Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Uniwersytet Warszawski
Wydział Zarządzania

Unia Europejska
Europejski Fundusz Społeczny

University of Warsaw

mgr Marta Kabut

False Respondents in Web Human Resource Surveys

Fałszywi respondenci w ankietach internetowych w badaniach z zakresu zarządzania zasobami ludzkimi

Doctoral dissertation
in the discipline of management and quality studies

Dissertation written under the supervision of
Prof. dr hab. Grażyna Wieczorkowska-Wierzbińska
Associate supervisor: dr. Anna Kuzminska
University of Warsaw, Faculty of Management
Managerial Psychology and Sociology Unit

Warsaw, 2021

projekt „Zarządzanie wielokulturowe w erze globalizacji" realizowany przez Wydział Zarządzania Uniwersytetu Warszawskiego na podstawie umowy nr POWR.03.02.00-00-I053/16-00 w ramach Programu Operacyjnego Wiedza Edukacja Rozwój 2014-2020 finansowanego z funduszy strukturalnych Unii Europejskiej

Oświadczenie kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie stopnia doktora.

Data Podpis kierującego pracą

Statement of the Supervisor on Submission of the Dissertation

I hereby certify that the thesis submitted has been prepared under my supervision and I declare that it satisfies the requirements of submission in the proceedings for the award of a doctoral degree.

Date Signature of the Supervisor

Oświadczenie autora pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca doktorska została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora pracy

Statement of the Author on Submission of the Dissertation

Aware of legal liability I certify that the thesis submitted has been prepared by myself and does not include information gathered contrary to the law.

I also declare that the thesis submitted has not been the subject of proceeding in the award of a university degree.

Furthermore I certify that the submitted version of the thesis is identical with its attached electronic version.

Date Signature of the Author

Zgoda autora pracy

Wyrażam zgodę ma udostępnianie mojej rozprawy doktorskiej dla celów naukowo-badawczych.

Data Podpis autora pracy

Consent of dissertation's author

I agree to make my dissertation available for research purposes

Date Signature of the Author

Abstract

Web surveys have become the dominant form of data collection in management sciences. Unfortunately, the ease of data collection does not accompany methodological diligence in data analyses. Some respondents do not read questions carefully, giving random answers, so they are called FALSE/inattentive respondents. Inattention may be global (whole or most of the survey) or local (when respondents were inattentive when answering a block of questions but answered others with due diligence).

FALSE respondents can be easily spotted during the interview but detecting them in an online survey is difficult. Unreliable data may change correlations, make the analysis and evaluation of the results difficult, decrease statistical power and effect size, and lower internal consistency.

The dissertation aims to develop and test PROCEDURE for detecting FALSE RESPONDENTS. Literature review ends with identification of four warning signs based on: WS1: answering time, WS2: attention check questions, WS3: response style and the number of Don't Know answers, and WS4: declarative and behavioural cooperation measures.

Two research tasks were carried out: (1) estimation of the level of inattention, (2) estimation of the consequences of ignoring the problem and testing the usability of the FLEXMIX procedure (finite mixtures of generalized regression models) for detecting FALSE respondents. Nine data sets from online surveys with the total of 5645 respondents and three offline (based on face-to-face interviews) data sets – Polish samples from World Values Survey and European Working Conditions Survey, with the total of 3169 respondents – were analysed. The percentage of FALSE respondents depended on the survey. The highest was 71% for data from the commercial online panel, lowest was 22.1%. Overall, percentages were the lowest for offline data files (from high-budget international surveys that were carefully designed and cleaned by international teams of researchers before they were made available to the public).

To estimate the consequences of ignoring the presence of FALSE respondents in data set, the reliability of the measurement (operationalized by Cronbach's Alpha) in 2 groups was compared. For the first group of ATTENTIVE (who passed all tests) respondents, Cronbach's alphas were congruent with theoretical assumptions. For the second group of FALSE respondents, the reliability of the same index was extremely low (in some cases even negative – which is the best proof that the FALSE respondent didn't notice reversed items).

The dissertation ends with presenting a procedure for detecting FALSE respondents, which can (and should) be used by all researchers. However, it should be emphasized that methods of data analysis and exclusion criteria should be planned before designing the survey to be sure that attention check questions have been introduced, answering time is measured, etc.

The results of my analyses cannot be generalized to all data collected in web surveys, because analysed data sets came from research thas has been carefully planned and conducted on groups of people who could participate by invitation only (employees registered in the commercial panel, employees studying at the Faculty of Management). All respondents were additionally motivated to participate and were informed in advance that their answers would be subjected to the procedure for detecting FALSE respondents. It can be predicted that data obtained from surveys that had links widely shared on social media platforms (like Facebook) contains a much higher percentage of FALSE respondents, who were answering randomly because e.g. they lost motivation while answering questions but were curious about the next page. These surveys are often poorly designed (e.g., without giving respondents the opportunity to choose "Don't Know" answer).

Analyses performed on uncleaned data could lead to FALSE conclusions, which, if incorporated into scientific circulation, harm the development of management research. HRM theories confirmed by biased (not reliable) data are not valid so FALSE respondents detection is an important pre-analysis task.

Key words:
false respondents, careless respondents, web surveys, flexmix, data cleaning

Abstrakt

Ankiety internetowe stały się dominującą formą gromadzenia danych w naukach o zarządzaniu. Niestety łatwość zbierania danych nie towarzyszy staranności metodologicznej w analizach danych. Niektórzy respondenci nie czytają uważnie pytań, udzielając przypadkowych odpowiedzi, więc nazywa się ich FAŁSZYWYMI/nieuważnymi respondentami. Nieuwaga może być globalna (całość lub większość ankiety) lub lokalna (gdy respondenci nie zwracali uwagi odpowiadając na blok pytań, ale odpowiadali na inne z należytą starannością).

FAŁSZYWYCH respondentów można łatwo zauważyć podczas wywiadu, ale wykrycie ich w ankiecie online jest trudne. Nierzetelne dane mogą zmienić korelacje, utrudnić analizę i ocenę wyników, zmniejszyć moc statystyczną i wielkość efektu oraz obniżyć spójność wewnętrzną.

Rozprawa ma na celu opracowanie i przetestowanie PROCEDURY wykrywania FAŁSZYWYCH RESPONDENTÓW. Przegląd literatury kończy się identyfikacją czterech sygnałów ostrzegawczych na podstawie: WS1: czas odpowiedzi, WS2: pytania sprawdzające uwagę, WS3: styl odpowiedzi i liczba odpowiedzi beztreściowych oraz WS4: deklaratywne i behawioralne środki współpracy.

Wykonano dwa zadania badawcze: (1) oszacowanie poziomu nieuwagi, (2) oszacowanie konsekwencji ignorowania problemu oraz przetestowanie przydatności procedury FLEXMIX (skończone mieszaniny uogólnionych modeli regresji) do wykrywania fałszywych respondentów. Przeanalizowano dziewięć zbiorów danych z ankiet internetowych z łączną liczbą 5645 respondentów oraz trzy zbiorych danych nieinternetowych (oparte na wywiadach bezpośrednich) – polskie próby z World Values Survey i European Working Conditions Survey, w sumie 3169 respondentów. Odsetek fałszywych respondentów zależał od ankiety. Najwyższy wyniósł 71% dla danych z komercyjnego panelu online, najniższy 22,1%. Ogólnie rzecz biorąc, odsetek ten był najniższy w przypadku zbiorów danych nieinternetowych (z wysokobudżetowych międzynarodowych badań, które zostały starannie zaprojektowane i oczyszczone przez międzynarodowe zespoły naukowców zanim zostały udostępnione publicznie).

Aby oszacować konsekwencje ignorowania obecności FAŁSZYWCH respondentów w zbiorze danych, porównano rzetelność pomiaru (operacjonalizowanego przez Alfę

Cronbacha) w 2 grupach. Dla pierwszej grupy UWAŻNYCH (którzy przeszli wszystkie testy) respondentów alfy Cronbacha były zgodne z założeniami teoretycznymi. Dla drugiej grupy FAŁSZYWYCH respondentów wiarygodność tego samego wskaźnika była wyjątkowo niska (w niektórych przypadkach nawet ujemna – co jest najlepszym dowodem na to, że FAŁSZYWY respondent nie zauważył odwróconych pytań).

Rozprawa kończy się przedstawieniem procedury wykrywania FAŁSZYWYCH respondentów, z której mogą (i powinni) korzystać wszyscy badacze. Należy jednak podkreślić, że metody analizy danych i kryteria wykluczenia powinny być zaplanowane przed zaprojektowaniem ankiety, aby mieć pewność, że wprowadzono pytania sprawdzające uwagę, mierzony jest czas odpowiedzi itp.

Wyników moich analiz nie da się uogólnić na wszystkie dane zebrane w ankietach internetowych, ponieważ analizowane zbiory danych pochodzą z badań, które zostały starannie zaplanowane i przeprowadzone na grupach osób, które mogły wziąć w nich udział wyłącznie na zaproszenie (pracownicy zarejestrowani w panelu komercyjnym, pracownicy studiujący na Wydziale Zarządzania). Wszyscy respondenci byli dodatkowo zmotywowani do udziału i zostali poinformowani z wyprzedzeniem, że ich odpowiedzi zostaną poddane procedurze wykrywania FAŁSZYWYCH respondentów. Można przewidzieć, że dane uzyskane z ankiet, do których linki były szeroko udostępniane na platformach społecznościowych (takich jak Facebook) zawierają znacznie wyższy odsetek FAŁSZYWYCH respondentów, którzy klikali losowo ponieważ np. stracili motywację podczas odpowiadania na pytania, ale byli ciekawi następnej strony. Ankiety te są często źle zaprojektowane (np. nie dają respondentom możliwości wyboru odpowiedzi „Nie wiem”).

Analizy przeprowadzone na nieoczyszczonych danych mogą prowadzić do FAŁSZYWYCH wniosków, które, jeśli zostaną włączone do obiegu naukowego, szkodzą rozwojowi badań z zakresu zarządzania. Teorie HRM potwierdzone przez obciążone (niewiarygodne) dane nie są prawdziwe, więc wykrywanie FAŁSZYWYCH respondentów jest ważnym zadaniem przed analizą danych.

Słowa kluczowe:

fałszywi respondenci, nieuważni respondenci, ankiety internetowe, flexmix, czyszczenie danych

Table of Contents

1. THEORETICAL PART

1.1 Introduction

The growing popularity of online surveys as a data collection tool is also present in social sciences, including management sciences, mainly because the developing world of technology and innovation has also caused changes in social sciences. The ability to study people without making them more inconvenienced than necessary has opened both new opportunities and new dangers for social scientists. One of such threats is FALSE respondents, people who decide to take part in an online survey but do not do it carefully: do not read the questions, answer randomly, play with the survey, or undertake other activities that, however, have little to do with attentive answering the questions (the truthfulness of the answers is a separate topic). Such people may introduce random noise to data collected in surveys, however, they usually do not answer entirely randomly, which results in a systematic bias in responses, and, as a result, a change in obtained results (obtaining statistically significant effects or, on the contrary, no results).

When it comes to management sciences, surveys, conducted more and more often online, are the most popular research method in human resource management. This happens because they are the most straightforward method - researching employees does not require more effort and commitment from participants than necessary, but sometimes even this is not enough. People taking part in the study may not even make a minimal effort to answer the questions mindfully (which does not have to be the subject's fault; it may be the researcher's fault, research topic, used survey tool, conditions). Data from such individuals are problematic: it may skew outcomes and make estimates of the sample/population overestimated or underestimated, influencing outcomes of almost every statistical analysis performed on such data. Therefore, FALSE respondents should be removed from the data sets (or analysed separately, if the sample size allows it) to avoid introducing systematic or non-systematic bias. This dissertation aims to present ways to detect FALSE respondents in human resource management surveys, identify and test the procedure of their detection, and propose a new detection method.

1.1.1 Structure of the dissertation

The dissertation consists of two parts: theoretical and empirical.

The first part of the dissertation focuses on internet research as a mode of data collection. The author presents advantages and disadvantages of internet research, online panels of respondents use, a model of a respondent answering survey questions, FALSE respondent description and impact of FALSE respondents on data quality. Then, classifications of FALSE respondent indicators are presented, and four warning signs and examples of studies using a part of the signs were presented. Other methods that were not used as part of the four warning signs will be briefly presented. At the end of this part, the author will construct a FALSE respondent model, the main aim of the research.

In the second part, the chosen methods of detecting FALSE respondents will be tested, and the results of this test will be presented. The author will show the overall and partial answering time as an indicator connected with reading speed, outcomes of the analysis of several data sets. Then, behavioural cooperation, low differentiation rating style and non-informative answers, declarative cooperation level, and logical consistency analysis will be shown. The dissertation ends with a summary, conclusions, and proposed analysis procedure designed to detect FALSE respondents.

General remarks about editing of the dissertation

In accordance with the supervisor's recommendation, the following standards were used to maintain the transparency of the argumentation and readability of the results:

1. Due to the exponential increase in the number of scientific publications on any topic, the literature review is limited to articles and other literature sources relevant to the research problem. From the point of view of synthesizing knowledge, the names of the research authors are the least important pieces of information, so instead of in parentheses – as the twentieth-century APA standard dictates – they are placed in footnotes. This way of referencing shortens the whole text by about 20% and makes it easier to focus on what is important (resulta) instead of on the history of research.
2. The volume of the first two parts of the doctoral dissertation should not exceed 100 pages. To facilitate perception of the content, the most

important concepts are distinguished using SMALL CAPS or bolding. New threads are separated in the American style by leaving free lines, instead of using uniform line spacing and indentation.

3. We do not avoid repeating the same words – scientific concepts – remembering that the doctoral dissertation is a scientific text, and the precision of the language is important. If we use synonyms, e.g. FALSE, inattentive, careless, it should be clearly indicated in the text.

4. When discussing the results of analyses, we focus only on the factors relevant to the interpretation. We do not enter statistics and significance levels into the text – if they are included in the tables. However, we introduce average values into the text even when they are presented in drawings, because the purpose of drawings is to illustrate the relationships found, so they can exaggerate the differences.

5. If the results of a series of studies are presented in a dissertation, the discussion of the results obtained can be presented together. Unless otherwise indicated under a specific table or drawing, graph, the source of all tables and figures presented in the dissertation is the work and own analysis of the author of the dissertation

1.1.2 Dissertation topic justification

Online surveys have replaced other ways of conducting studies and have already had a dominant position among quantitative research methods[1]. Growth dynamics are impressive: in 2006, about 20%, and **in 2013 above 50%** of all data collection expenditures were spent on online surveys.[2] Looking at a percentage share of scientific publications based on online surveys, we see a significant increase in the percentage of scientific publications, which would be much higher if we considered that it takes about four years from the study to its publication.

Years	Percent of search outcomes[3]
2012-2016	12.67
2017-2021	15.97

Table 1 *Outcomes of the Scopus bibliographic database search*

Most measures used in human resources are measured with surveys, so there is a need for data cleaning techniques that allow for control on respondents' engagement, which was seemingly less problematic in paper surveys[4].

Other examples of what can be researched using online surveys include customer, employee satisfaction surveys[5], market research[6], consumer preferences research[7].

Respondents tend to behave in different ways when taking such a survey. However, some of their behaviours are undesirable – they may not pay sufficient attention to the contents of the survey and consequently introduce bias to analysed data that can result in false results[8] and, therefore, false conclusions. It is crucial to distinguish data from attentive and FALSE (inattentive) respondents. While online surveys have become increasingly popular, new opinion polling companies have also sprung up. These companies bring together people in their research panels who view survey completion as an additional easy

[1] ESOMAR, 2014
[2] Vehovar & Lozar Manfreda, 2008; *ESOMAR, 2013*
[3] Number of search results for journal articles containing phrases ‘internet/online/web survey’ in abstract, keywords, or title, in Scopus bibliometric database, divided by the exact search containing just ‘survey’ word.
[4]Kiesler & Sproull, 1986.
[5]Kasvi, 2017; Barakat et al., 2015; Mitchell et al., 2021
[6] ex. Queloz & Etter, 2019; Kumar Mishra et al., 2016
[7] Molenaar et al., 2018
[8] Alvarez et al., 2019

job they are being paid for (often in the form of reward points). Having a more accessible way to get respondents means greater availability of respondents for researchers. Easiness comes with limited or lack of control over the behaviour and environment of the respondent. Lack of control means leaving honesty, when it comes to paying attention and the truthfulness of answers when filling out the survey, on the side of the respondent's good will. Relying on respondents' honesty has many downsides (the most problematic one being lack thereof), making it essential to check if the respondent behaved according to the researcher's intention. This makes FALSE respondents a vital issue, especially when people's work recommendations are made, which is the case in human resource research.

There are many studies on inattentive respondents that have been done on English-speaking samples[9]. However, this phenomenon has not yet been examined across disciplines for Polish samples. Although there is a related problem of turnout overreporting[10] (declaring participation in elections when one does not plan to participate), it is a specific case of social desirability bias. It can, but not necessarily occur along with the problem of FALSE respondents.

The author of this dissertation focuses on a specific case of online surveys in human resource management and research in the case of online surveys in which data come from Polish samples. The research gap to be filled by this research is determining the level of inattention of respondents, consequences of not excluding FALSE respondents from analysed data, and devising a relatively straightforward and easy to use/apply procedure to detect FALSE respondents in data sets from online human resource surveys.

[9] ex. Nichols & Edlund, 2020, Schneider et al., 2018, Bowling & Huang, 2018, Alvarez et al., 2019

[10] Górecki, 2011

1.1.3 Definitions of terms used in the dissertation

The literature review will start with the disambiguation of some terms used in the dissertation.

A **FALSE respondent** (FR, careless respondent, inattentive respondent, flagged respondent) is a person who voluntarily participates in a survey who, knowingly or not, does not cooperate and answers questions without thinking about the answer, chooses the first good enough answer, chooses random answers without reading the questions, chooses logically contradictory answers, or does not pay enough attention to answering the questions.

A **web/online/internet survey** is understood here as a self-administered, online questionnaire, explicitly used in human resources research, completing what is done by real people voluntarily, who are either being paid for the task in various ways or are unpaid participants. Although the title of the dissertation contains term ‘web survey’, it is a synonym to ‘online’ and ‘internet’ survey, and those two terms will be used in the main part of the dissertation interchangeably.

Overall answering time (OAT) is the time from the first load of the first survey page to the end page shown, which means it includes instructions and breaks that the respondent took (whether on a page explicitly instructing them that they can take a break or breaks which they took on other pages, indicated by answering significantly longer than other respondents to the question on that page).

Partial answering time (PAT) considers the time spent answering particular parts/parts of the survey, usually having the same type of questions, the same rating scale, and distinguished from other parts by either a set of new instructions or a break screen.

Words per minute (wpm) – an indicator of reading speed of the respondent, calculated by dividing the number of words presented to the respondent to read on a single survey page or specific survey part by the time that elapsed (in minutes for wpm, and the seconds in wps) between a single page, or first page of a set, fully loaded, and answering to the page, or a set of pages finished.

The behavioural cooperation level based on respondent's answers analysis means how respondent reacts (or not) to different types of attention check questions aimed at testing whether he read questions, comprehended what he has read, and acted accordingly to the when answering. Wrong answers to attention check questions mean a low level of behavioural cooperation, and correct answers can signify a high level of behavioural cooperation. The respondent can choose a correct answer by chance (with the probability of that depending on the number of answer choices available).

The rating style (RS, response style) is how a respondent uses the rating scale across many questions with the same rating scale. It can be represented by the standard deviation or variance of the responses. **The low differentiation rating style** is understood as a value of the answers of a standard deviation of a particular respondent's answers on a set of questions that is low or equal to zero. The rating style can also be viewed from a single question standpoint, not the respondent, but this will be specified for particular analyses.

Non-informative answers (**DK** answers, Don't Know, 'empty' answers, noninformative answers) do not convey any information about the opinion/thinking/facts regarding what the question asks them to give. It may be considered as an 'in the middle' answer if it is worded in a way that allows such an interpretation (i.e., 'Hard to say'), following the assumption that the respondent could choose not to answer at all and decided to choose that answer, but only when an explicit 'in the middle' option is not available.

Declarative cooperation level refers to answers given by the respondent on questions that were asked explicitly about his performance, i.e., how engaged they were and how tiring the task of answering the survey was for them, or would their answers change if it was a different day.

Logical consistency has two meanings: (1) choosing answers that do not contradict each other in logically related questions (that is, respondents respond 'I do not have a job currently' in one question but respond 'I like my job' instead of 'not applicable' later in the survey) or (2) choosing similar answers to questions that should correlate, either positively or negatively.

