechanisms used to answer them or strategies to avoid/trick them. Expanding researchers' knowledge on these phenomena would help to design better attention checks and also enlarge understanding of survey behaviours.

To be fully useful as screening against inattentive participants, attention checks need to **improve** their levels of **sensitivity** and **specificity** to C/IER. Additionally, they should not cause **any negative spillover effects**, i.e., unwanted changes in

participants' behaviour, such as socially desirable responding, negative survey attitudes, noncompliance with instructions, item-nonresponse or drop-outs. It seems that we do not know as much about attention checks, as we have thought. However, we possess enough knowledge to note that we need more valid (generating less false positives and false negatives) attention checks, which would also generate less unwanted survey behaviour (reduced spillover effects) and will not lead to biased samples. One of the ways to achieve it is to develop and validate new types of attention checks that would be less ostensive, e.g. infrequency/frequency items (Kay & Saucier, 2023) or mock vignette checks, which are attention checks disguised as short texts and a number of factual questions that follow the text to assess participants' understanding of it (Kane et al., 2023). Such checks could lead to the higher screening quality. To do so, we have to significantly broaden our knowledge about attention checks and mechanisms of responding to them.

4. POPULARITY AND USE OF ATTENTION CHECKS – (FOREGOING) PRACTICE AND RECOMMENDATIONS

Popularity of attention checks

As stated by Kam and Chan (2018), attention checks are massively used and will continue to be so. Their popularity is driven by ease of use and high face validity. Most researchers do not delve deep into methodological concerns around attention checks and simply use them to easily measure participants' (in)attentiveness. Failing attention checks is also used as a straightforward argument to withhold payments for crowdworking platforms or opt-in commercial panels participants (Alvarez et al., 2019; although each panel provider/platform has its own policies here, e.g. Prolific Team, 2023).

All that makes attention checks immensely popular - fundamental papers about this topic (e.g. Meade & Craig, 2012; Oppenheimer et al., 2009) were already cited more than 3000 times (as per Google Scholar). This method to account for C/IER is ubiquitously used in social, individual differences, or industrial and organisational psychology, marketing studies, political studies, and all other branches of social sciences (Baumgartner & Weijters, 2022). Despite the immense popularity, step-by-step practical recommendations are still lacking. Based on the evidence gathered to date, it can be concluded that:

Which type of attention checks should I use?

Rosenzweig et al. (2023) formulated a few components of a proper attention check: a) it should measure only attention, not other constructs like memory, knowledge,

etc.; b) it should have only one, undoubtedly and clearly correct answer; c) it should not be overly difficult or very easy, d) it does not require violating norms of human behaviour. On the basis of the above review it can be added: e) they should have the same task difficulty and motivation to respond as other survey items, f) yield no correlations with socio-demographic and cognitive characteristics.

It seems that long IMCs do not fulfil any of these requirements: they are too complex, they deceive participants, and they measure other constructs. They also seem more prone to participants' who browse surveys in the search of attention checks to trick them and respond to the rest of the survey inattentively as they are easily identified (Hauser & Schwarz, 2016; Lovett et al., 2018; Peer et al., 2017).

Short IMCs, especially in the form of mock vignette checks (Kane et al., 2023), seem to be a much better choice. However, they seem better suited for psychological experiments or surveys in which participants have to do other tasks (e.g. read and comprehend longer text instructions) than just answer rating scale items. IRIs and IFIs seem to be the best choice for the latter case. An especially interesting type of attention checks are subtle IFIs (Kay, 2023; Kay & Saucier, 2023). By their covert nature, they can be more difficult to identify and trick by participants screening the questionnaire for attention checks, but they can also reduce or even eliminate spillover effects – subtle IFIs are very similar to regular items and most participants probably do not perceive them as attention checks. Maths quiz items and other relatively newly proposed attention check types are so far insufficiently researched to be recommended.

How many attention checks should I use?