Odd answers to open-ended questions to open-ended questions mean answers that are too short or cannot be interpreted concerning the question content (no answer is a separate category, and its usefulness depends on whether the answer was required or not) – applicable only when there were any open questions in the survey.

1.2 Literature review

1.2.1 Advantages and disadvantages of internet research

An increasing number of pooling companies and researchers have realized that online surveys are less expensive than physically approaching respondents or even calling them.

The most popular technique of doing research online is a self-administered survey[11]. Internet is also suitable for experimental research, provides the ability to collect information other than survey data, and enables the possibility of integrating different data analysis methods (qualitative and quantitative).

Types of internet research can be distinguished based on several criteria:

- participants' awareness that they take part in research – they actively decided to participate vs they have been subjected to an experiment without their knowledge (Cambridge Analytica case)
- time of the research – real-time vs anytime
- level of participant's required engagement – active vs passive
- knowledge about participant's identity – anonymous vs identified

Strengths of internet research:

- availability of respondents
- easiness/fastness of reaching specific groups, hard-to-reach individuals
- saves time
- enables tracking of responding process
- cheaper (no need to hire interviewers)
- the research is done at home, without the need to leave to the research facilities
- ability to show more information to the subjects

Weaknesses, possible problems (and proposed solutions):

1. sample selection
    a. inability to conduct representative studies (not needed in experimental and qualitative research) – population using the

[11] Batorski & Olcoń-Kubicka, 2006

internet differs from the general population of a given country (although some countries, like the Netherlands, try to mitigate this by essentially giving their citizens devices to be used to complete surveys required by country's official statistical bureau)

b. impossible to research all internet users (there is no census containing information about all of them)

c. problem with the sample selection itself – does a target sample even use the internet?

    - Possible solutions of this problem: (1) precise determination of the population to be studied, (2) reaching the respondents offline

d. volunteers can be recruited, either through websites or by invitation, but they are not random, which makes it impossible to generalize outcomes, they differ from non-volunteers in a significant way, and there is no easy way to assess what type of people decide to volunteer

    - Possible solutions to this problem: (1) recruit large samples, (2) estimate the number of people who have seen the invitation, (3) increase validity through replication

2. response rate

a. low propensity to participate – due to not having time, no interest or other reasons

- Possible solutions to this problem: (1) offering rewards (financial, gifts, lottery, feedback), (2) invitation form (who organizes the research, way of inviting), (3) choosing the right time to send invitations, (4) specifying a deadline for completion of the study, (5) sending an invitation again.

b. study abandonment by the participant – because of length, lack of interest, problems with the survey itself, slow internet connection.

- Possible solutions to this problem: (1) rewards, (2) information on the progress, (3) saving participants answers as frequently as possible, (4) placing the most critical questions at the beginning, and (5) randomization delay in experiments.

3. online implementation

a. lack of control – over external factors, over participants' devices, over participant's internet skills, over their identity
b. electronic communication – usually anonymous, psychologically distanced, different when it is synchronous vs non-synchronous, different levels of technical skills, less credible, possibility of encountering pranksters

There are also some ongoing problems with internet research ethics, i.e., it rarely considers the respondent's current situation, mainly due to lack of information, so it needs to be assumed that participants are responsible adult people who take responsibility for their actions and decisions and can deal with research content accordingly assumption. Like almost every assumption, this may not be true in every case, especially when the topic of the study is sensitive or triggering. Participation of minors is also problematic, but this rarely poses a problem in management science. There is also a lot of already available data. However, it raises concerns about people's privacy and uncertain content status (can or cannot be used, who owns the rights, who can give permissions), making data anonymization an important issue.

1.2.2 Online respondent panels

The traditional definition of a panel refers to 'a longitudinal study in which the same information is collected from the same individuals at different points in time'[12]. Online panels are not exactly an online version of the traditional panel, but they can be used in such a way. They are 'a pool of registered people who have agreed to occasionally take part in web-based studies'[13], which means they may be a part of longitudinal studies but can also be used in one-time studies[14].

Most internet panels, whether nationwide or not, are not probability-based. Some respondents participate in more than one panel[15], which means that using more than one panel at a time will not necessarily provide more diverse samples, and there is a risk of duplicate data (for anonymous surveys, this is almost impossible to detect). There is also

[12] Göritz, 2010
[13] Göritz, 2010
[14] Göritz et al., 2002
[15] Tourangeau et al., 2013

a risk of the same respondent registering more than once in the same panel[16]. In less technically advanced respondent cases, current software for online surveying allows the detection of duplicate data (data coming from the same device and IP address). This can lead to rare FALSE positives – when two respondents use the same device but are two different people – so it is essential to analyse other characteristics that may indicate an actual duplicate, relying only on software seems to be a risky path.

There is also a problem of 'professional respondents' – respondents who decide to 'opt in' to take many surveys frequently, in exchange for money or some other form of compensation[17], but their motivation to participate is not always on the same level as their motivation to pay attention when answering. Frequent respondents are different from non-frequent respondents. However, research results and conclusions are conflicted as to how they differ. Some studies conclude that experienced respondents use timesaving strategies more often than non-frequent respondents[18], more often choose non-informative (DK) answers[19], are more likely to try to avoid follow-up questions[20] - generally more likely to behave in an undesired way. On the other hand, other studies indicate that frequent respondents are less likely to show these behaviours[21] (or at least there is no evidence they are more likely to do so). There is no clear consensus on how to deal with frequent respondents, but there is a concern about the quality of the data they provide[22].

Online panels are also recommended to be avoided when the researcher wants to estimate values for the population accurately[23]. Panels should also be chosen carefully because they differ in the number of available respondents and practices when verifying respondents[24]. These differences may prove the problem if the researcher aims to examine a particular and hard to reach sample (i.e., older respondents) – it will not always be available in every panel in required sample sizes.

There are also many advantages of using an online panel as a source of respondents. They often have their survey software which may be adapted to the requirements of the study.

[16] Göritz, 2010
[17] Baker et al., 2010; Gittelman & Trimarchi, 2009
[18] Toepoel et al., 2008
[19] Garland et al., 2012, as cited in Hillygus et al., 2014
[20] Nancarrow & Cartwright, 2007
[21] Smith & Brown, 2006
[22] Hillygus et al., 2014
[23] Baker et al., 2010
[24] Baker et al., 2010

However, this possibility depends on the closeness of collaboration with the panel's representative assigned to the project and if the software allows for easy to implement modifications (as any custom features may be expensive to design and integrate into the existing structure). Surveys done through online panels also have a higher response rate and faster data collection than surveys done online without using organised respondent sources[25].

1.2.3 Psychological model of answering survey questions

Respondent's behaviour differs depending on the mode of administration of the survey. Some studies show the difference is not significant or there is no difference[26]. In contrast, others show apparent differences: higher non-response rate in an online sample compared to face-to-face interviews[27], higher data quality in online samples[28], faster data collection in online surveys[29], different age groups represented by internet users (younger initial population available for sampling)[30].

A meta-analysis[31] of 2037 surveys from papers published from 1995 to 2008 in Psychology, Management, and Marketing, showed that survey response rates depend on who is asked to be the respondent – response rates are lower for respondents higher in the organizational hierarchy.

The general model of how respondent answers survey questions is presented in **Figure 1** below.

[25] Göritz et al., 2002
[26] Revilla, 2012; Dodou & Winter, 2014
[27] Christensen et al., 2014; Shin et al., 2012
[28] Shin et al., 2012
[29] Kwak & Radler, 2002
[30] Gigliotti & Dietsch, 2014
[31] Anseel et al., 2010

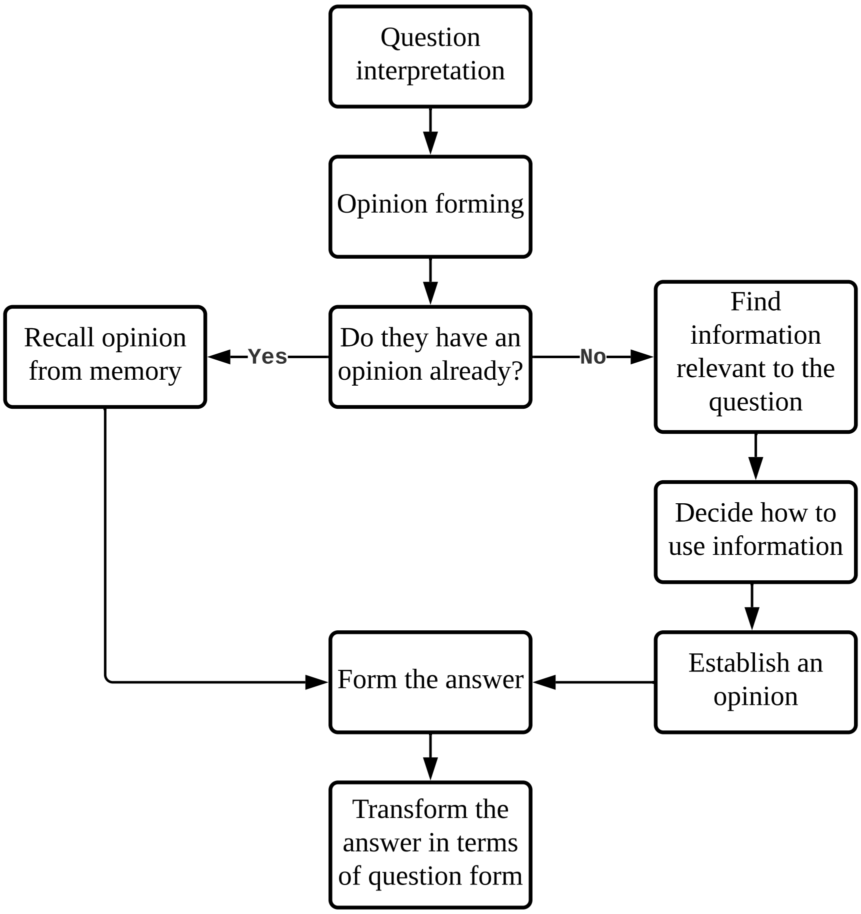

Question interpretation
Opinion forming
Do they have an opinion already?
Yes
Recall opinion from memory
No
Find information relevant to the question
Decide how to use information
Establish an opinion
Form the answer
Transform the answer in terms of question form

Figure 1 *Question answering process*
Source: adapted from Sułek, 2002

The four strategies of answering survey questions[32] are:

1. reproducing already formed assessments
2. motivated processing
3. simplified processing
4. analytic processing.

Reproducing an already formed assessment means that a respondent tries to find their opinion in their memory if they already had an opinion previous to being asked about this opinion. This strategy also applies to questions about facts, assuming that the respondents want to give truthful information[33].

Motivated processing is similar to the first strategy. However, it engages more cognitive resources than simply stating already known facts. If a respondent has a strong preference

[32] Forgas & Vargas, 2005; Wieczorkowska & Wierzbiński, 2011
[33] Wieczorkowska & Wierzbiński, 2011

regarding the opinion they know or something they think is, i.e., not a socially desirable opinion, they may intentionally change their answer to the questions so that they do not lose a positive image of themselves[34].

Simplified (heuristic) processing means that the respondent does not have a ready opinion, needs to formulate it 'on the spot' and consider the subject. In the case of simplified processing, they look for the first association that comes to their mind while thinking about the question content and answer accordingly to what came to their mind first, not engaging in further thinking[35].

Analytical processing is the most complicated and resource-consuming type of answering strategy, often used when the question is not usual, complicated, or the respondent wants to give an as accurate opinion as possible. In this case, the respondent tries to analyse and consider all relevant information about what the question asks and then forms their opinion[36].

When a respondent is answering survey questions, he can choose a strategy of answering, based on:

- their cognitive abilities
- their motivation
- time state of mind (fatigue, mood)
- difficulty/complexity of the questions[37]
    - difficulty of interpretation
    - the difficulty of recalling from the memory
    - the difficulty of the assessment task
    - the difficulty of the rating scale
- length of the questionnaire.

Respondent's ability means their cognitive sophistication, amount of practice in thinking about that particular question[38], and having (or not) opinion on a given topic before

[34] Wieczorkowska & Wierzbiński, 2011
[35] Wieczorkowska & Wierzbiński, 2011
[36] Wieczorkowska & Wierzbiński, 2011
[37] Krosnick, 1991
[38] Fiske & Kinder, 1981

someone asked a question about it[39] because easily accessible information is more accessible to call[40].

Motivation is needed to interpret a question carefully, search their memory and find relevant information, then process that information to form the answer and express that answer in a clear and precise manner as possible, the respondent must be adequately motivated[41]. Motivation is influenced by the degree to which the respondents feel the need for cognition, respondent's interest in the research topic, perceived importance of the survey, accountability, how long the survey is available to be completed, how long it takes to complete the survey.

Question difficulty relates to respondents' ability – questions easy for some respondents may prove difficult for others, even with additional information and instructions. Generally, longer questions, with more caveats and conditions, without reference points or any indication about correct interpretation, tend to be more difficult than short and straightforward questions containing easy-to-understand reference points. Formulation of rating scale provided for the question also influences answering style[42], changing the reference respondent uses to make a choice, and respondents also have difficulties when a question is quantitative rather than qualitative[43] (retrospective questions about exact numbers pose problems, so respondents prefer ranges rather than entering an exact value).

Use of the rating scales provided by the researcher

It is often the case that the person asking questions in a survey provides the scales used for evaluation. Therefore, we are automatically asked to transform our scale (e.g., two-valued: I like or do not like the object of assessment or more sophisticated 10-point scale) into a given rating scale[44]. This transformation produces differences in the way respondent uses the rating scale. All those using Likert-type scales differ in rating style (response style, evaluation style)[45] , which manifests in, e.g., tendency to use only specific points (answer options) of the scale. The respondent's rating style can only be defined

[39] Wieczorkowska-Wierzbińska, 2011
[40] Fazio, 1986
[41] Tourangeau & Rasinski, 1988; Krosnick & Presser, 2010
[42] Moxey & Sanford, 1986
[43] Moxey & Sanford, 2000
[44] Wieczorkowska, 1993, as cited in Wieczorkowska-Wierzbińska, 2021
[45] Wieczorkowska, 1993, as cited in Wieczorkowska-Wierzbińska, 2021; Hoyt, 2000

when the respondent assesses multiple items. Then it can be operationalized by a measure of central tendency and variation of the rating distribution.

Many studies have shown differences in evaluators' rating styles[46]. Systematic individual differences in the leniency of evaluators have also been shown[47], which means there are differences in information processing, personality traits and situational conditions[48]. The style of evaluating can be described in terms of leniency (some give, on average, higher ratings than others), differentiation of ratings (low when evaluated objects assessed similarly)[49], and using extreme ends of the scale (avoiding or using only extremes)[50].

1.2.4 FALSE respondent model

FALSE responding[51] has been called in literature in many different ways – as random responding[52], insufficient effort responding[53], careless responding[54], satisficing[55], inattentive responding/participant inattention[56], indiscriminate responding[57]. It can be defined broadly as happening when the respondent filling a survey does not behave cooperatively. The decision about their chosen way of answering may be

- intentional, and thus suggesting malicious uncooperative behaviour - they have the resources needed, just choose not to use them or use them to lie deliberately, or
- not intentional – because they lack resources, understanding, motivation, the survey itself is poorly designed, and many other reasons that are not the respondent's fault.

When the respondent does not behave cooperatively, they usually choose one of the satisfactory strategies[58]:

[46] see the summary in Wieczorkowska, Kowalczyk, 2021
[47] Dewberry et al., 2013
[48] Dewberry et al., 2013; Judge & Ferris, 1993, as cited in Wieczorkowska-Wierzbińska, 2021
[49] Król & Kowalczyk, 2014
[50] Clarke III, 2000, as cited in Wieczorkowska-Wierzbińska, 2021
[51] Levi et al., 2021
[52] Credé, 2010
[53] Huang et al., 2012; Huang & DeSimone, 2021
[54] Meade & Craig, 2012; Bowling et al., 2020
[55] Krosnick, 1991
[56] McKibben & Silvia, 2017; Beck et al., 2019; Steedle et al., 2019
[57] Holden et al., 2019
[58] Krosnick, 1991

1. **Selecting the first response alternative that seems reasonable**[59] – they do not read all answers and choose first that suits them; answers to such studies can be influenced by order of response choices because respondents tend to spend more time looking at the first few response options.
2. **Selecting option easier to see**[60] – in web surveys, we can choose more than one display of answers; comparing dropdown vs radio buttons vs scrollable dropdown, it turned out that used response format can affect choices more than primacy.
3. **Speeding** – occurs when the respondent answers so quickly that there is little to no time for thinking about the answers[61] (an example in the doctoral dissertation from the Faculty of Management was about customer satisfaction of students who were assessing their lecturers[62]).
4. **Agreeing with assertions (confirmation bias)**[63] – in agree/disagree, true/false and yes/no questions, there is a tendency to accept any given assertion, regardless of its content.
5. **Endorsing the status quo**[64] – when a question asks about increasing or decreasing something, respondents often choose base (starting) value when explicitly given to them.
6. **Non-differentiation in using rating scales**[65] – when using the same response options, in the same order, there is a danger that respondents will not differentiate between objects. Consequently, respondents will choose the same or almost the same options in each question.
7. **Answering ‘do not know’** - as ‘do not know’ is hard to interpret, but also does not require much thinking, when that answer is presented, satisficing respondents will choose to pretend they do not have an opinion rather than trying to put effort into creating one, although research shows, that providing this answer option increases data quality[66].

[59] Galesic et al., 2008
[60] Couper et al., 2004
[61] Conrad, et al., 2017
[62] Michałowicz, 2016
[63] Krosnick, 1991
[64] Schuman & Pressner, 1981
[65] i.e. Krosnick & Alwin, 1989
[66] Albaum et al., 2011

8. **Mental coin-flipping**[67] – choosing randomly from among the response alternatives.
9. **Omitting a whole set of questions**, either by losing one's attention or by purpose, does not mean that answers are worthless, but there are difficulties with determining what to do with them – include or not.

Respondents have different reasons to engage in the above strategies. As mentioned before, the reasons may lie in their cognitive/mental abilities, motivation, fatigue, and mood, which are all respondent-dependent. However, they may also come from respondent-independent factors, like the length of the questionnaire, the difficulty of questions, unexpected interruptions and distractions.

More complex survey formats cause respondents with lower ability to satisfice more[68], as they likely require more cognitive abilities than simple survey formats. The more difficult (complex) question, at all possible points, also the higher the chance of satisficing[69] - questions presented in a grid format tend to invoke more satisficing than questions presented separately[70]. Overall, in terms of cognitive load, the more demanding the survey, the higher the chance of a respondent becoming a FALSE respondent[71].

Motivation can be intrinsic or material – for the first type, the respondent is interested in the topic of the survey or has other reasons specific for themselves, which naturally makes them more attentive. However, respondents taking a survey for material reasons may speed through the survey to collect their reward at the end without exerting much effort[72]. Less motivated respondents give less reliable responses[73], but the causes of low motivation are inevitably different for different samples.

[67] Converse, 1964
[68] Roßmann et al., 2018
[69] Krosnick, 1991
[70] Roßmann et al., 2018
[71] Merritt, 2012
[72] Berinsky et al., 2014
[73] Bassett et al., 2017

Participants' fatigue while negatively worded items are present causes the unidimensional construct to become bidimensional[74]. Inattentive responding is also connected with dysphoric mood[75] and anxiety disorder, or anxiety combined with depression[76].

Survey length is thought to have only one way of influencing attentiveness – the longer the questionnaire, the less attentive respondents become[77], mainly because longer questionnaires are more demanding in terms of cognitive effort required[78]. Survey length also negatively affects completion rates[79]. Distractions that can often happen during multitasking (a common occurrence in respondents from crowdsourcing sites) can lead to false responses[80].

Several respondents' sociodemographic characteristics correlate with a higher probability of being a FALSE respondent – male gender[81], lower education level[82], rural place of residence[83], being younger[84]. There is also evidence of personality traits, either self-reported or reported by acquaintances, being correlated with FALSE responding, although correlations differ depending on what detection of FALSE respondents method (or methods) were used.

- personality traits - lower levels of conscientiousness, agreeableness, extraversion and emotional stability (acquiescent-reported)[85], lower levels of benevolent traits and higher levels of malevolent traits (self-reported)[86] (correlations depend on the indices of FALSE responding)
- (self-reported) external (extrinsic), rather than internal (intrinsic), motivation[87].

Some respondents may attempt to take a survey more than once[88], which may mean two things: the first attempt was the only one valid, and subsequent attempts should be

[74] Merritt, 2012
[75] Murphy et al., 2013
[76] Conijn, 2020
[77] Bowling et al., 2020; Gibson & Bowling, 2020
[78] Eisele et al., 2020
[79] Galesic & Bosnjak, 2009
[80] Necka et al., 2016
[81] Roivainen et al., 2016; Maniaci & Rogge, 2014
[82] Verbree et al., 2020; Lu et al., 2019
[83] Lu et al., 2019
[84] Maniaci & Rogge, 2014
[85] Bowling et al., 2016
[86] McKay et al., 2018
[87] Maniaci & Rogge, 2014
[88] Johnson, 2005

discarded, or the first attempt was made carelessly to quickly look through contents of the survey, and the second is actually attentive. It can be distinguished by looking at answering time – shorter time in the first attempt means inattentive responding. In the first case, the first attempt can be ‘saved’ and left for further analyses, so only subsequent duplicate attempts must be discarded. In the second case, as the respondent will inevitably respond quicker to questions they have already seen before, answering time cannot be used to assess attentiveness, so all attempts by the same respondent should be discarded.

1.2.5 Impact of FALSE responding on data quality

Extensive research exits on the ways that FALSE responding influences data quality – from minor, insignificant noise introduction to entirely changing the outcome of statistical analyses and, in consequence, conclusions drawn from the data.

Over a decade ago, a problem with replicating previous research became more visible[89], and this is especially a problem in social sciences – an attempt to replicate 100 classic studies in psychology resulted in replicating results in only 39 of them[90]. It seems to be a problem also in management[91]. Whether it may be actually caused by increasing numbers of FALSE respondents – is still a question that remains unanswered, but this may be a possible explanation for the problem of failed replications.

Another problem with the influence of FALSE respondents are false positives – FALSE responding may inflate correlations between variables[92], although it is usually thought to attenuate these relationships rather than inflate them[93].

In studies that use experimental manipulation, respondents identified as FALSE tend to not respond to manipulation in text content[94], making an assessment of manipulation effect virtually impossible, as it cannot be determined if the manipulation was ineffective or if a respondent simply did not read contents of question containing this manipulation. Excluding FALSE respondents may also increase statistical power, regardless of initial sample size[95] and increase effect sizes[96].

FALSE responding can also cause lower internal consistency of validated scales[97]. If overlooked as a cause of low reliability of the scale, analysing data containing observations coming from FALSE respondents can lead to false conclusions that the scale is no longer adequate and appropriate to measure a construct it was designed to measure.

[89] Ioannidis, 2005
[90] Open Science Collaboration, 2015
[91] Hensel, 2021
[92] Huang et al., 2015b
[93] McGrath et al., 2010
[94] Maniaci & Rogge, 2014
[95] Maniaci & Rogge, 2014
[96] Brühlmann et al., 2020
[97] Huang et al., 2012

Problems in questionnaire development and item analysis are also an important consequence of inattentive responding[98] - the usefulness of questionnaires can not be assessed based on meaningless data, and so can not be the usefulness of particular items – excluding or including them in the scales based on false data makes the scale itself less valuable. This problem is connected with analysing questionnaire dimensionality – the survey on 666 employees from different organizations had shown that the seemingly separate constructs of job satisfaction and dissatisfaction – their correlation was nearly -1 when FALSE respondents were excluded – actually form the same, unidimensional, construct of job satisfaction[99] (this study's more detailed description also appears later in the dissertation).

1.2.6 The magnitude of the problem

Percent of FALSE (careless) respondents varies from study to study, from around 4%[100], through around 10%[101] to about 20%[102]. Most careless respondents look like they are intermittent, and self-reported low level of cooperation can be as high as 50%[103].

As the studies are not consistent in their inattentiveness indicators' use, short summaries of the examples of studies on inattentive (careless) respondents are presented in **Table 2** (to demonstrate how different approaches can be).

[98] Johnson, 2005

[99] Kam and Meyer, 2015

[100] Johnson, 2005

[101] Kurtz & Parish, 2001; Meade & Craig, 2012

[102] Curran et al., 2010

[103] Baer et al., 1997

Year	Criterion	Methods of detection used in the study	Sample	N	Excluded respondents
2009[104]	Failed attention check question	Attention check question (IMC)	Students	144	35% on first try
2015[105]	Respondents were flagged as inattentive by at least one indicator	Self-reported low level of cooperation, Attention check question	MTurk	400	5.5%
2016[106]	No exact cut-off points	Respondent's Goodness of Fit	Purposive sample	205	10.81%[107]
2016[108]	Failed attention check question	Attention check question (IMC)	MTurk	396	5%
			Students	85	61%
		Attention check question (Novel IMC – long instruction with the hidden correct answer)	MTurk	185	4%
			Students	245	74%
		Attention check question (more difficult novel IMC – short instruction to mark two answers)	MTurk	239	74.5%
			Students	90	97.8%
2020[109]	Respondents were flagged as inattentive by at least one indicator	OAT, IRV, psychometric synonyms, odd-even consistency	Students	278	12.8%
				281	12.5%
				268	15.7%
2017[110]	Faster than 1 spi, consistency measure lower than 0.5	Answering time, Response consistency (correlations between related items)	MTurk	421	between 5 and 24%
	Faster than 1 spi, consistency measure lower than 0.43		Students	296	12%
2018[111]	10% of the sample on each measure	Attention check questions (infrequency type), Answering time, Even-Odd Consistency Index, Long String Index, Intra-Individual Response Variability	Students	199	30.15%

Year	Criterion	Methods of detection used in the study	Sample	N	Excluded respondents
2018[112]	Respondents flagged as inattentive for each measure separately	Mahalanobis distance, Psychometric synonyms, Psychometrics antonyms, Maximum LongString, Answering Time, Self-reported low level of cooperation, Self-reported interest, Attention check question (instructed response), Self-reported Single Item	Students	274	On average, 5.91% per method
		The same as 1st study + Even-Odd consistency		614	On average, 2,88% per method
				394	On average, 4.33% per method
2019[113]	Respondents were flagged as inattentive by at least one indicator	Contradicting answers to reversed items, Answering time	Students (online)	129	23%
			Students (paper and pencil)	101	27%
			MTurk	110	46%
2019[114]	Failed both attention check questions	Attention check questions (instructed response), Answering Time, Straight lining, Item nonresponse	German Longitudinal Election Study (panel)	5205	6.1%
2020[115]	Based on Latent Profile (Class) Analysis	Open-ended questions, Resampled Individual Reliability, Person-Total Correlation, Self-reported low level of cooperation, Attention check questions, Answering time, Long-String index, Odd-Even Consistency	Crowdsourced, FigureEight	394	45.9%

***Table 2** Examples of different exclusion rates*

[104] Oppenheimer et al., 2009
[105] Rouse, 2015
[106] Kountur, 2016
[107] exclusion by design of the research – group of 20 respondents was instructed to behave inattentive when responding
[108] Hauser & Schwarz, 2016
[109] Iaconelli & Wolters, 2020
[110] Wood et al., 2017
[111] Dunn et al., 2018
[112] Ward & Meade, 2018
[113] Aruguete et al., 2019
[114] Silber et al., 2019
[115] Brühlmann et al., 2020

1.2.7 Classifications of methods of detection inattentive responding

FALSE responding detecting methods can be classified by looking at the problem they are trying to solve from different points of view. Some of the most common classifications have been shown and described below.

Objective and subjective methods.

The objective category includes:

1) reaction times
2) eye movements and other physiological measures,
3) logical inconsistency
4) rating style.

In the subjective category:

1) attention check questions
2) non-informative answers
3) odd answers to open-ended questions
4) response latitude
5) respondent's goodness of fit
6) comparing with random data
7) accidental finding

Some of the objective methods are precise measurement and, in some cases, expensive equipment. Based on previous studies' findings, a link exists between eye movements and cognitive processes during normal reading, which is not present when mindless reading[116]. Specialized equipment can be used to track eye movements. However, there is also a possibility that standard webcams can generate the data having quality sufficient to analysis. For now, they have higher variance than eye trackers designed to do just that[117].

[116] Reichle et al., 2010
[117] Sammelman & Weigelt, 2018

Statistical and non-statistical methods.

The scientific community has already developed several methods of detecting those behaviours, and these methods can be divided into two seemingly broad categories: statistical and non-statistical ones. Scientists protect themselves against FALSE respondents by placing instructional attention check questions in polls (i.e. such as ‘Here check this answer’ and if the instruction is not completed, this respondent is suspected to be FALSE), the purpose of which is to detect whether the respondent reads the questions he or she answers, but those questions do not cover all possible strategies of responding that are considered to be fake. Therefore, a complete procedure that considers the highest possible number of FALSE responding strategies would be a valuable tool for improving the quality of data.

Statistical methods of finding FALSE respondents include:

1. analysis of answering times[118]
2. analysis of logical inconsistency (response consistency)[119]
3. analysis of rating style[120]
4. comparing with responses randomly generated by a computer[121]
5. measuring response latitude (when Likert scales were used)[122]
6. respondent’s goodness of fit analysis (RGF)[123].

Non-statistical methods for finding FALSE respondents include:

1. serendipitous finding[124]
2. self-reported low level of cooperation[125]
3. attention check questions[126].

The possibility of using statistical methods strongly depends on the quality of the data analysed[127], which means that the precautions for the detection of FALSE respondents

[118] Huang et al., 2012; Wood et al., 2017; Meade & Craig, 2012; Curran, 2016
[119] Merritt, 2012; Meyer et al., 2013; Weijters et al. 2013; Kam, Meyer, 2015
[120] Weathers & Bardakci, 2015
[121] Fronczyk, 2014; Dunn et al., 2018
[122] Lake et al., 2013
[123] Kountur, 2016
[124] Piferi & Jobe, 2003
[125] Merckelbach et al., 2010
[126] Conrad et al., 2017
[127] Meade& Craig, 2012; DeSimone & Harms, 2018

should be taken before the data is collected. There is evidence from meta-studies (based on a small sample to be clear) that biased responses are the source of error variance and that careless responding can underlay bias responses[128]. Also, insufficient effort responding is responsible for many other data distortions[129].

Analysis of an answering time variance seems to be especially interesting in this matter, as this can be gathered without respondents intention to tell the truth or not – the truth is always there.

Direct and indirect measures

Classification of directness and indirectness of detection methods comes from a recent study[130]. Direct measures can be described as having been placed in the survey in advance to explicitly measure respondents' attentiveness, and it is their only goal. Indirect measures rely on looking for unusual patterns in respondents' rating style, and they do not give any indication to the respondents that they are being 'checked' or watched by the researcher.

Direct measures:

- attention check questions
- self-reported low level of cooperation

Indirect measures:

- analysis of rating style
    - intra-individual response variability
    - long-string
- analysis of answering times
- analysis of logical compatibility

[128] McGrath et al., 2010
[129] Huang et al., 2015a
[130] Goldammer et al., 2020

1.2.8 FALSE respondents detection methods

Several technical issues may occur when dealing with designing a web survey, including the need to use ‘cookies,’ measuring answering time, avoiding missing data, randomizing items, and maximizing response rate[131]. Those are not the only problems that we have to face, although some are easier to handle than others. It is impossible to gather the data necessary to assess how respondents answered the survey questions in many cases. The reasons for that vary, a researcher may not have technical knowledge, not anticipating that the data may be useful, more sophisticated software solutions being too expensive, etc., but usually, lack of planning at the stage of survey design is the main culprit of a researcher being very restricted in what can be done with the data later.

Most of the methods of detecting careless respondents developed by researchers up to date require planning before the data collection stage. However, some can be used on data that has already been collected.

To make following the reasoning easier, methods of detection have been divided into two main groups: those, who are the part of 4 warning signs described in this dissertation, which will be stated and described first, and other methods, which were considered either non-applicable to the area of management science, or simply too complicated to be effectively used without spending excessive time on analysis and learning how to use them. The four warning signs will be described by at least one study example per sign. Other divisions of the methods will also shortly be described, along with study examples.

Four warning signs consist of one or more parts and are organised to make analysis and to follow the procedure easier. Not all of them must be used every time and in every study – decision about what can and should be used in a particular case needs to be made by the researcher designing methodology for the study.

The contents of each sign emerged during repeated analyses and reflected their weight in the overall exclusion rate.

[131] Eysenbach & Wyatt, 2002

WARNING SIGN #1 Too short answering time

There are a few interesting facts about response times collected from surveys. They can be used to detect 'bad questions' because questions with problems have longer answering times[132]. There is also a possibility to use them, in connection with other characteristics of a question (average word length, number of letters, number of words, number of possible response options, length of response options, etc.) to determine if a respondent paid attention when reading and answering, or not. That requires an approach to the problem of detecting FALSE respondents on many levels[133].

Warning Sign #1 (WS1) can consist of up to 5 parts: OAT including non-compulsory elements (like planned breaks etc.), OAT excluding non-compulsory elements, PAT for every main part of the study, PAT for every series of similar questions, PAT for every single question. Exclusion criteria usually depend on the contents of the study and the analyses that need to be conducted – as respondents' attention and patience usually decline toward the end of a survey and on parts that are not interesting for them, it may be helpful to exclude some of them only locally.

Usually, allowing a respondent to be flagged as 'suspect' by one of the parts and still not be excluded allows avoiding excluding overall attentive respondents who just happened to have a higher thought process than the average.

The answering time can be calculated at different levels of detail, depending on the measure of detail in the analysed data. Thresholds set for levels of detail can either be chosen arbitrarily (not recommended) or chosen based on some other characteristics of the study and/or question. It can, for example, depend on the features shown in **Figure 2** (below).

[132] Bassili, 1996

[133] Yan & Tourangeau, 2008

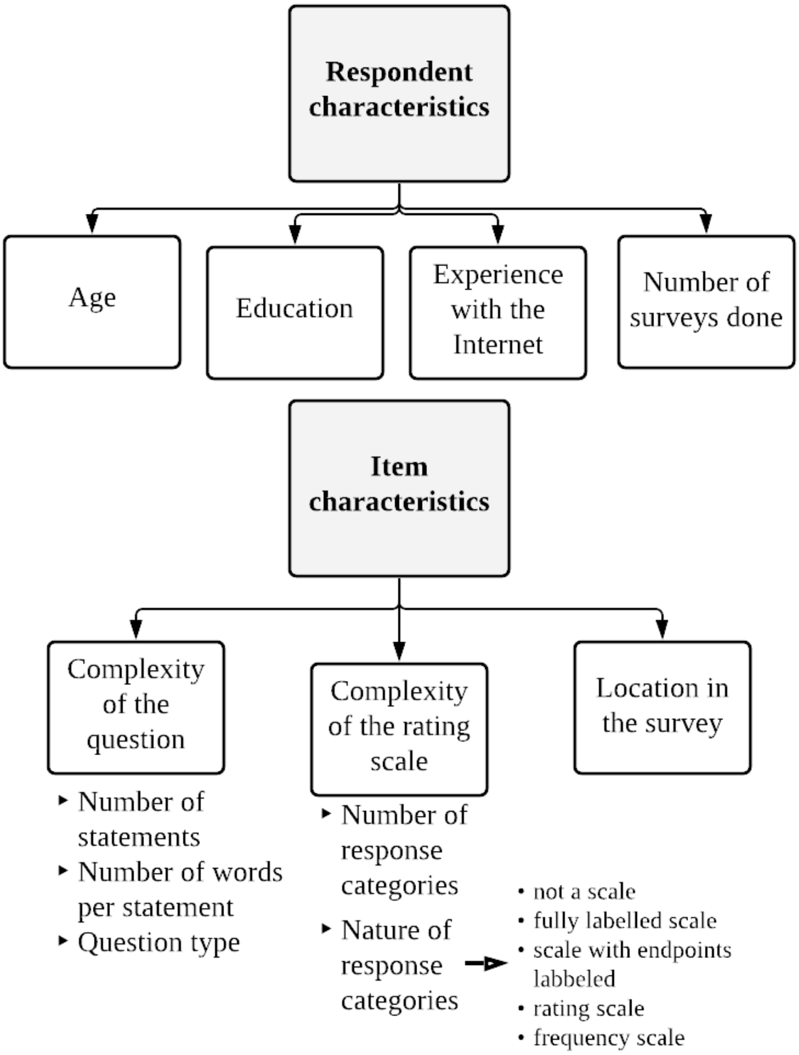

Respondent characteristics
Age
Education
Experience with the Internet
Number of surveys done
Item characteristics
Complexity of the question
Complexity of the rating scale
Location in the survey
Number of statements
Number of words per statement
Question type
Number of response categories
Nature of response categories
not a scale
fully labelled scale
scale with endpoints labbeled
rating scale
frequency scale

Figure 2 *Item and respondent characteristics influencing answering time, modified from: Yan & Tourangeau, 2008, p. 56*

Though many more factors influence respondents' behaviour, most of them are out of the researcher's control.

The use of reading speed as an indicator of respondents' attention is widely accepted by the research community[134], and it is well researched and described way of identification.

According to one of the most cited studies on reading speed[135], the typical reading speed for English is about 200-300 wpm for adults without any impairments, with a comprehension on an acceptable level, having an upper limit of 600wpm for most

[134] Conrad et al., 2017; Zhang & Conrad, 2014

[135] Carver, 1992

proficient readers. Polish is more complex and difficult to learn language, and in a study comparing the reading speed of native Polish readers and learners from other countries, findings of the study on the English language have been somehow confirmed – reading speed for the Polish language in native readers learning languages was between 160 and 240 wpm[136]. As the sample was very small (only 16 people) and consisted of students of language learning course, the actual reading speed for non-learners may be lower.

Estimating minimal reasonable duration time, either using the lowest value from a pilot study done on known and engaged participants or assuming a threshold based on reading speed. The most conservative threshold is 600wpm, but it can be achieved only by people trained in speed reading, and comprehension rates drop significantly for ordinary people at half this value, so it is safe to assume that reading speeds higher than 300wpm will rarely exclude respondents reading the questions. However, research shows that students and young people generally can read faster (about 250wpm), so it is reasonable to analyse diverse samples as separate age groups.

OAT including non-compulsory elements

The time it took a respondent from opening the survey to completing it. More objective if connected with a total number of words in the survey and reading speed threshold instead of assumed time threshold.

OAT excluding non-compulsory elements

If the survey allows respondents to take breaks (they can take a break regardless, and there is almost no way of controlling this), has optional questions, or has any elements that do not require attention for the attempt to be considered valid and completed observation, computing actual responding time requires excluding these elements from total time and total words count (if the wpm measure is used).

PAT for every main part

Almost every survey has a few main parts, usually organised by their topic. If a respondent is interested in one part of the survey and not in the other, it can be detected on the main part level by computing PAT (without non-compulsory elements) for each

[136] Moździerz, 2019

distinct part. This allows for excluding respondents from analyses done on specific main parts.

PAT for every block/series

Blocks/series of questions are different from the main parts, in a way. They usually have a similar question format, similar question answer options but do not necessarily pertain to the same topic. It is the main difference between parts – a part can consist of questions that look and behave differently. However, a series consists of questions that look and behave similarly. Respondents may still be attentive during answering questions that differ (as the change itself can be attractive to them) but inattentive when they answer longer strings of similarly looking items.

PAT for every question/page

The most detailed part of the sign allows for analysis on a single-item level (columns), besides the analysis on a respondent level (rows). Like the previous parts, it is more beneficial if reading speed is considered in the analysis.

An example of a study that used PAT for every question is studied[137] on data from German Longitudinal Election Study, from a nonprobability online panel, using quota sampling. It consists of four data sets, each one containing about 1100 respondents. The survey used was designed to be completed within a timeframe of 30 minutes.

Procedures for page-specific analysis used in this study were as follows. First, the authors analysed page specific answering time, and after estimating the page's median answering time, the authors compared specific respondents' time to the median time for the page. Respondents' times and page-specific times were analysed in the division to age groups to avoid flagging young, highly educated people as too fast. To avoid using an arbitrarily set threshold, the authors used three different thresholds: 30%, 40%, and 50% less time needed to answer for a given respondent on a given page, compared to that group's median for that page. To explain this method, let us assume that page median answering time for somebody's age and education group was about 20 seconds, and their answering time for that page was about 13 seconds; if 30% threshold were applied, they would be

[137] Greszki et al., 2015

flagged as too fast because their answering time is 35% below the median for this split group. However, they would not be flagged if a threshold of 40% was applied.

For respondent-specific analysis, the authors calculated the time spent by each respondent on each page. The respondent-specific sum of below-median page values (taken from page-specific analysis) was divided by the number of survey pages below the median (indicator taken from prior research – Rossmann, 2010[138]). In this part of the analysis also three thresholds were set – 30%, 40% and 50% below the median time. For 'No answer'/' Do not know' answers, the threshold of 450wpm was set – meaning that the respondents above this threshold were flagged as too quick.