This depends on the survey length, but the evidence strongly favours using more than one attention check in the survey (Berinsky et al., 2014; DeSimone et al., 2015). How many should be used? Meade and Craig (2012) suggest one attention check per every 50–100 items in a survey, while Kam and Chan (2018) propose slightly more: one per every 50–60 items, with both research teams advising no more than 3 to 5 overall per survey. Kay and Saucier (2023) recommend even more checks: one pair of IFIs per every 40 survey items with a minimum of two pairs of such attention check items in a questionnaire. Panel providers companies offer similar advice, e.g. Prolific allows two attention checks in surveys lasting 5 minutes or longer (Prolific Team, 2023). Some researchers and practitioners advocate mixing various types of attention checks (Keith et al., 2017; Peer et al., 2022; Prolific Team, 2023), e.g. more overt and more covert, or easier and more difficult, to capture "strong" and "weak" satisficers (Paas & Morren, 2018).

It is worth noticing that experimental evidence on which number and mixture of attention checks allows the best identification of inattentive participants is

largely lacking. The evidence shows, though, that it is recommendable to disperse the attention checks over the course of the survey in order to capture lapses in participants' attention (DeSimone et al., 2015). Prolific Team argues that at least one attention check should be placed at the beginning of the survey to screen out the inattentive respondents early on (2023). This is especially crucial if any type of dynamic reaction to participants' inattentiveness, e.g. prompting, is to be used.

How should I screen out the sample based on attention check responses?

Typically, participants who fail attention checks are eliminated (screened out) from the sample on which analysis is then performed. Prolific Team argues that, to limit false positives to a minimum, participants need to fail at least two attention checks to be screened out (2023). It seems that the best strategy is to use attention checks that have the highest validity and reduce spillover effects to a minimum, such as subtle IFIs, use more than one to reduce false positives, and do not be too conservative in removing flagged participants.

Before excluding any of the data, it is warranted to use at least one method of C/IER identification other than attention checks (Bauer et al., 2020; Curran, 2016; DeSimone et al., 2015). The response times analysis seems to be one of the best among such methods, provided that item-level response time indices are used (Kroehne et al., 2019; Ulitzsch et al., 2022a, 2022b). Unlike other *post-hoc* C/IER indices, e.g. longstring or intra-individual reliability, response times are easy to interpret and set valid cut-off points. They also seem to be able to validly identify inattentive participants in surveys of varying length, whereas other *post hoc* indices were found more suitable for long surveys only (DeSimone et al., 2015). However, some indices, such as Mahalanobis distance, even-odd reliability or longstring can be useful add-ons to attention checks and response time analysis (see Curran, 2016 for recommendations on these measures under specific survey conditions).

More technically advanced are response process indices based on computer-based paradata (e.g. Pokropek et al., 2023). Self-report measures of attention or commitment statements, in which participants are informed about the importance of data quality and are asked to agree to provide only attentive responses, are also useful supplements for other methods. Commitment statements or special instructions informing participants about the necessity of attentive responding can also serve as a preventive method, limiting C/IER from the early beginning (Geisen, 2022; Huang et al., 2015). An alternative to eliminating participants on the basis of C/IER indices would be to implement the information on attention checks responses into one of the models used to account for C/IER, e.g. based on response times or person fit indices (Niessen et al., 2016; Ulitzsch et al., 2022a,

2022b). However, to date, no such model using this information has been proposed.

To avoid data hacking, all data cleaning procedures should be planned beforehand and, preferably, communicated *a priori*, e.g. as a preregistered report. All data cleaning steps should also be comprehensively described in all scientific reports. As put by Mellis and Bickel (2020): *"To live up to the potential of highly transparent, reproducible science (...), researchers should clearly report inclusion/exclusion criteria, data quality checks and reasons for excluding collected data, how and when data were collected and both targeted and actual participant compensation."*

If failing attention checks or other inclusion criteria caused a reduction or withholding participants' compensation it should also be noted. Any criteria affecting participants' remuneration should be negotiated with panel provider companies and/or clearly communicated to participants before the survey starts, e.g. in task description or instructions. Communicating payment and its rules is a matter of retaining high ethical standards, along with asking for informed consent, providing an overview of the study and its estimated time (Bauer et al., 2020) or seeking consent to collect paradata (Henninger et al., 2022).