The results of this study: 31.1% to 100% of 'No answer' or 'Do not know' answers were given too quickly. This was calculated by taking the percent of a number of pages for a given respondent that was either 30, 40, or 50% below the median for every given page median. The authors used two approaches (page-specific and respondent-specific) to avoid false positives – respondents flagged as too fast who were not too fast (that is why there is a need to look from different perspectives). There is also evidence (based on those indices) that some respondents are faster in some parts but much slower in other parts of the survey. Page-specific and case-specific measures grasp two different, although related, phenomena – flagging different respondents as speeders.

The conclusion from this study: for simple models, irrespective of which technique or criterion was employed, correcting for speeding did not change marginal distributions of the variables included in the analysis. Speeders seem just to add random noise to data. For multivariate models, removing speeders can increase or decrease coefficients, but by 1 to 1.3 standard error.

In another study, the authors tried to test the impact of warnings on speeding[139] tried to reduce speeding by giving their respondents immediate feedback about their behaviour. A warning was triggered for reading speed faster than 350ms/word (about 171wpm). Respondents in this study were assigned to two groups: E1: group with no warning or E2: a group who received a warning when answering faster than a set threshold of

[138] Rossmann, Joss. 2010. Data Quality in Web Surveys o f the German Longitudinal Election Study 2009. Paper presented at the 3rd ECPR Graduate Conference, Dublin, Ireland.

[139] Conrad et al., 2017

350ms/word, and those who did not trigger any warnings during responding were treated equally with those in control (E1) group. Respondents in the experimental group, who sped at least once (and triggered a warning), sped subsequently on a fewer number of questions than respondents in a control group (reduction by about .2 to .6 questions). If they triggered more warnings than one, it did not change anything to increase or decrease the number of subsequent questions they sped on, which means that a single warning is as adequate as many warnings. However, it has more of an impact if shown earlier in the survey.

In this study, numeracy questions were used to test attention (accuracy), and six correct (out of 7) were used as the threshold. However, warnings did not change the accuracy of answering these questions, which may mean that respondents did not have a problem with attention, but with task difficulty.
The authors also presented results indicating that speeding correlates positively with straight-lining (choosing the same answer in many subsequent questions).

The authors of the study described the above-used warnings triggered only when the respondent was too fast. This inspired one of the research tasks for this dissertation: checking if forcing respondents to choose a correct answer to attention check questions will make them more attentive than respondents who can choose any answer without warning that the answer is incorrect.

WARNING SIGN #2 Errors in attention check questions

It has proven to be a useful measure of respondents' attentiveness. It relies on placing one or a few questions designed to check respondents' attentiveness during answering. It is usually done as a simple instruction to choose a particular answer of the options provided or in the form of instructional manipulation check, but it may also take on different forms, like arithmetic questions.

Examples of attention check question types:

- Type 1 – Attention check questions

Attention check questions are questions for which the only purpose of their presence in the survey is to check if respondent reads questions. An example of such a question is an arithmetic question: *What is the outcome of this operation? 13-3=?* There is only one correct answer, and it is 10, but the question itself does not have any connection with the topic of the study. It does not have the same rating scale as other survey questions, which makes it less ‘tricky’ for the respondents.

- Type 2 – Instructional manipulation checks

This type aims to check if the respondent reacted to the contents of the study in an expected way, i.e., if he read the instructions, reacted attentively to information presented; questions about facts. Example: *You read the description of a place on the previous page. What was this place?* Possible answers: A church, A school, A parking lot. Only one of them is correct, and it should be an obviously correct answer.

- Type 3 – Instructed response items

Questions with the same rating scale as a whole series of questions in the survey but having a clear instruction about which answer the respondent should choose, i.e., *In this question, choose ‘Strongly agree’ — o*ne of the most widely applied forms of attention check questions, and the easiest one to apply.

- Type 4 – Infrequency scales

Questions are designed in a way that they have only one correct answer (often absurd statements), and all respondents should choose that answer because other answer options are impossible to be true, i.e., *I have never drunk any water*.

Authors of the study on attention check questions[140] state that attention check questions (1) can induce more attentiveness in respondents, but (2) they also often violate the rules of cooperative conversation, which may teach participants that the researcher does not trust them and does not want to cooperate. If the second effect of an attention check question is valid, they hypothesize, its presence should decrease effects (correlations) in other parts of the questionnaire. Data for this study came from Amazon MTurk.

Results of this study show that 6.5% (52) participants did not pass the attention check question in Survey 1, and (1) attention check question did not influence the response order

[140] Hauser et al., 2017

effect; it also did not mitigate respondents' nondifferentiation behaviour (meaning that they responded to questions regardless of its content), did not change choosing non-informative answer frequency, and did not lessen acquiescence (agreeing with everything), (2) attention check order (whether it was presented as one of the last questions of the survey versus presented as one of the first questions) did not influence how much participants responded to question context, scale range effects, or cooperative conversation norms violation, but had slightly moderating effect on comparative judgement for scale range effects – if participants answered attention check question first, the effect disappeared.

Another study on attention check questions[141] was conducted on a sample of 666 full-time employees invited by email and paid for completing the survey. The authors used Job Satisfaction Scale: Illinois Job Satisfaction Index, which was shortened to 2 items per facet (a set of questions about the same latent construct). One of the items was negatively worded (ex. *I hate my job*), and one was positively worded (ex. *My job is OK)*. Their reversed counterparts were also created and added to the item list (ex. *I love my job* and *My job is not OK*), which gives a total of 16 items for the whole shortened scale. Study used attention check questions, long string, repeated responses, Mahalanobis distance and total survey time to measure carelessness. The authors used latent class analysis to estimate respondents' class membership, the study aimed to assess the influence of careless responding in combination with acquiescence bias (agreeing with questions regardless of content) on construct dimensionality (on example of job satisfaction and dissatisfaction). The authors assumed that acquiescent respondents pay more attention than careless respondents.

Bayesian information criterion (BIC)[142] was used to find a fitting model in LCA – a two-class initial solution was tested by the authors (careful respondents and careless respondents), then a three-class solution and four-class solution. In three-class solution, the careful class had about 69% of respondents; first careless class, identified and described by the authors as patterned careless responding, had about 14% of respondents, and was characterized by long-string and repeated answering patterns, the second careless class, described as unpatterned careless responding, had a share of about 17% of

[141] Kam & Meyer, 2015

[142] Meade & Craig, 2012

respondents and was characterized by higher Mahalanobis distance scores (unusual response pattern). After merging to careless classes the total was about 31% of the sample. Study results have shown that (1) correlation between job satisfaction and job dissatisfaction was stronger for careful respondents that for careless respondents, (2) job satisfaction and dissatisfaction constructs became unidimensional after controlling for acquiescence, (3) negative constructs correlated significantly more strongly positively with job dissatisfaction, and positive constructs correlated positively significantly stronger with job satisfaction, (4) correlations, that should be positive, were negative for careless respondents, and (5) deflation of effect sizes was stronger than inflation due to presence of FALSE respondents in data.

Conclusions from this study: (1) careless more likely than careful to give identical answers to many items in a row, (2) there are multiple ways of careless responding, (3) there is an interaction between careless responding and type of survey measures, (4) careless responding goes in combination with acquiescence bias.

WARNING SIGN #3 Too many non-informative answers and low differentiation rating style

This sign consists of two parts (which can have many instances in a single survey): (1) percent of non-informative answers in a single similar questions series – having the same rating scale and the same question format; (2) too low differentiation rating style (or standard deviation).

To many non-informative answers

Usually, the rating scale contains an option that allows 'an escape' – if the respondent cannot decide which answer to choose, they can choose 'It is hard to say' or 'I do not know' or 'I do not want to answer' or other similar answer option (if available). In literature, choosing this option is usually treated as item nonresponse[143] and one of the forms of inattentiveness[144]. There are cases when answering 'Don't know' is a way of expressing genuine attitudes[145], but in many cases, these options may be used as avoiding too much thinking – respondents failing attention check questions more frequently choose

[143] Kuha et al., 2018
[144] Beatty & Herrmann, 2002
[145] Krosnik, 2002

‘Don’t know’[146]. Research[147] has shown that the number of non-informative responses negatively correlates with the self-esteem of the respondent's effort to answer questions, suggesting that avoidance is often due to laziness.

However, if the researcher wants to ‘force’ a respondent to choose something meaningful, such an option may not be available[148]. The respondent is then forced to choose randomly. It is necessary to include non-informative answer options in questions about opinions – where there is a high possibility that the respondent does not have an answer. There is also an option of using filter questions[149] - asking respondents whether they have an opinion/knowledge on the question’s topic, but personality surveys usually use many similarly formatted questions and asking a filter question before every main question is not feasible.

Too many non-informative answers (above 50% at least) usually render computing respondent’s scores on question scales a questionable result. If possible, non-informative answers can be recoded to represent the middle of the scale (if the middle option is not already used by an answer option with other content, like ‘Neither A nor B’, etc.) – which is an approved approach in many cases. This, however, does not change the fact that using non-informative answers too much defies the purpose of personality research – correlating personality traits (used often in human resources research) with other measures when the respondent has basically no personality (if we assume non-informative answers reflect how they really behave) makes no sense, hence using these answer options as a proxy indicator of FALSE responding.

Low differentiation rating style

Requires at least five questions with the same rating scale and length. Questions cannot be worded in the same direction if there is an expected strong positive correlation between them (they are a part of the same factor) – at least one question should be reversely worded for this part to be effective. Cases having variance (or SD) equal to 0 are almost certain to be FALSE respondents. The threshold for acceptable minimal variability depends on the number of questions used to calculate it and the length of the rating scale (i.e., 3-

[146] Gummer et al., 2018
[147] Krosnik et al., 2002
[148] Krosnick 2002
[149] Allen, 2017

option Likert scale will have less variability than the 5-option Likert scale)., but also on the respondents rating style.

In an example study on low differentiation rating style,[150] authors propose Individual Response Variability (IRV) index as a measure of insufficient effort responding (inattentive respoding), as an extension to (LSI) Long String Index (occurs when respondent gives the same response to an unusual number of consecutive items). They propose that IRV should be calculated as a standard deviation of a set of consecutive item responses for a given respondent.

Advantages of this approach are clear – it detects long strings of similar answers within an examined set, it can detect less apparent forms of insufficient effort responding (ex. alternating patterns), this approach is easier to calculate than LongString Index, and it can be calculated for many sets across the survey. The approach has also disadvantages – for the pattern of variation in standard deviation to show itself a certain number of questions is required, but, on the other hand, including too many items reduces the sensitivity of identifying careless responding. The approach is also sensitive to respondent's rating style.

The study was conducted on university students who were given bonus points for taking part in the research. Participants completed short personality measures questionnaire (50 items) and a lengthy questionnaire about two weeks after (325 items) to increase the likelihood of FALSE responses occurring. The authors of this study used several methods of inattentive responding detection:

1) attention check questions – precisely the type called 'infrequency scales', (sometimes) absurd statements, having only one correct answer (i.e., 'I have never used a pen' has only one correct answer, 'No true' or 'Disagree'); 25 of such questions were placed randomly among real survey questions;
2) answering time (on each page);
3) odd-even consistency (negatively worded items were part of the survey) – they calculated the within-person correlation between a person's vector score on

[150] Dunn et al., 2018

negatively and positively worded parts. They hypothesised that attentive responding should lead to stronger correlations;

4) Long String Index;
5) IRV, as described earlier: standard deviation of the last 150 items from the 325-item questionnaire.

Results of the study were show that participants with low IRV scores do not vary according to how items were worded (positive vs negative). The values of IRV were significantly correlated with other insufficient effort responding indices, except answering time (answering time did not correlate with any other measure), the strongest correlation was present between IRV and LongString Index. Individual response variability also flagged different respondents than other insufficient effort responding measures. The authors recommend reducing insufficient effort responding before data collection by considering possible causes of its occurrence when designing a survey, using multiple measures of insufficient effort responding, paying attention to multiple parts of the survey, especially parts closer to the end, as respondents get tired. One of more important recommendations is to pay attention to cutoff values for used to distinguish careless from attentive respondents, because there is a risk of removing attentive respondents accidentally. Providing comparison of full and clean sample in a footnote could also be useful.

Another study[151] had the sample for this study consisted of 18578 high school students. Data collection happened started in 2009 and ended in 2013. The survey was self-administered and had 108 items, forming a total of 10 subscales, answering scale was a Likert-type scale. The authors used nine indices of FALSE responding: (1) per cent of items left blank (more than 15% left blank), (2) intra-individual response variability (simulating null dist.), (3) inter-item standard deviation (simulating null dist.), (4) long string (scree approach, different thresholds), (5) mean absolute difference between positively and negatively worded items (simulating null dist.), (6) psychometric antonyms (simulating null dist.), (7) average item-rest regression residual (simulating null dist.), (8) squared Mahalanobis distance (1% Type 1 error cut-off), (9) the standardized log-likelihood (1% Type 1 error cut-off).

[151] Steedle et al., 2019.

Results of this study were that 3% of respondents most likely to exhibit insufficient effort responding based on each index (total of 15.5%) were removed.

WARNING SIGN #4 Too low declarative cooperation level, logical inconsistency, and odd answers to open-ended questions

This sign consists of three different parts: (1) declarative cooperation level – how respondent evaluates their engagement and effort; (2) logical inconsistency – do answers to logically related questions coming from the same respondent contradict each other or not; (3) odd answers to open-ended questions – answers than cannot be interpreted in a way that allows connecting them with the question content (weird answers).

Declarative cooperation level

Suppose there was a question about the respondent's approach to the research included at the end of the survey (in a face-to-face interview, this can be assessed directly by the interviewer, but in online research, it is the respondent who needs to admit what he did, or did not do, during the survey). In that case, it gives a respondent the last chance of admitting that they were not answering carefully, and data coming from them should be discarded – if, of course, they decide to be honest with the researcher.

There is a high chance, especially when they were informed that they are being 'watched' and will not get their reward if caught, that respondents will not answer such questions honestly, so answers indicating high effort should be treated instead as a usual way of avoiding responsibility in some cases, especially those indicated as 'suspects' by other warning signs.

In the study on declarative cooperation (self-reported low level of cooperation)[152] sample consisted of 3490 Germans recruited through Google AdWords on searches related to elections. The authors hypothesized that serious (attentive) participants would provide more coherent and valid data than non-serious participants. The survey's content was concentrated around voting intentions in incoming elections and political attitudes. The authors used a simple question about the self-assessed seriousness of answers at the end of the survey. The question had only two possible answers: respondents could say that they took part seriously or not.

[152] Aust et al., 2013

Results of this study show that 3.2% of respondents of the sample admitted that their answers were not serious. For serious respondents (by this method of detection) correlation between political attitudes and intention of voting was significantly higher than for non-serious respondents, which improved data quality even after excluding duplicates based on IP addresses.

Suggestion for other researchers: include seriousness checks in online studies.

Potential risks associated with exclusion pointed out by the authors: (1) it may exclude valid data (proposed solution: ask participants to elaborate about their non-seriousness), (2) non-differentiating between different types of FALSE responding (solution: include other methods), (3) multiple screening methods increase 'researcher degrees of freedom'[153] (solution: a priori consideration of exclusion criteria).

Logical inconsistency

This part depends on the availability of questions that can be compared. If there are none, it cannot be used.

In the first variant of the sign, it should use answers to directly connected questions, and contradictory answers indicate a lack of attention (or, on the contrary, playing around on purpose, which can be equally worrying). The best and most straightforward method to do this is using crosstabulation and select cells, for which a combination of the answers is impossible to be true for the same person at the same time. The limitation of this part is that it does not say anything about other parts of the survey – only about this part which had the questions that contradict each other, so it should not be used as an indication of inattentive responding on its own.

In the second variant of this sign, instead of checking particular answers through crosstabulation, at least two questions, for which expected relations are known, either from theory or previous research, should be checked in terms of correlation coefficients and categorised accordingly separate groups. This approach is the one that has been tested in this dissertation.

[153] Simmons et al., 2011

In a set of five studies on logical inconsistency[154], the author tried to determine if a two-factor solution to Affective Commitment Scale (ACS) was substantive or methodological (because negatively worded items were included in the scale). The aim of the studies was to determine experimentally if a two-factor solution is more likely to show under conditions of different levels of cognitive fatigue (if it is consistent across conditions, it is substantive, if it is not consistent, it is methodological and caused by careless responding or negatively worded items being harder to process).

In Study 1, participants were randomly assigned to complete ACS before or after completing a 60min experimental task. The sample of the study consisted of 184 students from US universities for extra course credit. Results of the study show that CFA for before condition had a close fit for the unidimensional model, but two-factor model gave statistically similar fit, with the correlation between two factors close to 1; for after condition CFA had a poor fit for one factor solution: positively and negatively worded items loaded two different factors but failed to have an acceptable fit even for the two-factor solution, with reduced correlation (0.74) between two factors. This means that the two-factor solution is methodological.

Study 2 had two conditions – original ACS vs modified ACS (original had 50% negatively worded items, modified had none), at the end of another long task for all participants. The sample consisted of 369 psychology and business students. The result of this study shows that, in the modified condition, the unidimensional model had a borderline fit, two-factor similar, a correlation between two factors 0.98; in the original condition, the unidimensional model did not fit the data, but two-factor did fit data better, with the correlation between factors equal to 0.78. This also confirms that the second factor is methodological.

In Study 3, reversed direction for previously positively and negatively worded items were used. The author also added a attention check question - manipulation check type, cognitive workload scale, NASA Task Load Index, check fatigue), and repeated measures (completing ACS twice – at the beginning and the end). The sample consisted of 118 students that could earn bonus points for the course. The result of this study was that (1) negatively worded items had lower correlations – participant had difficulty responding to

[154] Merritt, 2012

negatively worded items; (2) tasks meant to increase participant's fatigue indeed increased mental fatigue; (3) unidimensional model fit poorly at Time 1, but two-factor fit better, although with a lower correlation between factors (0.96 vs 0.88), meaning that factor structure was affected by negative wording even being early in the study, median split into groups of lower and higher reported fatigue at Time 1 showed, that unidimensional model fits well for lower fatigue, but poorly for higher fatigue; (4) for Time 2 level of fatigue did not matter, one-factor model had a poor fit, but two-factor sig improved fit.

Study 4 used the same method as Study 3. The sample consisted of 313 full-time employees recruited through a market research company. The result shows that at Time 1 unidimensional model fit well according to part of the statistics (good NNFI and CFI stats, but poor CFI and χ^2), two-dimensional fit equally well, a correlation between factors 0.99, in Time 2 unidimensional fit poorly, but two-factor fit significantly better, correlation 0.66. Results from students confirmed on employees.

In Study 5, a potential testing solution to negative wording items was used – highlighting them so that they stand out (1st condition) and no highlight (2nd condition). The sample consisted of 262 adults (working at least 15h/week) who were also students. Results show that not highlighted had poor fit to the model, but highlighted also had a poor fit, so the conclusion is that highlighting is insufficient to overcome fatigue.

Odd answers to open-ended questions

When the respondent does not want to cooperate, they usually do not want to exert more than minimal effort, which results in giving one-word replies, connected with the question or not, weird strings of unrelated, nonmeaningful signs, question marks or other things that do not convey any meaning nor value for the researcher. This is the most subjective of the parts, as the researcher has no way of telling if a respondent had something meaningful in mind or not by giving such an answer (besides obvious cases) – it is, often, a form of guessing, especially in borderline cases.

An example of a study that used the quality of answers to an open-ended question[155] had this method combined with other methods of detection, namely with a self-reported low

[155] Brühlmann et al., 2020

level of cooperation, attention check questions (infrequency type and instructed response item), answering time, LongString index, Odd-Even Consistency, Resampled Individual Reliability[156], and Person-Total Correlation[157]. The authors used an experiment that previously produced significant results and analysed the consequence of excluding FALSE respondents. The sample consisted of 394 participants coming from a crowdsourcing platform, with monetary compensation as a reward.

Criteria for the quality rating of answers to open-ended questions came from another study[158]and were as follows: (1) if the response was thematically substantive, (2) if the required minimum of words was met (50 words as per instruction in the case of this study), (3) if the answers provided were complete sentences (as instructed in the study), (4) number of sub-questions answered, (5) number of sub-questions further elaborated.

All answers were coded manually by one of the authors, and a random subset of 100 answers was coded by another author to check inter-rater reliability (which was very good, 0.96 for the index based on the five aspects enumerated above). The codes for the answer quality index were 'insufficient', 'high', or 'excellent'.