It is also relevant to check the consequences of screening on data quality and sample composition (DeSimone et al., 2015; Keith et al., 2017). All analyses should be performed before and after the screening and both sets of results need to be reported. Analysing participants who failed attention checks before discarding them completely is crucial as sometimes their data can also have good qualities (Waites & Ponder, 2016). Both sample size and composition will likely be affected by screening, as filtering decreases sample size and can systematically change its composition, especially if passing rate is correlated to socio-demographic characteristics or other variables, e.g. cognitive abilities, attitudes or personality (Rubio Juan & Revilla, 2021). Hence, the socio-demographic representation of the sample should be checked before and after screening out (Geisen, 2022; Keith et al., 2017). Lowered sample size and decreased sample generalisability after screening out can be remedied with sampling more respondents to retain power and sample composition (Curran & Hauser, 2019). Alternatively, participants, who fail attention checks, can be eliminated in real-time, if the data provider enables this option. In this way no oversampling would be needed, also this method allows to fill quotas as designed (Geisen, 2022).

Most important recommendations are summed up in Table 2:

Table 2. Practical Recommendations for Using Attention Checks in Computer-based Surveys

Question	Recommendation	Justification
What type of attention checks should I use?	Use IFIs, IRIs or short IMCs. Long IMCs are not recommended.	Attention checks should be adjusted to the design and need of a particular survey. They will work best if they resemble regular survey items that you are going to use. They should also validly identify inattentive participants with limited false classification and spillover effects.
How many attention checks should I use?	More than one, more are needed if the survey is long.	Multiple attention checks should be used; one is not enough to validly identify inattentive respondents. Longer surveys need more attention checks, e.g. two for every five minutes of a survey.
How should I use attention checks to screen out participants?	Use more than one criterion. Analyse data from participants flagged as inattentive. Report consequences of data screening for sample composition and statistical analyses.	More than one criterion of inattentiveness should be used to eliminate a respondent. Response time indices are a good choice for a complementary index. Data from flagged participants should be analysed before elimination.
What should I describe in my paper?	All decisions should be clearly described. It's best to plan data cleaning criteria beforehand.	Criteria to eliminate any participants should be clearly stated and described. Consequences of cleaning for sample size, composition, psychometric qualities of measures, and statistical analyses should be presented. Key analyses should be performed on both full and cleansed samples.
What should I tell the participants?	If failing attention checks result in losing remuneration, participants should be informed about that fact. Explain why you use attention checks.	Participants should be clearly informed about the possible consequences of being identified as inattentive respondent. Information on the presence of attention checks in a survey can work as a preventive method against C/IER. Providing such explanations is also recommended if overt attention checks, e.g. IRIs, are used.

5. FUTURE DIRECTIONS AND CONCLUSIONS

Future directions in attention checks research

The future research projects should aim to fill in pivotal research gaps, such as determining in a systematic way, whether using attention checks to screen out participants who failed them leads to enhanced data quality. A systematic re-analysis of studies using attention checks to screen samples is recommended.

Additionally, an analysis of relations between attention checks' passing rates and other variables, such as socio-demographic traits, personality, and cognitive abilities, should be performed to identify unwanted relations between attention checks and other variables (Rubio Juan & Revilla, 2021).

Comparing attention checks validity across different samples should be also performed as different groups of participants differ in motivational states affecting data quality (Litman et al., 2015; Maniaci & Rogge, 2014; Shamon & Berning, 2020). Opt-in panels participants and respondents from crowdworking platforms are often motivated economically, while student or intercept samples, as well as probability-based panels respondents, are driven more by interest and internal incentives (Lundmark et al., 2023).

Qualitative data, for example coming from cognitive labs or both individual and group interviews, can also bring core information on attention checks, interpretation, use and validity, as seen by the respondents (cf. Curran & Hauser, 2019; Lovett et al., 2018). More in-depth knowledge of how participants process attention checks and respond to them, can be acquired by asking participants to describe their mental processes and emotions towards surveys with different attention checks embedded (Jobe & Mingay, 1989). To date, as to the best of the author's knowledge, no such study has been conducted.

Furthermore, auxiliary data, such as computer-based paradata (log-data) or eye-tracker data, should be also used to comprehend mechanisms of careless responding better. Log-data indices were not employed to study attention checks before, but their potential role in uncovering mechanisms of responding to attention checks was noticed before (e.g. Liu & Wronski, 2018). Paradata indices seem a better way to study non-compliance than asking dichotomous yes/no questions used to study before, e.g. whether participants have noticed an attention check (Liu & Wronski, 2018; Silber et al., 2022). Moreover, any differences in participants' responding processes before and after spotting an attention check can be tracked by log-data in order to find any change in behaviour.

Employing attention checks to model-based approaches to account for C/IER is a recommendable research step. Such investigation could examine whether attention checks offer any incremental validity over and above information used in these models. Models such as latent class analysis (Morren & Paas, 2020), person fit (Niessen et al., 2016) or mixture models (Ulitzsch et al., 2022a, 2022b) could be used as a first step.

Future studies should also investigate the role of the number of attention checks in a survey in their ability to identify inattentive participants correctly. This is relevant, as the larger the number of attention checks is, the lower remains the possibility of false positives and false negatives. However, a precise ratio needs to be established due to the time efficiency and prevention of unwanted participants' reactions.

Moreover, an experimental test of whether attention checks are really capable of differentiating between attentive and inattentive participants, should be conducted to provide experimental, not only correlational, evidence that attention checks are in fact capturing C/IER. This would enable to determine **why** they work. Comparisons between the validity of different attention checks and validation of new attention check types, e.g. maths quiz items and mock vignettes (Kane et al., 2023) are also needed.

Attention checks can also serve to **detect bots**, but this possibility has received limited research attention. Bots and automated scripts pose potential threats to the validity of web surveys (Goodrich et al., 2023; Ilagan & Falk, 2023) and are difficult to identify due to a constant "arms race" and continuous bot refinement. Previously used bot detection methods, such as CAPTCHA items or honey pots (items hidden from human participants but visible to bots), may not be as effective as they once were (Storozuk et al., 2020). Recent research has indicated the limited effectiveness of simple attention checks in bot detection, but more research is still needed in this area (Saad, 2021; Storozuk et al., 2020; Zhang et al., 2022).

Finally, most of the attention check research is concentrated on attitudinal surveys and using checks to identify inattentive participants in more factual surveys is underresearched. Answers to factual questions (e.g. number or frequency of certain actions, e.g. visiting a doctor in the last year) can be at least sometimes verified, which offers additional indicators of C/IER. The logic of attention checks can be also used in cognitive tests, especially when log data such as response times are not available. Response times and participants' actions, such as response editing are often and successfully used to model C/IER in tests (Ulitzsch et al., 2020; Welling et al., 2023). However, some modification of attention checks could prove useful in situations when log data are unavailable.

Conclusions

Careless responding is one of the most serious threats to data quality in self-report research, especially in online surveys (Berinsky et al., 2016). Because attention checks are one of the most popular methods to account for it, they exert a huge influence on online studies' evidence quality and should be well-validated to justify their use to account for C/IER. More extended knowledge of attention checks and refining the practice of their use, offers one of the best possible "silver bullets" against C/IER. This is because attention checks have certain, undeniable, and unique assets: they are quick and effortless to use, easier to interpret than other methods proposed to screen out inattentive participants (such as indicators of straightlining, random responding or person fit indices; Ward & Meade, 2023).

Recent research points out that, when appropriately used, some attention checks can prove very useful in data cleaning (Kay & Saucier, 2023; Maniaci & Rogge, 2014; Meade & Craig, 2012). It appears that problems with attention checks stem from an insufficient understanding of how respondents process, interpret and answer them. Moreover, many disparate types of attention checks were proposed, albeit only some were widely used and tested. More evidence on processes engaged in responding to attention checks is needed too.