Results of this study showed that 25.4% of participants had insufficient open-ended answer quality. This measure only weakly correlated with other measures of FALSE responding (correlation coefficients from 0.13 to 0.26). This outcome may have different possible causes: odd answers to the open-ended question may simply detect another type of FALSE responding, so the low values of correlation coefficients are not surprising, but this outcome may have its source in the way answer quality was assessed – some respondents are used to online communication, which is often based on using part of sentences, or even single words and emojis to convey a message, so they may not value using full sentences, elaborating, using many words, or answering all sub-questions, even when explicitly asked. This behaviour may indicate a lack of cooperation but does not necessarily indicate a lack of attention.

[156] Curran, 2016
[157] Curran, 2016
[158] Holland & Christian, 2009

Other methods of detection

Other methods found in the subject literature include response latitude (rating scale use), respondent's goodness of fit, comparing responses with randomly generated data, and serendipitous findings.

Examples of studies and reports using these methods have been described below.

The authors of the study on **response latitude** (rating scale use)[159] started with a general research problem, that wide response latitudes are likely to be associated with degraded psychometric properties and decided to compare how the participant's involvement (engagement) in the survey influences participant's response latitudes. To answer questions (hypotheses) connected with this problem, they designed and conducted several survey studies designed to be high involvement (HI) or low involvement (LI) surveys. They expected that respondents in LI surveys are more likely to be satisficing, and therefore answers to these surveys will have poor psychometric properties. The study had four hypotheses.

Hypothesis 1 was that HI attitudes relate to significantly narrower response latitudes than LI attitudes, operationalized as distances between graded response model *b* parameters (option response boundary). To test it, the authors compared LI (attitudes about employment testing) vs HI topics (attitudes about drinking. The sample consisted of students, N=971. T-tests showed significant differences in mean distance between b parameters, which means that LI surveys wider response latitudes, which means that hypothesis was supported, and respondents discriminate between rating scale options less when their involvement in the survey's topic is higher.

Hypothesis 2 was that response latitude width relates to discrimination in test characteristic curves. Narrow-latitude surveys are associated with greater discrimination than wide-latitude surveys. The authors used a median split based on involvement on a survey with high variability of involvement (attitudes towards sex). The sample consisted of students, N=503 HI, N=442 LI. H2 was supported – curves were steeper for discrimination parameter *a* (proportional probability that people are likely to select

[159] Lake et al., 2013

Strongly Agree option at various points in the survey) in HI surveys, which means higher discrimination between answer options in HI surveys.

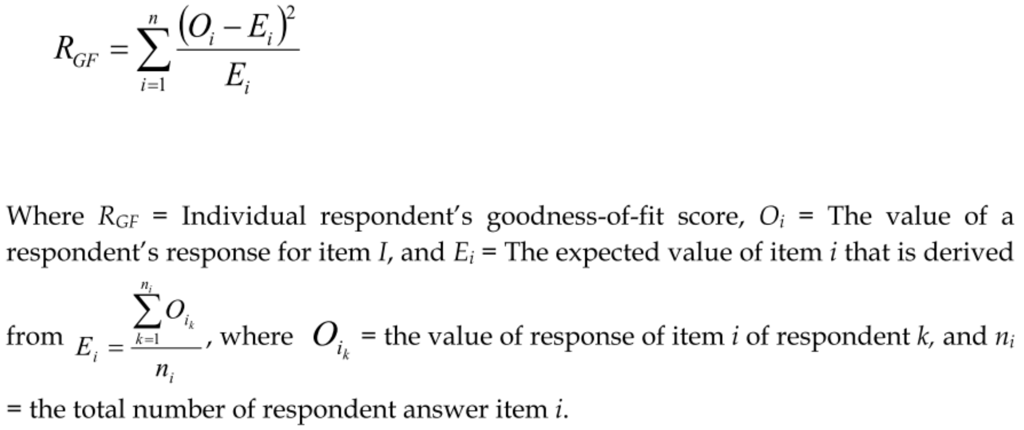
Hypothesis 3 was that response latitude width relates to testing information. Narrow-latitude surveys are associated with more information than wide-latitude surveys. Experimental manipulation was used to test this hypothesis – researchers forced participants to become HI from LI (the survey was about attitudes towards prisoners, the experiment had a description about prisoner release program in their state, stating that the prisoners will be working in common businesses, like local shops – meaning that participants would likely encounter prisoners in their daily life; the control group had a different state in the description – scenario irrelevant to participants because they would not encounter prisoners in their daily life). The sample consisted of students, N=484 for high involvement, N=487 for low involvement conditions. Results showed that H3 was partially supported – greater test information means reduced measurement error; for 2/3 comparisons, test information was greater for HI. That means that when the involvement of a participant is higher, they also provide more relevant answers.

Hypothesis 4 was that narrow-latitude attitude scores have greater validity than wide-latitude attitude scores. Comparison of attitude scores correlated with single-item self-reported attitudes in HI and LI condition and Z-tests was used to test if that was true.
Results of testing showed that H4 was supported – correlations were stronger in HI conditions than in LI conditions. All three *z* scores for comparison of correlations between HI and LI were significant (.05 level). Cronbach's Alphas were higher for the HI condition.

A study on respondent's goodness of fit[160] as a method of FALSE respondent detection is an interesting example of a measure more complicated than most of the detection methods, as it uses expected values for a given item as a benchmark for respondent's answers. The goodness of fit score (R_{GF}) is used by the author of this study to check the consistency of observed frequency compared to expected frequencies of response to items in a questionnaire. A small score on the R_{GF} means that the respondent is considered consistent; a large score means that the respondent is likely inconsistent in their answers.

[160] Kountur, 2016

The author used Cronbach's Alpha to assess the consistency of items – this indicator relies on the assumption that respondents are consistent in responding. That means that if the questionnaire was found to be reliable during tests, inconsistency found later must be caused by respondents. The assumption is that careless respondents do not fit expected responses, and careful respondents do fit.



$$R_{GF} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where R_{GF} = Individual respondent's goodness-of-fit score, O_i = The value of a respondent's response for item I, and E_i = The expected value of item i that is derived from $E_i = \frac{\sum_{k=1}^{n_i} O_{i_k}}{n_i}$, where O_{i_k} = the value of response of item i of respondent k, and n_i = the total number of respondent answer item i.

***Figure 3** Respondent's goodness of fit index and its description*
Source: Kountur, 2016, p. 3

The method used by the author was experimental manipulation. One group (E1) was instructed to give proper and given enough time to answer the questions mindfully; the other group (E2) was given very little time to complete the same survey and was encouraged to answer randomly. Group E1 was given 30 minutes to complete a questionnaire about natural medicine. They were observed to be sure they truly completed the questionnaire. Those who completed it too fast were later rejected. Group E2 was given 5 minutes to complete this questionnaire and were encouraged to give random answers. If there are FALSE respondents in the data, the distribution of R_{GF} will be skewed to the right; when they are true, the distribution should be normal. The score of R_{GF} that separates true and careless is the area between normal and skewed distribution. The sample consisted of 185 respondents in the E1 group, 20 respondents in the E2 group and was purposive (participants selected on the criterion that they are familiar with natural medicine).

Results showed that group E2 had significantly higher values of R_{GF}, a frequency distribution of careless responses starts closely at the point when the tail of the skew begins: range of R_{GF} between 09.7 and 12.5 may be used as a boundary of valid responses, R_{GF}s higher than that comes from FALSE respondents. This means that at least based on values derived from a group instructed to be careless, the index can be used as an interesting additional measure of FALSE responding.

An important limitation of this study is the use of limited self-report questionnaires to measure affective domain in humans, and the index has no exact cut-off score – it is necessary to identify a model that can determine the cut-off score first.

The questionnaire used by the authors of the study on comparing with responses randomly generated by a computer[161] was PID-5 (DSM-5 questionnaire measuring five domains of maladaptive personality). Initially, this questionnaire had 220 items, shortened version (used in this research) had 100 items, with a short (10-item) version of response inconsistency scale (INC-S) designed by other authors (the original INC[162] consists of 41 item pairs and reflects the sum of absolute difference scores within each similar pair – high scores=greater inconsistency). The study aimed to check whether a 10-pair version of the questionnaire distinguishes between FALSE and attentive respondents. The sample consisted of 246 students and 209 respondents from Amazon's MTurk. Participants were excluded from analyses if they had omitted more than 5% of items, had any item of short INC unanswered, or the survey was completed too quickly. The authors used the total score on INC short as a method of detecting FALSE responses. It was calculated as a sum of absolute differences within the item pair for all ten pairs. Random and non-random data were compared by generating random response sets (using uniform random response distribution in MS Excel) and comparing them to answers given by respondents. Three data sets were used for comparisons (50, 30 and 10% of randomly generated responses).

Results show that randomly generated data scored higher on INC-S than students and MTurk respondents – INC-S has a high level of discriminability between random and non-random data, which can detect random responses. The remaining two studies aim to cross-validate short scale with another sample and with a longer version of the INC. Both validations turned out to be acceptable for further use. This means that comparing with meaningless data generated by a machine may be a useful way of FALSE respondent detection – at least for survey answers done by bots instead of real people.

[161] Lowmaster et al., 2020

[162] Keeley et al., 2016

1.3 Proposed procedure for detecting FALSE responding

Based on the literature review, detecting FALSE respondents has been proposed and will be described below. The procedure consists of several steps, each of which aims to detect a different kind of FALSE respondents. Shortened version of this procedure is presented at the end of the dissertation.

WS 1. Answering time. It requires prior planning (not every survey software allow the collection of data on the duration of the test), and is based on the duration of the test and the speed of reading. We cannot define the minimum thinking time, therefore, the assumption that the respondent has ready answers to the questions in his head and he has enough reading time basically excludes the possibility of rejecting people who have at least read the questions.

Step 1. Calculate the Overall Answering Time of the study.

Step 2. Calculate the total number of words in the survey, remembering that the respondent does not read the repeated rating scale every time (it should be included in this number only once). For this purpose, we can use the automatic word count function available (in example in MS Word).

Step 3. Calculate the minimum time needed to read the questions, and assume the maximum reading speed (enabling comprehension) considering the characteristics of the respondents (students read faster, etc.). - in this dissertation, the assumed maximum reading speed is **300wpm** for all datasets.

Step 4. Divide the respondents into those who read faster than the minimum time and those who are below this threshold. Extremely long times are not a problem because in surveys we do not ask people to hurry.

Step 5. If the survey had optional questions, breaks, or other elements that the respondents might have omitted, but the omission of which does not affect the main objectives of the survey, repeat steps 1-4 also for the version that does not include these elements.

Step 6. If time data are available for individual question blocks / single questions, repeat steps 1-4 for each question / question block, remembering that the repeating scale of answers only counts towards the number of words in the first question in the series, and that the answering time to the first question after changing the topic / type of question is always longer than the times of subsequent answers.

Step 7. Check the time distributions both globally (OAT , OAT without optional elements) and locally (PAT for question blocks, PAT for individual questions if data is available) - respondents may be 'FALSE' only in some parts of the survey.

Step 8. Set the percentage threshold for the number of blocks / questions answered too fast, above which a respondent should be excluded from the survey globally (not only from individual analyses) - it depends on the type and purpose of the survey.

WS 2. Attention check questions. It requires prior planning, like WS1, both in terms of number, location, and content of the questions.

Step 1. Count the number of incorrect responses to attention check questions for each respondent.

Step 2. With the number of questions above 3, we can allow 1 error in the attention check questions, if we need to be more lenient (smaller data sets, etc.) or allow no errors, we can be stricter. The decision as to whether to make a mistake or not is up to the researcher.

WS 3. Low Differentiation Rating Style and Non-Informative (DK) Answers. It requires partial planning - including DK answers in the scales ('Don't know', 'It's hard to say', etc.). It is not possible to do this for individual questions.

Step 1. Checking DK anwers for each respondent.

Determine which series of questions can be used to analyse DK answers and variance - e.g., with the same rating scale, on the same topic, matrix questions (multiple statements on the same page) and calculate the number of DK answers in each group of questions with the same rating scale.

Step 2. Check rating style.

Step 3. Decide what percentage of DK answers is acceptable respondents with too high number of DK should be flagged.

Step 4. Calculate the standard deviation of the respondent's answers for the same groups of questions. If it is 0, the respondent should be excluded (except for question scales where identical answers to subsequent questions do not constitute contradictions).

Step 5. Both sub-signs can be global or local (if there is more than one question group), so we can exclude respondents based on this sign globally or locally.

WS 4. Low declarative cooperation level, logical inconsistency, odd answers to open-ended questions. It requires prior planning - placing direct questions about the respondent's participation in the study at the end of the study and questions of 'twins' to which the answers should be consistent / correlated in a certain way.

Step 1. Determine what level of commitment indicated by the respondents is sufficient (e.g., on a 6-point scale, where 1 is a complete lack of commitment, and 6 is a very high commitment - answer:
three or more?).

Step 2. If possible check the logical consistency (e.g., The question about the number of children and the question about satisfaction with the relationship with children), check whether the respondents gave consistent answers - inconsistent answers are a problem in the analyses.

Step 3. Code answers to open-ended questions into 5 categories: (1) no answer, (2) answer not connected with the topic of the question, (3) too short answer, (4) informative answer, (5) refusal. If questions were obligatory, no answer counts as a too short answer. Whether too short answers should also be treated as non-informative depends on the decision made by the researcher. The safest option is to flag only respondents who clearly gave noninformative (odd) answers.

Filtering out FALSE respondents. Depending on the amount of data available, the criterion of exclusion can be either lenient, allowing one of the signs to be identified, or strict, excluding respondents flagged by any of the signs.

The lenient form is more applicable with smaller samples because it allows leaving a larger number of respondents for further analysis, but it should be remembered that they may disturb the relationships between the studied phenomena, and with smaller numbers of respondents, the impact of each of them on the results is greater and may change them completely. The researcher should analyse the advantages and disadvantages of excluding fewer or more people in relation to the purpose of the study. If it is decided that FALSE positives (excluding possibly attentive respondents) are unacceptable, a lenient criterion (allowing flagging by one of the signs) should be applied. If potentially unreliable data is unacceptable, strict criteria (excluding respondents flagged by any sign) should be applied.

2 EMPIRICAL PART

2.1 The aim of the empirical part

The starting point was to determine the operationalization of 4 warning signs and test them on 12 data sets – 9 samples collected online and three collected offline.

The first research task is to determine the level of respondents' inattention – that is, what percentage of respondents should be excluded from further analyses.

Respondent's inattention may be global or local.

We talk about global inattention when a respondent 'plays' by clicking and not giving any attention to the answers given – such a respondent can be easily captured during an interview but is extremely difficult to detect during an online survey.

Local inattention is when a respondent loses their attention, ponders, or deliberately ignores a block of questions but answers other blocks/questions with due care.

The second research task is to determine the consequences of ignoring the problem of FALSE respondents. For this purpose, the reliability of measurement was checked in two groups: attentive and FALSE respondents. As part of this task, the usability for the new method of detecting FALSE respondents was also tested using the FLEXMIX procedure (finite mixtures of generalized regression models).

2.2 Description of data sets

The summary of descriptive statistics for all data sets is presented in Attachment 1. General description of all analysed data sets is presented below – three data sets (A1, A2, C) being employee samples, six data sets were students samples (B1-B6), and last three (D, E1, E2) offline general population samples from big surveys, used here as a comparison for online samples.

Data set A1 (N=1 421)

Data comes from a survey done through a Polish commercial online panel. Participants collect points in exchange for participation in the survey, and can later exchange these points for various rewards.

Sample selected purposefully to consist of participants with at least secondary education, between 24 and 42 years old, and 96% of the sample had at least secondary education – as these were the characteristics needed for the main research goal, not connected with FALSE respondents study.

Data was collected in 2018.

Data set A2 (N=1 497)

Data comes from a survey done through a Polish commercial nationwide panel (the same as data set A1). Participants collect points in exchange for participation in the survey and can later exchange these points for various rewards.

Sample selected purposefully to consist of employed participants with at least secondary education, coming from Mazovian voivodeship. As the requested quota was not met for this administrative area, additional respondents were invited from two cities: Lublin and Łódź. Participants who had work experience but were currently unemployed (short-term unemployment) were also allowed to take part in the study, as they had required work experience.
The influence of offering personalized feedback on the FALSE responding rate was tested on this data set.

The study was conducted in July of 2020.

Data set C (N=287)

Employees with at least three years of work experience were invited to fill in the survey using the snowball method. Students who invited employees to participate in this survey could get bonus points for the course ‘Sociology in business’ (obligatory) and were warned that the data will be screened for FALSE respondents and no points will be awarded for those invited employees who fail the screening.

Data was collected in April and May 2020

Data set B1 (N=740)

Students of the University of Warsaw participating in the ‘Psychology in business’ course (obligatory) (2018) and other courses at the Faculty of Management, who for participation in the research could receive bonus points at their respective courses, and employees invited.

Data was collected in 2018, partially in 2017.

Data set B2 (N=341)

Dataset consists of students of the Faculty of Management, University of Warsaw, participating in the ‘Sociology in business’ course (obligatory). Participants could get bonus points in the course for participation in the research and were warned about data screening FALSE respondents – those who did not pass the screening did not get any bonus points for participation.

The study was conducted in April-May 2020.

Data set B3 (N=414)

Students of the Faculty of Management, University of Warsaw, participating in various courses, some got bonus points for participation in the research, some did not, depending on the course instructor. Students were warned about FALSE respondents screening only at the beginning of the survey. If there were extra points in their respective course, getting those points depended on the outcome of the screening. The study was a pilot study, and the influence of disciplining reminders on FALSE responding rate was tested on this sample.

The study was conducted in January of 2021.

Data set B4 (N=308)

Students of the Faculty of Management, University of Warsaw, participating in the ‘Psychology in finance’ course, offered bonus points and feedback for participation in the survey and personalized profiles based on their answers (optional) to increase interest and engagement in attentive responding to questions. Students were warned about FALSE respondent screening by the course instructor, and at the beginning of the survey, awarding of points depending on whether they passed the screening or not.

The study was conducted in May-June 2021.

Data set B5 & data set B6 (N_{B5}=140, N_{B6}=497)

Students of the Faculty of Management, University of Warsaw, participating in various courses, got bonus points for participation in the research, but how much points it was worth depended on individual course instructors. Students were warned about FALSE respondent screening at the beginning of the survey.

The study was conducted in May-June 2021.

Data set D (N=1203)

European Working Conditions Survey (EWCS, 2015, N=43 850) is a survey based on interviews with working people; data comes from 28 EU Member Countries, the five candidate countries (Albania, Macedonia, Montenegro, Serbia, Turkey), Switzerland and Norway. An interviewer recorded answers given by a respondent. 90% of the sample had at least secondary education.

Analyses in the main part of the dissertation were done on a Polish sample.

Data set E1 (N=1000)

Data comes from World Values Survey, Wave 5 (2005, N=83 975), 58 countries. Personal face to face interviews in 53 countries, postal interviews (self-completed, pen and pencil) in 3 countries (Australia, Japan, New Zealand), mixed in one country (Taiwan), and electronic (online, non-voluntary) in one country (USA).

Analyses in the main part of the dissertation were done on a Polish sample.

Data set E2 (N=966)

Data comes from World Values Survey, Wave 6 (2010, N=89 565). Personal face to face interviews (explicitly stated) in 15 countries, postal interviews in one country (Australia), phone interviews in one country (Nigeria), online (non-voluntary) in one country (USA), missing information about the procedure in 41 countries.

Analyses in the main part of the dissertation were done on a Polish sample.

Availability of data in data sets.

In **Table 3** below, data was marked 'yes' only if the information met the criteria of usability described earlier in the dissertation, as not all datasets contained measurments for all warning signs.

Data set	**Time**		**Attention check questions**	**Differentiation style**	**Declarative cooperation**	**Logical consistency**	**Open-eneded questions**
	OAT	PAT					
A1	Yes	Yes	Yes	Yes	Yes	Yes	Yes
A2							Yes
C		No					Yes
B1		Yes	No		No		Yes
B2		No	Yes		Yes		No
B3						No	No
B4		Yes				Yes	Yes
B5							No
B6							Yes
D	No	No	No			No	No
E1							
E2							

***Table 3** Availability of data in analysed data sets*

2.3 Four warning signs

2.3.1 WS1 Too short answering time

Responding to questions too quickly (‘speeding’) may concern the entire survey or individual blocks of questions (questions formulated in a similar way, with the same rating scale), or even the questions themselves.