FUNDING

This research is financed by the National Science Centre (NCN) research grant (2019/33/B/HS6/00937) *Understanding response styles in self-report data: consequences, remedies and sources*, awarded to Artur Pokropek.

6. REFERENCES

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, *53*, 63-70. https://doi.org/10.1016/j.jom.2017.06.001

Alvarez, R. M., Atkeson, L. R., Levin, I., & Li, Y. (2019). Paying attention to inattentive survey respondents. *Political Analysis*, *27*(2), 145-162. https://doi.org/10.1017/pan.2018.57

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497-519. https://doi.org/10.1093/ijpor/edw007

Antoun, C., Couper, M. P., & Conrad, F. G. (2017). Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel. *Public Opinion Quarterly*, *81*(S1), 280-306. https://doi.org/10.1093/poq/nfw088

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-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

Arthur Jr, W., Hagen, E., & George Jr, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 105-137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Babakhani, N., Paas, L., & Dolnicar, S. (2022). Do instructional manipulation checks measure inattention or miscomprehension?. *International Journal of Social Research Methodology*, 1-12. https://doi.org/10.1080/13645579.2022.2111064

Barends, A. J., & De Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality*

and individual differences, *143*, 84-89. https://doi.org/10.1016/j.paid.2019.02.015

Bauer, B., Larsen, K. L., Caulfield, N., Elder, D., Jordan, S., & Capron, D. (2020). Review of best practice recommendations for ensuring high quality data with Amazon's Mechanical Turk. https://psyarxiv.com/m78sf/download?format=pdf. https://doi.org/10.31234/osf.io/m78sf

Baumgartner, H., & Weijters, B. (2022). How to Identify Careless Responders in Surveys, Baumgartner, H. & Weijters, B. (Eds.) *Measurement in Marketing (Review of Marketing Research, Vol. 19)*, Emerald Publishing Limited, Bingley, pp. 121-141. https://doi.org/10.1108/S1548-643520220000019007

Beach, D. A. (1989). Identifying the random responder. *The Journal of Psychology*, *123*(1), 101-103. https://doi.org/10.1080/00223980.1989.10542966

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739-753. https://doi.org/10.1111/ajps.12081

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers?. *Journal of Experimental Social Psychology*, *66*, 20-28. https://doi.org/10.1016/j.jesp.2015.09.010

Brosnan, K., Babakhani, N., & Dolnicar, S. (2019). "I know what you're going to ask me" Why respondents don't read survey questions. *International Journal of Market Research*, *61*(4), 366-379. https://doi.org/10.1177/1470785318821025

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), 218–229. https://doi.org/10.1037/pspp0000085

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, 100022. https://doi.org/10.1016/j.metip.2020.100022

Camus, K. A. (2015). *Once careless, always careless? Temporal and situational stability of insufficient effort responding (IER)*. [Master's thesis, Wright State University]. CORE Scholar. https://corescholar.libraries.wright.edu/etd_all/1617/

Carden, S. W., Camper, T. R., & Holtzman, N. S. (2018). Cronbach's alpha under insufficient effort responding: an analytic approach. *Stats*, *2*(1), 1-14. https://doi.org/10.3390/stats2010001

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*, 2022-2038. https://doi.org/10.3758/s13428-019-01273-7

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias?. *Public Opinion Quarterly*, *79*(3), 790-802. https://doi.org/10.1093/poq/nfv027

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, *64*(4), 464-494. https://www.jstor.org/stable/3078739. https://doi.org/10.1086/318641

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*(4), 596-612. https://doi.org/10.1177/0013164410366686

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

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & 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

DeSimone, J. A., & Harms, P. D. (2018). Dirty Data: The Effects of Screening Respondents Who Provide Low-Quality Data in Survey Research. *Journal of Business and Psychology,* 33, 559-577. https://doi.org/10.1007/s10869-017-9514-9

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171-181. https://doi.org/10.1002/job.1962

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

Freitas-Lemos, R., Tegge, A. N., Craft, W. H., Tomlinson, D. C., Stein, J. S., & Bickel, W. K. (2022). Understanding data quality: Instructional comprehension as a practical metric in crowdsourced investigations of behavioral economic cigarette demand. *Experimental and Clinical Psychopharmacology*, *30*(4), 415-423. https://doi.org/10.1037/pha0000579

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

Geisen, E. (2022, March 29). *Improve data quality by using a commitment request instead of attention checks.* Qualtrics.com. https://www.qualtrics.com/blog/attention-checks-and-data-quality/

Goodrich, B., Fenton, M., Penn, J., Bovay, J., & Mountain, T. (2023). Battling bots: Experiences and strategies to mitigate fraudulent responses in online surveys. *Applied Economic Perspectives and Policy, 45*(2), 762-784. https://doi.org/10.1002/aepp.13353

Göritz, A. S., Borchert, K., & Hirth, M. (2021). Using attention testing to select crowdsourced workers and research participants. *Social Science Computer Review*, *39*(1), 84-104. https://doi.org/10.1177/0894439319848726

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics, 3: Speech acts*, (pp. 41-58). New York: Academic Press. https://doi.org/10.1163/9789004368811_003

Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, *50*(1), 238-264. https://doi.org/10.1177/0049124118769083

Hargittai, E., & Shaw, A. (2020). Comparing Internet experiences and prosociality in Amazon Mechanical Turk and population-based survey samples. *Socius, 6*, 2378023119889834. https://doi.org/10.1177/2378023119889834

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary?. *Frontiers in psychology*, *9*, 998. https://doi.org/10.3389/fpsyg.2018.00998

Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2022). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*, 1-12. https://doi.org/10.3758/s13428-022-01999-x

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

Henninger, F., Kieslich, P. J., Fernández-Fontelo, A., Greven, S., & Kreuter, F. (2022). Privacy attitudes toward mouse-tracking paradata collection. *Osf preprint:* https://osf.io/6weqx/download. https://doi.org/10.31235/osf.io/6weqx

Horwitz, R., Kreuter, F., & Conrad, F. (2017). Using mouse movements to predict web survey response difficulty. *Social Science Computer Review*, *35*(3), 388-405. https://doi.org/10.1177/0894439315626360

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*, 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

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 2053168015604648. https://doi.org/10.1177/2053168015604648

Ilagan, M. J., & Falk, C. F. (2023). Supervised classes, unsupervised mixing proportions: Detection of bots in a Likert-type questionnaire. *Educational and Psychological Measurement, 83*(2), 217-239. https://doi.org/10.1177/00131644221104220

Jobe, J. B., & Mingay, D. J. (1989). Cognitive research improves questionnaires. *American Journal of Public Health*, *79*(8), 1053-1055. https://doi.org/10.2105/AJPH.79.8.1053

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

Jones, M. S., House, L. A., & Gao, Z. (2015). Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys. *Public Opinion Quarterly*, *79*(3), 687-709. https://doi.org/10.1093/poq/nfv015

Kam, C. C. S., & Chan, G. H. H. (2018). Examination of the validity of instructed response items in identifying careless respondents. *Personality and Individual Differences*, *129*, 83-87. https://doi.org/10.1016/j.paid.2018.03.022

Kane, J. V., Velez, Y. R., & Barabas, J. (2023). Analyze the attentive and bypass bias:

Mock vignette checks in survey experiments. *Political Science Research and Methods*, *11*(2), 293-310. https://doi.org/10.1017/psrm.2023.3

Kay, C. S. (2023). Validating the IDRIS and IDRIA: Two scales for detecting careless and insufficient effort survey responders. *Psyarxiv preprint: https://psyarxiv.com/us7bm/download?format=pdf*. https://doi.org/10.31234/osf.io/us7bm

Kay, C. S., & Saucier, G. (2023). The Comprehensive Infrequency/Frequency Item Repository (CIFR): An online database of items for detecting careless/insufficient-effort responders in survey data. *Personality and Individual Differences*, *205*, 112073. https://doi.org/10.1016/j.paid.2022.112073

Keith, M. G., Tay, L., & Harms, P. D. (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in psychology*, *8*, 1359. https://doi.org/10.3389/fpsyg.2017.01359

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214-233. https://doi.org/10.1177/0894439317752406

Kroehne, U., Buchholz, J., & Goldhammer, F. (2019). Detecting carelessly invalid responses in item sets using item-level response times. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*. Toronto, Canada.