There are several ways to establish a minimal reading time:

Globally:

1. Arbitrary minimum time threshold - may be based on the estimates of the software used or the authors of the study
2. The threshold established based on the pilot tests, how long does it take to complete the survey, requires test participants about whose involvement and attention we are sure, the threshold is the time of the fastest such participant
3. Empirically - based on the distribution of participants' answering times - the trials differ in the level of motivation, so the data are also different
4. Reading speed threshold (multiple thresholds possible for different subgroups) based on the total number of words in the study and the total duration
    a. Conservative approach: participants can process the information from the question while reading it, have ready answers, do not take time to think, assumes the upper limit of reading speed at 600 words per minute [wpm] (limit from studies of people training speed reading), or closer to reality the limit of 300 wpm, assuming that the average person reads at a speed of 150-250 wpm
    b. Realistic approach: participants can understand the information while reading but do not have ready answers to it (questions require at least a moment of reflection, lowers the threshold to 250wpm)
    c. Rigorous approach: participants who do not have ready answers are not fast processors, questions require more reflection/analysis, the threshold is between 180-200wpm.

Locally:

1. Selecting arbitrary thresholds for individual blocks/parts / individual questions (generalization at a different level of detail)
2. Setting thresholds for blocks/parts/questions with pilot tests, without division into groups
3. Threshold determined by the speed of reading (as above), possible to use one or more methods of checking, depending on the purpose of the test, from 20% to 50% of acceptable determined as ‘suspected’ of clicking through:
    a. For each major part of the study
    b. For each block of questions with the same rating scale (which may be the same as a major part or subpart of the major part)
    c. For each question
    d. Mixed approach: different thresholds for different types/complexities of questions, for questions about facts (not requiring reflection/recall/analysis, e.g., sociodemographic, work), higher thresholds (400wpm?), for more complex questions lower (250wpm).

Calculating reading speed:

- the entire final version of the survey to a text editor (MS Word or other software with word count is sufficient) and using the word count function to determine the number of words in the entire survey, part of the survey, block, or individual questions; it should be remembered that in blocks with a repeating rating scale / in matrix questions, the respondent reads the rating scale most often only once, in subsequent questions he does not pay attention to its description, even if it is repeated at each subsequent question/point of the matrix question. Therefore, words describing the scale should be counted only once for a block/question - taking this fact into account requires consciously assigning a specific number of words to specific questions.
- Calculation/estimation of test times (in seconds) corresponding to parts/questions (depends on the tool used, usually possible down to seconds/milliseconds - the difference in measurement accuracy can significantly affect the results)

- Convert times from seconds to minutes and divide the number of words by the results
- Depending on the chosen strategy (conservative vs realistic vs rigorous), scores greater than 600, 300, 250, 180-200 are flagged as suspect and scores less than these values as normal.

2.3.2 WS2 Errors in attention check questions (arithmetic, instructed response)

The attention check questions (checking the participant's attention) can take various forms:

- A request to enter a specific word/phrase in response to an open-ended question (e.g., In this question, please enter the word ‘grass’)
- A request to select a specific answer in a closed question (e.g., In this question, select the answer ‘I strongly disagree’)
- Question about the topic/fact about the research, the answer to which should not be a problem if the respondent has read the previous instructions/descriptions (e.g., How many people were involved in the situation you just read the description of?)
- A question for which there is only one correct answer (e.g., I worked 48 hours a day last week)
- Simple ‘thinking’ questions (e.g., arithmetic) or riddles (e.g., completing well-known sayings such as ‘Don't praise the day before sun....’)

Not all questions of this type are neutral - questions that are absurd, tricky, or explicitly showing the respondent that he is being neglected have a greater chance of causing him to stop cooperating and start paying less attention to the study, so in subsequent studies, questions asking for a specific answer were replaced arithmetic questions, which in the study instructions were indicated as breaks to break the monotony of long series of similar questions.

In the research for this dissertation, two types of attention check questions were used: instructed response items and arithmetic questions. The former is a more obvious way of checking respondents’ attention. The latter is less obvious – in the research, it was presented and used as ‘break’ questions to make the survey more interesting to participants.

As there was always more than one attention check question per data set, thresholds of permitted errors depended on the number of questions used, but generally, this sign was treated as gradable – meaning that respondent could be less or more suspect of being inattentive based on the number of errors.

The lenient criterion for this sign usually allowed between one to three errors. Strict criterion did not allow any errors.

There should not be too many questions of this type, nor should they be questions that clearly violate the rules of cooperative conversation (the questionnaire is a form of conversation between the researcher and the respondent, even if the first one is not physically present) - for example, many studies on inattentive respondents, the so-called trick questions (infrequency scales, dummy questions, etc.) having only one correct answer because the question is absurd (‘I work 48 hours a working day’), the answer hidden in a longer instruction (‘What is your favourite colour. (...) If you read the questions carefully, choose the answer <<green>> in this question’), which may cause the respondent to stop taking the survey seriously because of the feeling that the researcher does not trust him. For this reason, we can use, for example, simple arithmetic questions as attention check questions and longer series of similar questions as breaks. As subjects are more likely to be tired at the end of the study and therefore have a greater tendency to be inattentive, at least one attention check question should be included at the end of the study.

2.3.3 WS3 Too many non-informative answers and a low differentiation rating style

Non-informative answers do not usually pose a problem in surveys, but it is well known that presenting ‘an escape’ from giving a meaningful answer to the question along with other answer options increases the percent of respondents who decide to choose such an option compared to allowing such answer to be given only spontaneously[163]. Using a non-informative answer without offering it explicitly is easier in offline surveys than in online

[163] Cichomski & Morawski, 1996

surveys, as there needs to be some form of ‘other answer’ button to allow respondents to react to it, and this in itself may induce more non-informative answering.

In this dissertation, all data sets analyzed and collected online were based on Survey of Activity Styles (SSA, described shortly later in this part) – which means that they contain many questions with exactly the same rating scale and the same question format. To not force respondents into choosing a random informative answer, the ‘It’s hard to say’ option was always presented at the bottom of the list[164]. This means that in all data sets, subsign concerning non-informative answers is based on SSA questions. The percent of non-informative answers allowed for a respondent to not be flagged as suspect depended on the number of questions in the survey, but it should not be higher than 50%.

The low variance component of this sign is also based on Survey of Activity Styles questions and is broader than just non-informative answer analysis, as it includes the whole rating scale and allows for detection of non-differentiating between scale answer options. This problem may appear as very obvious zero variance, which renders such a respondent basically useless (no variance makes it impossible to conduct statistical analyses), or less obvious very low variance. It is difficult to establish a reasonable threshold for the variance value to be acceptable, and that is the reason this subsign should not be used on itself – it is very easy to make variance value appear normal – in example, a respondent choosing answer at one end of the rating scale in a series of ten questions will have zero variance, but it is sufficient for them to respond just once with the option at the opposite end of the scale, and their variance will not look anything different from a respondent who chooses different options – and example of such a simulation is presented in Attachment 2. The threshold for variance is also dependent on the data itself. The author of this dissertation decided to use two standard deviations below the mean as a guide threshold for the analyses.

2.3.4 WS4 Low level of cooperation

This warning sign is based on (1) answers to questions asked to respondents about (direct) involvement in the research, (2) on giving contradictory answers to logically related questions (e.g., conflicting answers to questions about whether they like being around

[164] In data set C also ‘I do not want to answer this questions’ option was presented, but very few respondents actually decided to use it.

other people) and (3) giving odd (strange) answers to open-ended questions (or not providing them, if they were mandatory).

Declarative cooperation

Questions about respondents' declarative engagement/motivation were identical in online data sets and differed for offline data sets. Questions' content is presented below. Original (Polish) versions of the questions are presented in Attachment 3.

For all online data sets:

1. How do you assess the degree of your commitment to this task?
   Rating scale: 1 - very low, 2, 3, 4, 5, 6 - very high
2. To what extent was this task tiring for you?
   Rating scale: 1 - very tiring, 2, 3, 4, 5, 6 - not tiring at all
3. If you were to participate in the survey again (e.g. tomorrow), would your answers be:
   Rating scale: 1 - the same, 2 - could differ slightly, 3 - could differ diametrally

It is worth mentioning that previously discussed studies have shown that respondents tend to not admit their cooperation was poor.

For face to face data sets:

Data set D - Respondents' cooperation and understanding coded by an interviewer:

1. Respondent cooperation
   Rating scale: Very good, Good, Fair, Poor, Very Poor
2. Did the respondent ask for clarification or have difficulty answering any questions?
   Rating scale: Never, Rarely, Sometimes, Most of the time, Always

Data sets E1 and E2: Respondent's interest coded by an interviewer
Rating scale: very interested, somewhat interested, not interested

The lack of interest can be the sign of FALSE responding, but the interpretation will actually depend entirely on what the interviewer assessing this behaviour of the respondent had in mind as a comparison, whether the respondent was first or last in that day, at which point the interviewer decided what code is correct, and so on. This ambiguity means that this sign should not be used on its own as a way of determining that a particular respondent is FALSE or not – they might have been interested in the beginning, but not in the end, or not in the middle – and without additional data, it is impossible to tell.

Logical inconsistency

Logically inconsistent (contradictory) answers in online data sets (besides B3) were based on Survey of Activity Styles questions that should either correlate negatively or positively (questions that were logically connected, i.e. were identical besides switching original person A description so it would be person's B description in the second question, or came from the same scale and answers to them should correlate in some way according to theory). The analysis was done by merging 'Like person A' and 'More often like person A' (similarly with person B) rating scale options, leaving 'It's hard to say' in the middle, and calculating the absolute value of the difference between two questions. If the absolute value was equal to 2, a person was flagged as suspect by this subsign.

Data sets B3, D, E1 and E2, did not contain any questions suitable for checking the logical consistency of respondents.

Odd answers to open-ended questions

If this type of question was used in the data set, it could be either obligatory questions or optional questions. In the first case, all answers that are (a) not connected thematically with the question, (b) too short, (c) empty (if the answer was required, not optional), (d) answer is a random collection of words or other meaningless content. If any of these criteria were met by answers, the respondent was flagged by this sign on the particular question – being flagged on any open-ended question (if there was more than one) meant being suspect.

Data sets B2, B3, B5, D, E1 and E2 did not contain open-ended questions.

2.4 Survey of Activity Styles

SSA (Survey of Activity Styles) is an online tool based on ISA (Activity Style Inventory) questions[165].

It consists of a few question blocks (usually about a dozen, but can be less or more, depending on the purpose of the study); each block is focused on the measurements of various question scales.

The SSA editions used in the research in subsequent years are modified depending on the purpose of the study and the studied sample. Each edition includes a measure of interval activity style and reactivity.

Answers to the SSA's questions are subject to a FALSE respondent detection procedure.

Most of the SSA questions is of choice type, so they are more immune to rating style distortions.

The procedure for detecting FALSE respondents will be shown in the example of SSA-type studies. The procedure used in other surveys rejects many respondents due to the WS 3 - too little variance of the answers and too many non-informative answers, and a difference in the way of using the rating scale (e.g., average = 4.8 vs 2.3 on a five-point rating scale).

[165] Wieczorkowska, 1998

2.5 Warning sign #1 – Answering time

This warning sign was analysed on 9 data sets. In data six data sets [A1, A2, B1, B4, B5, B6], reading speed was used to indicate which respondents were too fast, and in three data sets [C, B2, B3], pilot studies on a small number of attentive participants were used to establish minimal time needed to complete the survey.

For **reading speed** data sets, the overall answering time (OAT) and partial answering time (PAT) for question blocks/series or individual questions were used.

If a respondent was faster than the given threshold (300 wpm) for OAT or PAT, they were flagged by this sub sign. In case of too many questions, respondent was flagged if they speeded through more than 50% of pages/questions.

OAT considered all the words that appeared on any page of the survey. If the respondent does not read the rating scale each time in the matrix questions / in blocks of questions with the same rating scale, the words included in the description of the rating scale were counted only once in these word sums, at the first appearance of a given rating scale.

PAT for blocks excluded non-obligatory parts of survey and breaks and was only calculated for blocks of questions having similar form (same rating scale, same format), without matrix questions or questions requiring typing. The rating scale was also counted once in word count for the whole block.

PAT for individual questions also excluded non-obligatory parts of the survey and breaks, and is the most detailed part, but also based on a different exclusion criterion. In this part, the respondents were flagged as too fast for individual pages (regardless of whether they contained one or more questions or whether it was a matrix question), considering that the rating scale was read only once in the case of blocks of matrix questions and questions. Respondents who were too fast more than 50% of the question pages displayed to them were flagged.

For **minimal time determined by attentive participants** assessment, a small sample of trusted participants was asked to complete the survey reading and answering questions attentively, but as fast as they could without skimming or skipping questions. Depending on survey length, minimal time obtained this way was later lowered by about 5 minutes to avoid excluding attentive respondents who may have been faster readers than the fastest reader of a pilot sample.

Details of analysis methods used are specified case by case below – otherwise, data sets were analysed the same way as described above.

Overall, for data sets with reading speed, this sign flagged between 3.2% and 52.5% of the respondents, and for data sets using minimal time determined by attentive participants, between 6.3% and 8.7% of the respondents.

2.5.1 Reading speed

Overall answering time

Descriptive statistics for OAT (in minutes) are presented in **Table 4**. All answering times have been truncated to 3 hours, as even with breaks, it should not take respondents more than 3 hours (180:00 minutes) to complete any of these surveys.

Data set	N	Mean	Median	SD	Min
A1	710	25:10	13:48	35:07	0:59
	711	22:50	14:08	30:28	1:00
A2	749	40:32	27:53	39:35	4:23
	748	38:20	25:12	39:40	3:59
B1	740	55:39	42:39	38:52	2:22
B4	308	64:37	53:33	58:31	19:14
B5	140	45:40	37:47	26:11	10:12
B6	497	21:07	17:54	15:06	3:58

Table 4 *Descriptive statistics of overall answering time for six data sets [A1, A2, B1, B4, B5, B6]*

Outcomes of OAT analysis for all six data sets are presented in **Table 5** below. As has been mentioned earlier, the reading speed threshold was 300 words per minute for all data sets.

Data set	Respondents below threshold [count]	Respondents below threshold [%]
A1	614	43.2
A2	171	11.4
B1	25	3.4
B4	4	1.3
B5	3	2.1
B6	23	4.6

Table 5 *Results of overall answering time analysis on six data sets [A1, A2, B1, B4, B5, B6]*

It is apparent from the above outcomes that student samples were slower than panel samples – as much as 43.2% of respondents were flagged just by the overall answering time, without taking into account that it's easy to pass this threshold by just taking a break. As the reading speed threshold and method of counting words were exactly the same for all six datasets, the difference seems striking.

Overall answering time without breaks and open-ended questions

Respondents can elongate their OAT just by taking a break during the survey. Thus, by calculating the indicator excluding breaks and open-ended questions, respondents who achieved acceptable time in the first measure (general OAT) will be excluded if they speeded only at closed questions. Excluding breaks leaves only the time spent on survey answering, and removing the open-ended question from analysis mitigates differences that respondents may have in typing skills.

Data set	Respondents below threshold [count]	Respondents below threshold [%]
A1	698	49.1
A2	228	15.2
B4	4	1.3
B5	3	2.1
B6	31	6.2

Table 6 *Results of analysis of overall answering time without breaks and open-ended questions on six data sets [A1, A2, B4, B5, B6]*

As can be seen by comparison of **Table 5** and Table 6, excluding breaks and open-ended questions slightly increases the number of flagged respondents in three out of five data sets, but again, the difference is bigger in commercial panel data sets – 5.9 p.p. for data set A1 and 3.8 p.p. for data set A2.

Data set B1 did not have any open-ended questions or breaks, and therefore repeating analysis, in this case, was redundant.

Partial answering time – blocks

Block analysis was done on two data sets [A1, B1], as these data sets came from studies that had clearly distinguished parts that could be treated as blocks. The remaining studies did not have the same type of questions (formatted the same way, similar content, the same rating scale, etc.) combined into blocks[166].

Data set	Number of question blocks	The threshold for the number of blocks speeded	Respondents above threshold [count]	Respondents above threshold [%]
A1	20	10 blocks (=50%)	774	54.5
B1	15	8 blocks (≈53%)	29	3.9

***Table** 7 Results of answering time block analysis for data sets A1 and B1*

Results presented in **Table 7** show a big difference between percent of respondents flagged by this sub sign. Again, it is clear that the paid panel sample has had a much worse rate of FALSE responses than the student sample.

Partial answering time – individual questions (pages)

Individual question time analysis was done on four data sets [A2, B4, B5, B6]. The analysis excluded breaks and open-ended questions for the same reason as stated above in the case of overall answering time analysis. In this part, the respondents were flagged as too fast for individual pages. Thresholds for each study, in a number of pages, are given in **Table 8**, along with the percent of a number of questions given in brackets.

[166] This was an intentional decision made after data from data sets A1 and B1 have already been collected – to increase respondents' attention and engagement by avoiding having them to answer high number of questions looking almost exactly the same. It would be possible to divide these data sets into blocks, but it was decided that a better approach would be to analyse individual pages, as it would take into account more information about respondent's behaviours in case of non-monotonous survey design that was used in those studies.

Data set	Number of questions	The threshold for the number of pages speeded	Respondents above threshold [count]	Respondents above threshold [%]
A2	84	42 (=50%)	181	24.2
	93	47 (≈51%)	95	12.7
	Total	**-**	**276**	**18.4**
B4	237	119 (≈50%)	10	3.2
B5	181	91 (≈50%)	10	7.1
B6	66	33 (=50%)	45	9.1

Table 8 *Results of individual questions (pages) answering time analysis for four data sets [A2, B4, B5, B6]*

The same patterns seem to be present in this sub sign – panel sample has a higher rate of speeders than student samples.

Comparison of results of answering time analysis on different levels of detail, on the example of data set A2

Below is a comparison of all three parts of Sign #1, on the example of data set A2. It shows that the individual parts of the sign only partially flag the same respondents, but the greater the number of sub signs the respondent has been flagged with, the more certain we are that he really belongs to the group of globally inattentive respondents.

		OAT too fast	OAT acceptable	Total
PAT too fast, above 50% of pages	**OAT too fast, only obligatory**	**6.28%**	**1.34%**	7.62%
	OAT acceptable, only obligatory	**-**	**10.82%**	10.82%
PAT acceptable, below 50% of pages	**OAT too fast, only obligatory**	**5.14%**	**2.47%**	7.62%
	OAT acceptable, only obligatory	**-**	**73.95%**	73.95%
Total		11.42%	88.58%	100%

Table 9 *Comparison of three ways of analysing respondent's answering time, data set A2*

In the case of data set A2, all three parts of Sign #1 flagged a total of 26.1% of the respondents, assuming that the respondent was not allowed to be flagged by any sub sign, or 10.8% assuming that the respondent could be flagged by only one of the sub signs to be flagged. 6.28% of the respondents were flagged by all three sub signs.

2.5.2 Completion time determined by a pilot study

Overall answering time

Descriptive statistics for OAT are presented in **Table 10**.

Data set	N	Mean	Median	SD	Min
C	1421	61:45	38:04	54:49	6:45
B2	341	40:48	29:09	37:30	1:02
B3	191	42:12	27:08	41:06	8:37
	223	34:31	26:59	28:37	7:12

Table 10 *Descriptive statistics of overall answering time for three data sets [C, B2, B3]*

In the case of all data sets, minimal completion time is too small to attentively respond to a survey (all surveys were quite long, which is reflected in mean and median of completion time. As was done previously, the overall time has been truncated to 180 minutes to diminish the influence of extreme outliers.

Outcomes of OAT analysis for all three data sets are presented in **Table 11** below.

Data set	Number of participants in pilot testing	Minimal time determined by attentive participants [minutes]	Threshold time used in the analysis [minutes]	Respondents below threshold [count]	Respondents below threshold [%]
C	15	20:00	15:00	25	8.7
B2	20	16:40	16:40	29	8.5
B3	10	20:00	15:00	33	8.0

Table 11 *Results of overall answering time analysis on three data sets [C, B2, B3]*

For all three data sets, the percent of flagged respondents is similar, regardless of not changing time for dataset B2 to a value lower than determined by participants. The value was not changed because the distribution of OAT data showed a clear point below which there is a significant decrease in the number of observations having a similar answering time – and this happens to be about 16:40 minutes (6.91 in natural logarithm value, see Attachment 4), which is why this value was deemed an accurate cut-off point, while for data sets C (distribution also in Attachment 4) and B (**Figure 4** below) there was no clear cut-off point.

For data set B3, there were two groups (as described earlier).