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

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236. https://doi.org/10.1002/acp.2350050305

Kung, F. Y., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity?. *Applied Psychology*, *67*(2), 264-283. https://doi.org/10.1111/apps.12108

Ladini, R. (2022). Assessing general attentiveness to online panel surveys: the use of instructional manipulation checks. *International Journal of Social Research Methodology*, *25*(2), 233-246. https://doi.org/10.1080/13645579.2021.1877948

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519-528. https://doi.org/10.3758/s13428-014-0483-x

Liu, M., & Wronski, L. (2018). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, *60*(1), 32-49. https://doi.org/10.1177/1470785317744856

Lovett, M., Bajaba, S., Lovett, M., & Simmering, M. J. (2018). Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology*, *67*(2), 339-366. https://doi.org/10.1111/apps.12124

Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-

based panel survey: Effects on survey measurement error. *Social Science Computer Review*, *34*(1), 78-94. https://doi.org/10.1177/0894439315574248

Lundmark, S., Silber, H., Roßmann, J., & Gummer, T. (2023, July 17-21). *Comparing different types of respondents' attentiveness measures: Experimental evidence from the German Internet Panel and the Swedish Citizen Panel.* [Conference presentation] European Survey Research Association Conference, Milan, Italy.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

Mansolf, M., & Reise, S. P. (2018). Case diagnostics for factor analysis of ordered categorical data with applications to person-fit measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 86-100. https://doi.org/10.1080/10705511.2017.1367926

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

Mellis, A. M., & Bickel, W. K. (2020). Mechanical Turk data collection in addiction research: Utility, concerns and best practices. *Addiction*, *115*(10), 1960-1968. https://doi.org/10.1111/add.15032

Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, 1-27. https://doi.org/10.1007/s11336-023-09901-0

Morren, M., & Paas, L. J. (2020). Short and long instructional manipulation checks: What do they measure?. *International Journal of Public Opinion Research*, *32*(4), 790-800. https://doi.org/10.1093/ijpor/edz046

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*(1), 58-88. https://doi.org/10.1086/297739

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2), 239-250. https://doi.org/10.1002/1097-4679(198903)45:2%3C239::AID-JCLP2270450210%3E3.0.CO;2-1

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

Olamijuwon, E. O. (2021). Characterizing low effort responding among young African adults recruited via Facebook advertising. *Plos one*, *16*(5), e0250303. https://doi.org/10.1371/journal.pone.0250303

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872. https://doi.org/10.1016/j.jesp.2009.03.009

Paas, L. J., & Morren, M. (2018). Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters*, *29*, 13-21. https://doi.org/10.1007/s11002-018-9448-7

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023-1031. https://doi.org/10.3758/s13428-013-0434-y

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A. et al. (2022). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, *54*, 1643–1662, https://doi.org/10.3758/s13428-021-01694-3

Pokropek, A., Żółtak, T., & Muszyński, M. (2023). Mouse Chase: Detecting Careless and Unmotivated Responders Using Cursor Movements in Web-Based Surveys. *European Journal of Psychological Assessment, 39*(4), 299–306. https://doi.org/10.1027/1015-5759/a000758

Prolific Team (2023, March 29). *Prolific's Attention and Comprehension Check Policy.* Prolific.co. https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Prolific-s-Attention-and-Comprehension-Check-Policy

Przybyłowska, I., & Kistelski, K. (1986). The social context of questionnaire interview. *The Polish Sociological Bulletin*, (75/76), 21-26. https://www.jstor.org/stable/44816641

Rosenzweig, Ch., Edelman, J., & Moss, A. (2023, March 29). *Examples of Good (and Bad) Attention Check Questions in Surveys*. CloudResearch.com. https://www.cloudresearch.com/resources/blog/attention-check-questions-in-surveys-examples/

Rubio Juan, M., & Revilla, M. (2021). Comparing respondents who passed versus failed an Instructional Manipulation Check: A case study about support for climate change policies. *International Journal of Market Research*, *63*(4), 408-415. https://doi.org/10.1177/14707853211023039

Saad, D. (2021). Nowe narzędzia i techniki zwiększające trafność badań internetowych. *Com. press, 4*(1), 106-121. https://doi.org/10.51480/compress.2021.4-1.248

Savage, S. J., & Waldman, D. M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics*, *23*(3), 351-371. https://doi.org/10.1002/jae.984

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, *82*(1), 29-56. https://doi.org/10.1177/00131644211004708

Schwarz, N. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. *International Statistical Review/Revue Internationale de Statistique*, 63(2), 153-168. https://doi.org/10.2307/1403610

Shamon, H., & Berning, C. (2020). Attention Check Items and Instructions in Online Surveys with Incentivized and Non-Incentivized Samples: Boon or Bane for Data Quality?. *Survey Research Methods*, 14(1), 55-77. DOI/10.18148/srm/2020.v14i1.7374. https://doi.org/10.2139/ssrn.3549789

Silber, H., Roßmann, J., & Gummer, T. (2022). The Issue of Noncompliance in Attention Check Questions: False Positives in Instructed Response Items. *Field Methods*, *34*(4), 346-360. https://doi.org/10.1177/1525822X221115830

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice*, *38*(2), 101-111. https://doi.org/10.1111/emip.12256

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology, 16*(5), 472-481. https://doi.org/10.20982/tqmp.16.5.p472

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197. https://doi.org/10.1016/j.chb.2017.08.038

Toich, M. J., Schutt, E., & Fisher, D. M. (2022). Do you get what you pay for? Preventing insufficient effort responding in MTurk and student samples. *Applied Psychology*, *71*(2), 640-661. https://doi.org/10.1111/apps.12344

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, *55*(3), 425-453. https://doi.org/10.1080/00273171.2019.1643699

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022a). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593-619. https://doi.org/10.1007/s11336-021-09817-7

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

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

Varaine, S. (2022). How Dropping Subjects Who Failed Manipulation Checks Can Bias Your Results: An Illustrative Case. *Journal of Experimental Political Science*, 299-305. https://doi.org/10.1017/XPS.2022.28

Voss, N. M. (2023). The effects of careless responding on the fit of confirmatory factor analysis and item response theory models. *Behavior Research Methods*, 1-23. https://doi.org/10.3758/s13428-023-02074-9

Waites, S. F., & Ponder, N. (2016). May I have your attention please? The effectiveness of attention checks in validity assessment. In *Celebrating America's Pastimes: Baseball, Hot Dogs, Apple Pie and Marketing? Proceedings of the 2015 Academy of Marketing Science (AMS) Annual Conference* (pp. 475-476). Springer International Publishing. https://doi.org/10.1007/978-3-319-26647-3_97

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74,* 577-595. https://doi.org/10.1146/annurev-psych-040422-045007

Welz, M., & Alfons, A. (2023). I Don't Care Anymore: Identifying the Onset of Careless Responding. *arXiv preprint:* https://arxiv.org/abs/2303.07167

Welling, J., Gnambs, T., & Carstensen, C. H. (2023). Identifying Disengaged Responding in Multiple-Choice Items: Extending a Latent Class Item Response Model with Novel Process Data Indicators. *Educational and Psychological Measurement*, 00131644231169211. https://doi.org/10.1177/00131644231169211

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 186-191. https://doi.org/10.1007/s10862-005-9004-7

Zhang, Z., Zhu, S., Mink, J., Xiong, A., Song, L., & Wang, G. (2022). Beyond Bot Detection: Combating Fraudulent Online Survey Takers. In *Proceedings of the ACM Web Conference 2022*, pp. 699-709. https://doi.org/10.1145/3485447.3512230