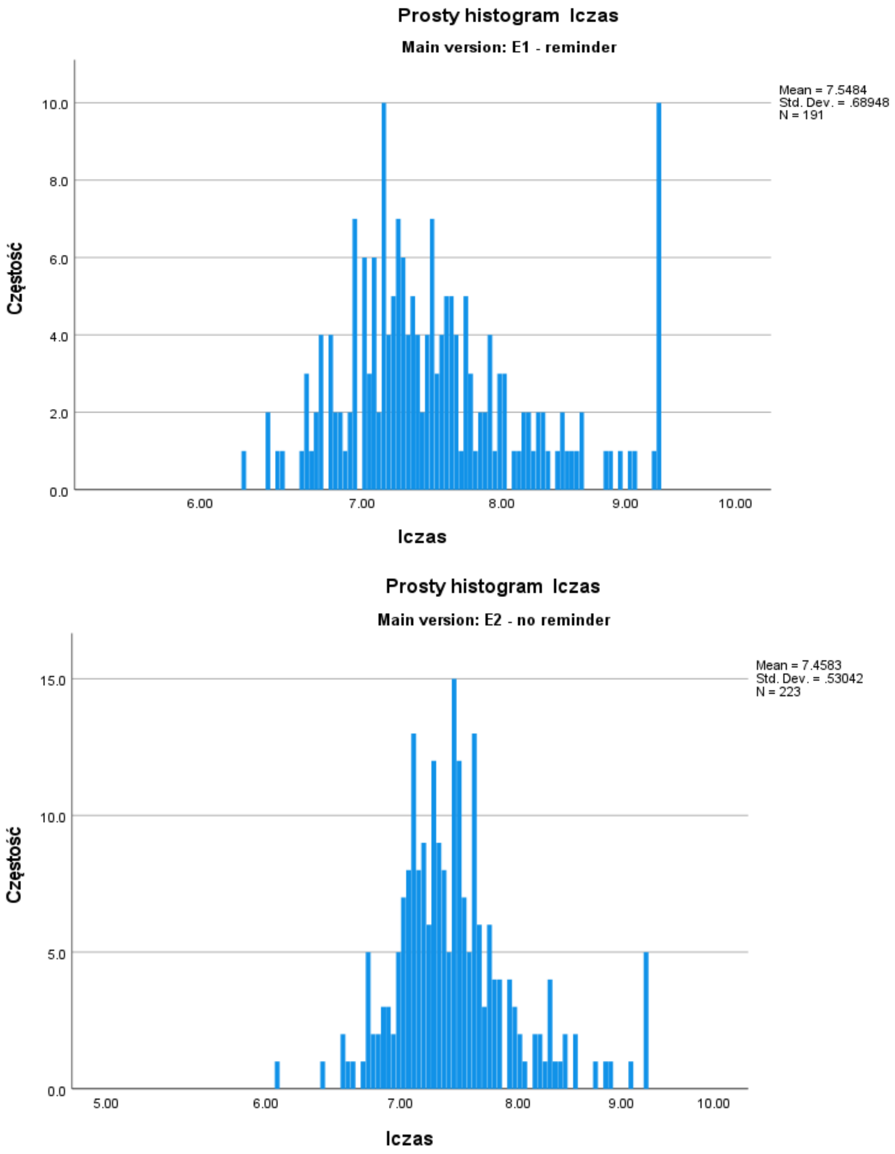


Figure 4 *Distributions of the natural logarithm of time in seconds, separated according to the experimental group, data set B3*

Although the mean time for the first group is higher than for the second group (the difference is significant), the reason for this may be different contents of questions presented to each group and the fact that an attempt to make both versions similar in terms of length – this attempt may have been unsuccessful. However, excluding outliers from both groups caused the difference to become insignificant, to which possible explanation is that the reminders used in the first group may have increased time spent on it, as they did not allow to simply proceed to the next question, and some respondents may have been possibly annoyed by them, taking longer to complete the survey (leaving it open and coming back later).

Partial answering time – test of threshold based on positional measures (median, interquartile range)

Using positional measures as a way of determining too low answering time was tested on data set B3.

In **Table 12** below, ten blocks are the main parts of the survey with a typical, closed rating scale (blocks of questions requiring additional actions like typing were excluded from the analysis to avoid the influence of different skill levels in typing, which inevitably influenced time spent on answering). The minimal time required to answer was determined as the interquartile range (IQR) subtracted from the median value (to avoid the influence of outliers).

Block number	Number of questions	Median [sec]	IQR [sec]	Minimal time [sec]	Respondents below minimal time [%]
1	6	65.06	44.43	20.63	0.2
2	5	41.31	27.11	14.2	0.2
3	10	31.85	21.29	10.56	0.5
4	5	64.00	36.07	27.93	1.0
5	9	116.73	65.83	50.9	1.0
6	4	42.56	20.50	22.06	2.2
7	4	55.45	34.91	20.54	0.2
8	8	97.55	55.04	42.51	0.2
9	3	39.17	29.39	9.78	0.5
10	5	44.19	22.89	21.3	2.7

Table 12 *Percent of respondents below cut-off value of time in seconds for ten blocks of questions*

The majority of respondents, 94.9%, did not speed through any block of questions (98.8% in the E1 group and 97.1% in the E2 group). In **Figure 5** percent of respondents who speeded through a given number of the block is shown.

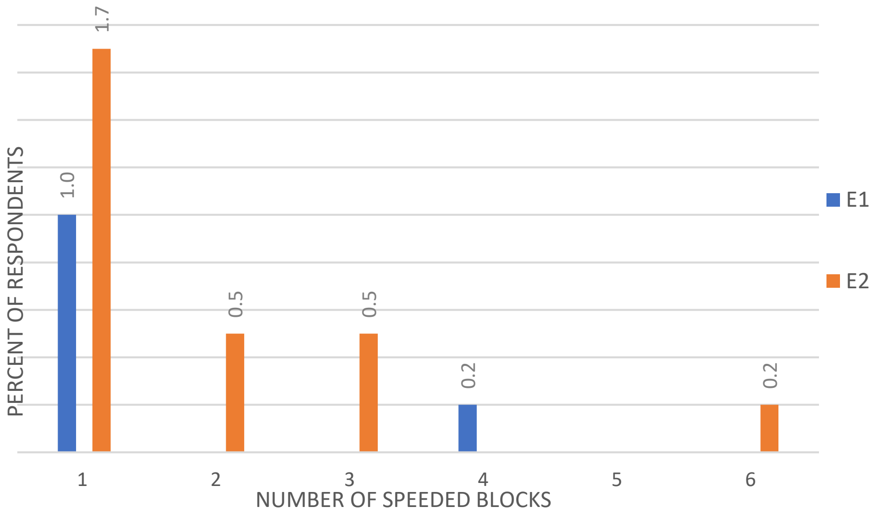


Figure 5 *Distribution of respondents that had too fast reading speed, data set B3*

In each group, there was one respondent who sped through all blocks (4 in E1 and 6 in E2), which means that in total, two respondents were flagged by this sub sign. If there was a need for analysis to be done on questions from a particular block, respondents should be excluded locally if they speeded through a given block.

Distributions of partial answering times for ten blocks are included in Attachment 5.

2.6 Warning sign #2 – Attention check questions

In seven data sets [A2, C, B2, B3, B4, B5, B6], attention check questions had a form of simple arithmetic questions (i.e., *21-4*, *15+7* etc.) with five possible answers to choose from (one of the answers was always correct). In these datasets respondents were flagged if the answer to attention (arithmetic) check question was incorrect. In data set B3, only one experimental group of the respondents could make an error in attention check questions (contents questions can be found in Attachment 6) – because there was an experimental manipulation that did not allow one group to proceed to the next question if the incorrect answer was chosen. Therefore, analyses on frequency of attention check questions errors were performed for the group that was allowed to make errors.

In data set A1, attention check question type was different – three attention check questions were instructed response items (see Figure 6).

In three cases [data sets A1, C, B2], the lenient criterion of flagging – allowing for one error in attention check questions - was used, although the percent of respondents flagged by strict criterion – no errors allowed – was also calculated. The decision that using a lenient criterion is better was based on the difficulty of questions [data set A1, see **Table 13**] or survey software not allowed to go back to previous question [data sets C, B2]. In five data sets [A2, B3, B4, B5, B6] going back to correct a mistake was allowed, and accounting errors in attention check questions as mistakes were much less justifiable.

Data set	Study version (group)	N	Number of attention check questions in a data set	Respondents flagged by strict criterion		Respondents flagged by lenient criterion	
				[count]	[%]	[count]	[%]
A1	-	1421	3	609	42.9	445	31.3
A2	1	749	5	47	6.3	13	1.7
	2	748	6	73	9.8	20	2.7
	Total	1487	-	120	8.0	33	2.2
C	-	287	11	45	15.7	22	7.7
B2	-	341	11	21	12.0	9	2.6
B3	2	223	5	5	2.2	0	0.0
B4	-	308	14	18	5.8	1	0.3
B5	-	140	13	5	3.6	1	0.7
B6	-	497	3	1	0.2	1	0.2

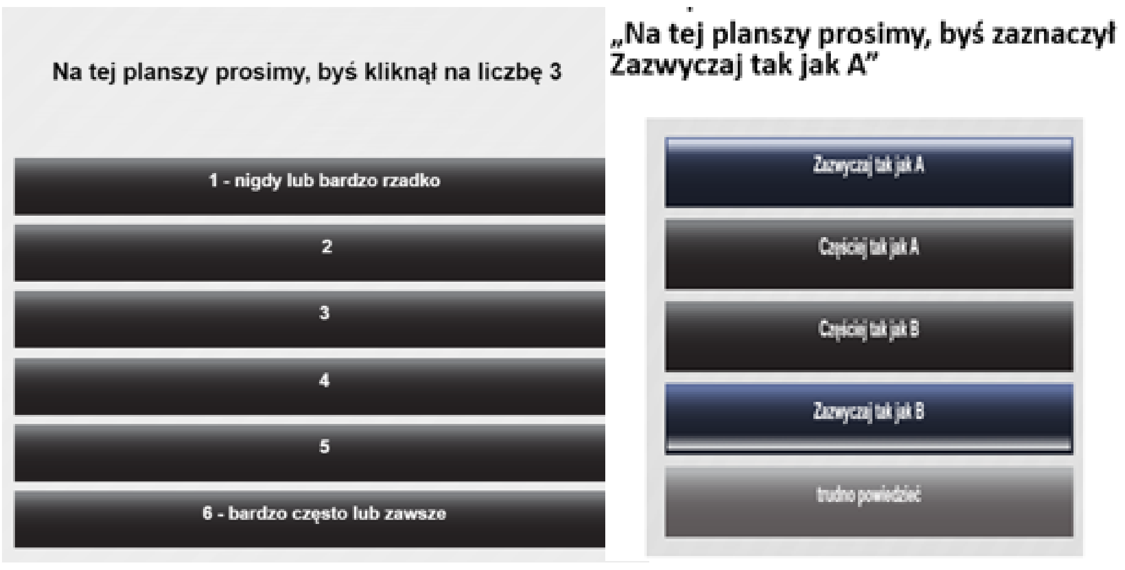
Table 13 *Results of errors in attention check question analysis on eight data sets [A1, A2, C, B2, B3, B4, B5, B6]*

Due to the fact that excluding respondents globally on the basis of an incorrect answer to any of the attention check questions would exclude 42.9% of the respondents in the case of data set A1 (see **Table 14** below), it was decided (as mentioned above) that a lenient criterion would be applied in the case of this sign. It allowed reducing the size of the excluded group to 31.3% of the sample.

Type and number	Answered correctly	Answered incorrectly
AC1	77.69%	22.31%
AC2	64.81%	35.19%
AC3	69.32%	30.68%

***Table 14** Percent of correct and incorrect answers to attention check questions, data set A1*

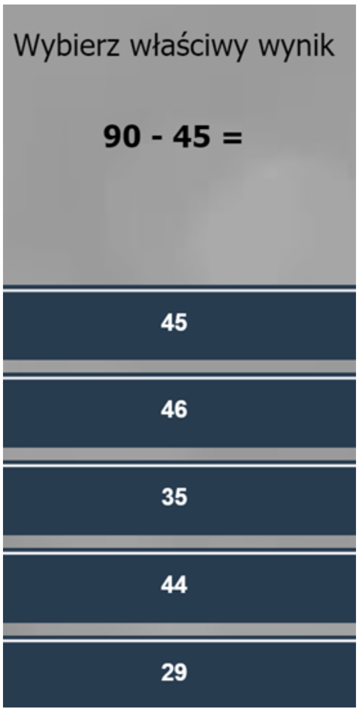
In **Figure 6** below, an example of how attention check questions looked like in data set A1 is presented. In **Figure 7**, an example of an arithmetic question used in data set C and B2 is presented. The remaining data sets were collected using different software, and an example question in these cases is shown in **Figure 8**.



Na tej planszy prosimy, byś kliknął na liczbę 3

„Na tej planszy prosimy, byś zaznaczył Zazwyczaj tak jak A"

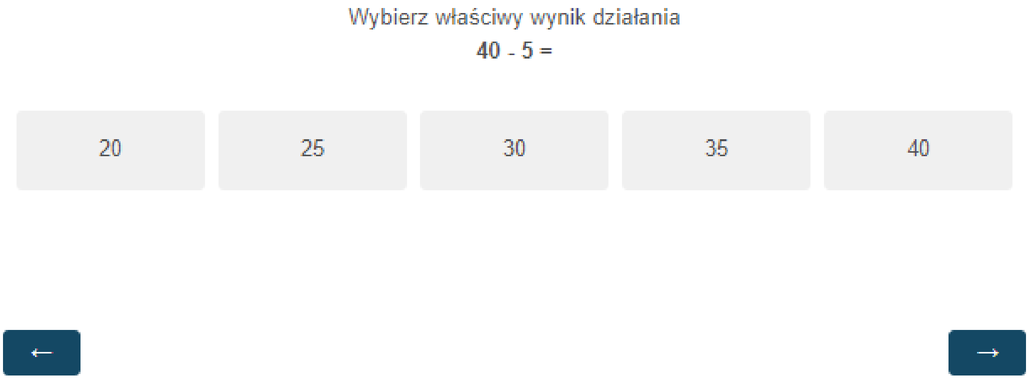
***Figure 6** Examples of AC questions used in study A1*



Wybierz właściwy wynik

90 - 45 =

***Figure** 7 Example of arithmetic question used in survey C*



Wybierz właściwy wynik działania

40 - 5 =

***Figure 8** Example of arithmetic question used in data sets A2, B3, B4, B5, B6*

2.7 Warning sign #3 – Non-informative answers and rating style

The third warning sign consists of two sub signs:

1. how many non-informative (ex. *Hard to say*) answers respondents gave in questions that had this answer option explicitly available to choose, and
2. if they were under the threshold for minimal variance across a series of questions with the same rating scale.

Thresholds for how many non-informative answers were accepted was 50% for seven datasets [A1, A2, C, B1, B2, B4, B5], 55% for data set B6 (because of small and uneven question number), and about 30% for data set B3. Thresholds for non-differentiaition rating style (too low variance) were determined individually for each dataset as two standard deviations below the mean for each series of questions with the same rating scale.

In the A1 study, the record-holders (86 respondents, 6.1% of the sample) chose a non-informative answer for all questions in the series.

In data set D, there were many blocks of questions with the same scale, but for most of them, answering consistently with the same answer could be considered a true state of matters (unlikely, yet possible), so only one block of questions was chosen (exposure to different stressing conditions at work), which consisted of a series of 18 questions. Value of 7 (‘Never’) was excluded - the contents of the questions and explanations as to why the value needed to be excluded can be found in Attachment 7.

2.7.1 Non-informative answers

The non-informative answers part of the sign was not used in this analysis, as these answer options were not part of the rating scale presented by the interviewer – they were coded only if given spontaneously by the respondent, which decreases the chance of choosing such answer significantly.

In the E1 and E2 data sets, ten questions about Schwartz values (content and the rating scale can be found in Attachment 8 were decided to be a criterion of local exclusion[167]. The analyses were done only on the Polish sample (N1=1000, N2=966) to allow for comparison with online data sets.

Data set	Number of questions in series	The threshold for non-informative answers	Respondents flagged	
			[count]	[%]
A1	30	15 (50%)	118	8.3
A2	40	20 (50%)	45	6.0
	61	30 (≈49%)	45	6.0
	Total	-	90	6.0
C	80	40 (50%)	2	0.7
B1	59	30 (≈50%)	2	0.3
B2	80	40 (30%)	4	1.2
B3	26	8 (≈31%)	2	1.0
	28	9 (≈32%)	3	1.3
	Total	-	5	1.2
B4	136	64 (50%)	1	0.3
B5	136	64 (50%)	0	0.0
B6	11	6 (≈55%)	1	0.2

Table 15 *Results of non-informative answers analysis across nine data sets [A1, A2, C, B1, B2, B3, B4, B5, B6]*

As is shown in **Table 15**, again, in the paid panel, respondents' rates of choosing non-informative answers are higher than in student samples. Data set C is an interesting case here because it is also an employee sample (like A1 and A2), but participant source was a snowball method – and this seems to make respondents more willing to put more cognitive effort instead of just choosing an easy way out by choosing 'Hard to say' option.

2.7.2 Low differentiation rating style

In **Table 16**, the results of variance analysis are presented. In the case of data sets D, E1 and E2, the threshold of 2SD were negative, so this part of the sign was not used in this way – only respondents with variance less than 0.15 were flagged – because of variance equal to 0.1 corresponding to a string of the same numbers with answers to just one out of ten questions changed.

[167] These values were used directly in the analyses for the other dissertation

Data set	Number of questions in series	Threshold for variance	Respondents flagged by variance	
			[count]	[%]
A1	30	0.21	126	8.9
A2	40	0.60	25	3.3
	61	0.57	26	3.5
	Total	-	51	3.4
C	80	0.44	1	0.3
B1	59	0.60	4	0.5
B2	80	0.56	1	0.3
B3	26	0.38	1	0.5
	28	0.63	0	0.0
	Total	-	1	0.2
B4	136	0.82	3	1.0
B5	136	0.46	3	2.1
B6	11	0.50	3	0.6
D	18	=0	44	3.7
E1	10	0.15	28	2.8
E2	10	0.15	30	3.1

Table 16 *Results of low differentiation rating style analysis across nine data sets [A1, A2, C, B1, B2, B3, B4, B5, B6]*

In the case of too low variance, student samples turned out to be more differentiating (mean % flagged equal to 0.78) than offline samples (mean % flagged equal to 2.1), and both of these types had more differentiation in answering than panel samples (mean % equal to 6.15). Panel samples results are consistent with other signs – flagging the highest percent of respondents.

2.8 Warning sign #4 – Declarative cooperation, logical consistency and open-ended questions

Almost all data sets contained some form of declarative cooperation questions – in the case of eight online data sets [A1, A2, B2, B3, B4, B5, B6], contents of three questions were the same, in the case of three data sets [D, E1, E2], cooperation was assessed by the interviewer. Study B1 did not contain any such questions.

Logical consistency was possible to check for eight data sets [A1, A2, C, B1, B2, B4, B5, B6]. Respondent was considered to be inconsistent when in question pair if they chose answers that were entirely not consistent (i.e. chose person A in one question and person B in the other question).

Open-ended questions were a part of seven data sets [A1, A2, C, B1, B4, B5, B6]. Respondent was flagged when theirs answers to those questions were not what was expected in a specific study.

For online data sets, too low a level of declarative cooperation was determined by answers to two questions

How do you rate the degree of your involvement in this task?
1 - very low, 2, 3, 4, 5, 6 - very high

If you were to participate in the survey again (e.g. tomorrow), would your answers be:
1. identical
2. could be slightly different
3. could be diametrally different

to be an answer to the first question

- 2 or less for data sets A1, A2, C[168]
- 3 or less for data set B1 to B6[169]

and/or answer to second question equal to 3 – *could be diametrally different.*

For the data set collected using face to face interviews method, cooperation and interest were assessed by the interviewer. In the case of data set D, there were two questions

[168] These samples were expected to be less engaged, as they were employees taking survey to earn points [A1, A2] or asked to take part by students [C].

[169] There were student samples, and students, in most cases, could get bonus points for their respecitve courses, so they should be engaged.

(about cooperation and how often respondents asked for clarification), in case of data set E1 and E2 there was one question (about respondent's interest during interview). For data set D, respondents who were assessed to cooperate poorly or very poorly and / or were asking for clarification most of the time or always were flagged. For data sets E1 and E2, respondent who were not interested during interwiev were flagged.

Six studies [A1, A2, C, B1, B4, B6] had open-eneded questions (respondents free to type whatever they want). Answers to these questions were coded into four (if question was obligatory) or three categories:

- no answer (if question obligatory)
- informative response (in terms of content),
- non-informative answer (e.g. typed characters that do not form any words, jokes, meaningless clusters of words, answers 'I don't know', 'hard to say', 'I don't have an opinion', etc.)
- answer is too short (e.g., one or two words, relevant to the question or not, only if longer answer expected).

2.8.1 Declarative cooperation

In **Table 17** the results of analysis for eight online data sets are presented.

Data set	Respondents who were not engaged		Respondents whose answers would be different		Total	
	[count]	[%]	[count]	[%]	[count]	[%]
A1	185	13.0	227	16.0	372	26.2
A2	20	1.3	45	3.0	62	4.1
C[170]	3	0.8	10	3.9	13	4.5
B2	23	7.0	5	1.5	27	7.9
B3	10	2.4	4	0.9	13	3.1
B4	5	1.6	2	0.6	7	2.3
B5	5	3.6	3	2.1	4	2.9
B6	17	3.4	3	0.6	19	3.8

Table 17 *Results of declarative cooperation analysis in eight data sets [A1, A2, C , B2, B3, B4, B5, B6]*

[170] N=257 for this analysis because of missing data fo 30 respondents.

Besides data sets A1 and E1, declared (assessed) cooperation tends to be rather high – in single digit order of magnitude. This is probably the sign of social desirability bias – respondents do not want to admit that they were not as diligent as they should be.

Table 18 contains results of analysis for offline data sets. As these have different operationalization (assessment by the interviewer instead of self-assessment), they should be a more accurate measure of attention.

Data set	Respondents who had low level of cooperation / were not interested		Respondents who asked for clarification most of the time or always		Total	
	[count]	[%]	[count]	[%]	[count]	[%]
D	61	5.1	16	1.3	69	5.7
E1	132	13.2	-	-	132	13.2
E2	55	5.7	-	-	55	5.7

Table 18 *Results of declarative cooperation analysis – data sets with cooperation assessed by interviewer*

There is no difference between data set D and E2 in terms of flagging rate despite data set D being based on two questions instead of one. Data set E1, however, has more than double flagged rate than both E1 and D. As the operationalization of too low cooperation was not changed (low interest during interview), this may be caused by many things, but two most probable are (1) change in the way interviewers were instructed to assess respondents' cooperation or/and (2) change in respondents behaviours towards interviewers in 5 year time gap between two waves of the survey. It is also possible that respondents became more interested in the interview, or the survey was conducted differently, but information about the mode used in Polish sub sample is missing from study documentiation. Therefore, explaining this difference was not possible in this case.

Summarised results for all countries are presented in Attachment 9.

2.8.2 Logical consistency

For logical consistency check, a pairs of questions (presented in **Table 19**) that should correlate in a certain way according to theory, were chosen to be compared. More question can be chosen if possible.

In case of studies A1, A2, C and B2 two pairs of questions were analysed – pair of questions from the same scale, but not identical, and pair of questions that were identical, with stwitched characteristics of person A and person B. In datasets B4, B5, and B6 the same pair of questions was used.

Data set	Question pair
A1	*Person A starts the task only after he has thought out **exactly how to perform it**. Person B starts the task even when he does not know exactly how to do it and is counting on ideas to come in the process.* *Person A starts writing an essay without having an exact vision of what he will write. Person B first creates a **mental vision of what he wants to write**, and only then begins to write.*
A2	*Person A **usually directs the course of the conversation with others**. Person B is often silent in the company of other people.* *Participants in a business dinner or social gathering may believe that A has **dominated the conversation**. Of person B, they think she said little and that it was the others who had to keep the conversation going.*
C	*When doing teamwork, person A feels **best in the role of a leader**. Person B is not bothered when someone else decides how to carry out the team's tasks.* *Person A in the group will **gladly play the role of the leader**. Person B if he can prefer to avoid the responsibility of being a leader.*
	*Person A thinks she has **many good qualities**. Person B feels she has little to be proud of.* *Person A feels she has little to be proud of. Person B thinks she has **many good qualities**.*
B1	*Person A feels she **has little to be proud of**. Person B thinks she has many good qualities.* *Person A sometimes **thinks she's useless**. Person B is generally pleased with herself.*
B2	*When doing teamwork, person A feels **best in the role of a leader**. Person B is not bothered when someone else decides how to carry out the team's tasks.* *Person A in the group will **gladly play the role of the leader**. Person B if he can prefer to avoid the responsibility of being a leader.*
	*Person A thinks she has **many good qualities**. Person B feels he has little to be proud of.* *Person A feels she has little to be proud of. Person B thinks she has **many good qualities**.*
B4	*Person A **could possibly achieve more**, but sees no reason to try more than necessary. Person B works more than other people.* *Person A **at work often exceeds his abilities**. Person B has no exaggerated ambitions, preferring a quiet and comfortable life.*
B5	
B6	

***Table 19** Question pair used in logical consistency check in eight data sets [A1, A2, C, B1, B2, B4, B5, B6]*

Results of comparing answers to these question pairs are presented in **Table 20** below.

Data set	Pair of questions	Respondents who gave inconsistent answers	
		[count]	[%]
A1	1	366	25.8
A2	1	371	24.8
C	1	41	14.3
	2	26	9.1
B1	1	135	18.2
B2	1	40	11.7
	2	41	12.0
B4	1	64	20.8
B5	1	17	12.1
B6	1	79	15.9

Table 20 *Results of logical consistency analysis across eight data sets [A1, A2, C, B1, B2, B4, B5 , B6]*

Rates of inconsistent answers range between 9.1% for one pair of questions in data set C to 25.8% in data set A1. The highest rates are, again, present in commercial panel samples: A1 and A2.

2.8.3 Odd answers to open-ended questions

Respodents were assessed on their behaviour regarding open-ended questions for the study as a whole – meaning that no answer (if required) or non-informative answer to any open ended question is treated as a reason for flagging a respondent. Number of questions depended on survey and is given in **Table 21** below. Refusals are not included in analysis. For data set A2 open-ended questions were optional, therfore no answer is not considered a reason to flag a respondent.

Data set	Number of open-ended questions	No answer		Non-informative answer		Too short answer		Informative answer	
		[count]	[%]	[count]	[%]	[count]	[%]	[count]	[%]
A1	3	315	22.2	116	8.2	136	9.6	868	61.1
A2	3	647	43.2	65	4.3	2	0.1	270	18.0
C	1	0	0.0	1	0.3	33	11.5	253	88.2
B1	2	304	41.4	4	0.5	-	-	430	58.1
B4	2	78	25.3	0	0.0	-	-	226	73.4
B6	2	18	3.6	9	1.8	-	-	471	94.8

Table 21 *Results of open-ended questions analysis across six data sets [A1, A2, C, B1, B4, B6]*

Percents may not add to 100 because of different combinations of categories for particular questions present – full statistics with division by question are presented in Attachment 10.

For the purpose of this analysis, all open ended questions have been coded but the author. The data in 'too short answer' for the last three data sets is missing because open-ended questions used in those data sets were were a short answer questions, and one-word answers were coded as informative answers. For data sets A1, C, B1, B4, B6 questions were obligatory, so 'no answer' category should count as lack of behavioural cooperation, along with 'non-informative answer', but for data set B1 this may actually not be the case, as some respondents had not had two open ended questions shown to them (this depended on filter question about whether they had a job currently), so in this case 'no answer' category was not used for flagging. For data set A2 open ended questions were not obligatory, so only 'non-informative answer' category means that respondent was flagged.

Rates of non-informative answering to open-ended question are not high, but this result is partially caused but the nature of the questions – for all data sets but A2 the questions were meant to be short answer type, meaning that it is difficult to assess whether a particular answer is substantial enough to be considered informative – and the author has chosen a safe approach to the matter, coding only gibberish answers as non-informative (ex. 'sfjehs3', '?', etc.).

2.9 Test of Flexmix analysis as a method of detecting FALSE respondents

Comparing pairs of questions that should correlate with each other.
It is worth having ‘twins’ questions - because they allow for additional tests.

The FLEXMIX (finite mixtures of generalized linear regressions) analysis allows to divide the respondents according to their fit to different regression lines. It is therefore an iterative program combining regression analysis and cluster analysis.

The research task was to check the usefulness of using this analysis method to detect FALSE respondents.

It was assumed that the respondents classified by the algorithm to the cluster with a positive correlation are attentive respondents (in the case of questions that should correlate negatively, the values were reversed so that the expected correlation was positive). Others are suspected of not being mindful in answering these questions.

Since the number of clusters depends on the person conducting the analysis, 3 cluster solutions were checked first.

Using the flexmix model in R, the correlations between the question sr14 and sr20, coming from the SSA methodicality index, were checked.

sr14 ***Person A starts the task only after he has thought out exactly how to perform it.*** *Person B starts the task even when he does not know exactly how to do it and is counting on ideas to come in the process.*

sr20 Person A starts writing an essay without having an exact vision of what he will write. ***Person B first creates a mental vision of what he wants to write, and only then begins to write.***

The analysis of the content of the questions shows that they should be negatively correlated, therefore the question sr20 was inverted in such a way that the expected correlation was positive.

The classification into 3 groups indicated 2 groups with a positive correlation and one distinct group with a negative correlation.

Cluster	Correlation coefficient	Cronbach's Alpha	N of items
1	-1.00	-2.841	3
2	0.17	0.317	3
3	1.00	0.824	3

***Table 22** Reliability statistics for 3-cluster solution for methodicality index, data set A1*

Two clusters have the correct correlation, also confirmed by the correct Alpha values. Respondents with a negative correlation constitute 32.1% of the sample (N = 456), with a weakly positive correlation (coefficient 0.17) - 30.8% of the sample (N = 437), and with a strongly positive correlation - 37.2% of the sample (N = 528).

It was decided to combine two clusters with positive correlation (as both are correct).

Cluster	Cronbach's Alpha	N of items
Positive correlation	0.631	3
Negative correlation	-2.841	3

***Table 23** Reliability statistics for 2-cluster solution for methodicality index, data set A1*

After combining the 2 clusters with a positive Cronbach's alpha, the combined group decreased, however, the attentive respondents represent 67.9% of the sample based on this indicator.

For the excluded (N = 456) Cronbach's alpha is minus 2.841 (which is an absurd value), for the attentive ones (N = 965) it is 0.631.

Similarly, for data set A2 the relationship between the sr2 and sr10 question coming from the SSA extraversion index was checked.

The analysis of the content of the questions shows that they should be positively correlated.

sr2 ***Person A usually directs the course of the conversation with others.*** *Person B is often silent in the company of other people.*

sr10 Participants in a business dinner or social gathering may believe that ***A has dominated the conversation****. Of person B, they think she said little and that it was the others who had to keep the conversation going.*

The classification into 3 clusters revealed 2 groups with a positive correlation and 1 group with a slightly negative correlation close to zero, which divided the respondents into those who generally had a positive correlation between the answers to both questions, and those who had a correlation close to zero.

Below (**Table 24**) is a comparison of the correlations for questions from the extraversion scale in the two groups determined by the model after combining the groups with a positive correlation. The expected correlations (based on the content of the questions) are generally positive, with the exception of question sr6, which should correlate negatively with all others.

	sr6		sr10		sr12	
	false	attentive	false	attentive	false	attentive
sr2	-0.262	-0.385	0.050	0.800	0.192	0.328
sr6	-	-	-0.182	-0.383	-0.148	-0.257
sr10	-	-	-	-	0.213	0.369

Table 24 *Correlations between items from extraversion scale from A2 data set for FALSE respondents flagged by flexmix and for attentive respondents separately*

As shown in above **Table 24**, correlations in attentive respondent's group are stronger than in FALSE respondents group, when division is based only on the results of the flexmix model.

The analysis of the reliability of the extraversion scale for those identified by flexmix showed that in the group of attentive respondents Cronbach's Alpha is higher than in the group of FALSE respondents.

Group	Cronbach's Alpha	N of items
FALSE	0.456	4
Attentive	0.737	4

Table 25 *Reliability statistics for 2-cluster solution for extraversion index, data set A2*

For FALSE respondents (N = 509) Cronbach's alpha of 4 questions from the extraversion scale was 0.456, for attentive respondents (N = 986) it was 0.737.

Comparison of FALSE respondents flagged by flexmix with FALSE respondents flagged by warning signs is shown in **Table 26** below.

	Data set A1			Data set A2		
	WS1-4 – attentive	WS1-4 – FALSE	Total	WS1-4 – attentive	WS 1-4 – FALSE	Total
Flexmix – attentive	34.34%	33.57%	67.91%	54.45%	11.51%	65.95%
Flexmix – FALSE	11.33%	20.76%	32.09%	28.03%	6.02%	34.05%
Total	45.67%	54.33%	100.00%	82.47%	17.53%	100.00%

Table 26 *Comparison of percent of respondents flagged by flexmix and WSs, data sets A1 & A2*

In general, flexmix flagged fewer respondents than warning signs (32.09% vs 54.33%) in data set A1, and more respondents as FALSE than warning signs (34.05% vs 17.53%) in data set A2. Both flexmix and warning signs agree in about 55% and 60% of cases whether a respondent is FALSE or not.

2.10 Consequences of ignoring the problem - reliability for groups of FALSE and attentive respondents

Test of impact of not excluding FALSE respondents was done on two data sets: A1 and A2.
For data set A1, reliability statistics (Cronbach's Alpha) for 3 items from the methodicality scale was calculated for two groups separately: attentive and FALSE respondents. In this analysis, 4 signs were used to identify FALSE respondents, by lenient criterion of exclusion.

Group	Cronbach's Alpha	N of items
FALSE	-0.380	3
Attentive	0.550	3

Table 27 *Reliability statistics for methodicality scale from A1 data set, in groups of FALSE and attentive respondents based on four signs*

In the case of FALSE respondents (n = 652), Alpha is -0.380, and in the case of attentive respondents (n = 769), 0.550.

For data set A2, reliability statistics for 4 items from the extraversion scale, the reliability statistic was calculated the same way as for dataset A1.

Group	Cronbach's Alpha	N of items
FALSE	0.288	4
Attentive	0.717	4

Table 28 *Reliability statistics for extraversion scale from A2 data set, in groups of FALSE and attentive respondents, based on four signs*

For the same items from the extraversion scale, in the case of FALSE respondents (n = 261), Alpha is 0.288, and in the case of attentive respondents (n = 1233), 0.717.

2.11 Summary and Discussion of the results

The empirical part of this dissertation begins with defining the operationalisation of 4 warning signs. This part ends with a **description of the FR procedure for detecting FALSE respondents** and the comparison of 2 groups of respondents.

To remind – there were 4 Warning Signs [WS], tested on 12 data sets:

- WS1 (TIME) is based on a combination of all types of answering time analysis.
- WS2 (ATTENTION TEST) is based on the number of errors in attention check questions placed in the survey to directly check the respondent's attention to the content of the question.
- WS3 (VARIANCE) is based on the analysis of the number of DK (Don't Know, non-informative, empty) answers and the respondent's rating style.
- WS4 (LOGICAL) is based on measures of respondent's behavioural (logical consistency, both in closed and open-ended questions) and declarative engagement.

The distribution of warning signs was analysed in nine datasets from the author's research[171]:

- two data sets consisting of commercial panel users: A1 (1421 employees) + A2 (1497 employees)
- six data sets B1- B6 based on responses from 2399 participants who, in the overwhelming majority, combine studies at the Faculty of Management with professional work
- one data set C, based on responses from 287 employees with at least three years of work experience

and 3 pre-existing data files:

- Data set D, European Working Conditions Survey, personal interviews, 1203 Polish employees

[171] WS1 tested on 9 datasets, WS2 tested on 8 datasets, WS3 and WS4 tested on 12 datasets

- Data sets E1 + E2, World Values Survey, two waves (5+6), 1000 + 966 Polish respondents.

Four Warning Signs were computed for all participants, so the subsequent decision is about which criterion (strong vs lenient) of exclusion should be used.

The use of strict criterion means the **GLOBAL** exclusion of the respondent (we remove them from the entire data set). It is mostly justified when we can detect **global inattention,** when the respondent 'plays' with the survey and does not pay attention to the answers given. Such a respondent can be easily spotted during the interview; however, it is challenging to detect them during an online survey.

The use of a lenient criterion means **LOCAL exclusion.** If we suspect that respondents speeded or lost attention in some of the questions, we turn their answers into missing data only in these questions. An interesting example comes from the study conducted by us in a big company. There was a large amount of missing data in the responses to the question about the year of birth. Employees were afraid that their year of birth would allow to identify them and the information would be delivered to the employer – assertion that this is ANONYMOUS research conducted by the University of Warsaw has not reduced their identification anxiety.

Comparing the year of birth (in cases where the respondent answered) with the years of seniority in the company revealed many logical contradictions (e.g., seniority indicating that somebody started to work in the company at 6 years old). Missing data is a much smaller threat to the accuracy of analyses than false data. Suppose that we flag such respondents as globally FALSE and excluded from the data. If that is the case, we can lose valuable respondents who were very honest in their answers as they claimed in post-survey interviews. By delivering false information about the year of birth, they try to protect themselves from identification. We have to emphasize that the identification anxiety was utterly unfounded. Even if we had tried very hard, we would still not be able to identify the respondent.

Therefore, logical inconsistency between age and seniority should cause LOCAL exclusion (converting age value into missing), not global one.

The local vs. global decision could also depend on the length of the survey, sample size, etc. Longer surveys tend to have different levels of respondent engagement. For some respondents, lower at the end because of the limitation of cognitive resources. However, it is not necessarily linear; different parts of the survey can be more or less attractive to some respondents.

Suppose that particular blocks differ in terms of respondent engagement. It may be rational to use a lenient criterion, allowing the respondent to be flagged by one of the WARNING SIGNS and not exclude them from further analyses. For example, errors in attention check questions can be 'forgiven', especially at the end of the survey.

It is reasonable to retain more respondents in smaller samples, as a minimal sample size is often required to use specific statistical tests.

For example, in the case of WS1 (answering time), the strict exclusion criterion assumes that respondents flagged as suspects by any partial (for one block of items) anwering time are excluded; the lenient criterion allows for being flagged as one of the subsigns.

In summary: the decision regarding exclusion criteria can be different for different warning signs.

Three offline data sets: The World Values Survey (Wave 5 & 6) and the European Working Conditions Survey should have been cleaned up, because they were carefully prepared by an international team of researchers, not publicly available on the Internet or for anyone wishing to participate.

Many surveys are posted on Facebook or other publicly available platforms; anyone can participate, regardless of their motivation to do so. Even if they are motivated with some rewards, in most cases researchers do not check attentiveness of their respondents, which may lead to a high number of FALSE respondents in data collected that way.

The internet surveys analysed in this dissertation were carefully prepared and conducted in the Managerial Psychology and Sociology Unit, and therefore these issues were taken into account. The students were motivated with bonus points and threatened with the algorithm, and yet, many of them were still rejected by the algorithm. We also care about giving respondents a way to avoid answering – respondents almost always have the 'It's difficult to say' option available on the response scale.

In the next section, I will summarise the completion of the results of the research tasks.

2.11.1 Research task #1: FALSE respondent scope

The first research task was to determine the scope of respondents' inattention.

The percentage of respondents flagged as "FALSE" depended on the survey (see Table 29 below).

Data set	Group	WS #1 – Answering time	WS #2 – Attention check questions	WS #3 – Rating style	WS #4 – Declarative, logical, open-ended
A1		56.4	31.3	11.0	48.1
A2		26.1	2.2	6.0	30.3
C		8.7	7.7	0.7	25.4
B1		4.1	Nd.	0.8	18.2
B2		8.5	2.6	5.3	21.4
B3	E1	10.5	Nd.	1.6	4.2[c]
	E2	6.3	2.2[b]	1.3	2.2[c]
B4		3.2	5.8	1.0	20.8
B5		7.1	3.6	2.1	20.0
B6		10.1	0.2	0.6	17.3
D		Nd.	Nd.	3.7	5.7
E1		Nd.	Nd.	2.8	13.2
E2		Nd.	Nd.	3.1	5.7

Table 29 *Percent of respondents rejected by warning signals in the analysed sets*

In Table 30 below, percentages are calculated separately for lenient and strict criteria of exclusion:

The strict criterion means that respondents flagged by any of the four warning signs were excluded so that it would case GLOBAL exclusion.

The lenient criterion could allow for LOCAL exclusions, for example, inconsistency between age and seniority and too many DK answers in one block of the survey. We can accept **local inattention** when the respondent becomes lost in thought, pondering, or deliberately ignoring a specific block of questions, but answers others with due diligence.

In my dissertation, a lenient criterion means that the respondents can be flagged by one warning sign – excluded are those who are flagged by 2, 3, or 4 warning signs.

Data set	Year	N	Sample	% of respondents excluded	
				Lenient criterion	Strict criterion
A1	2018	1421	Panel respondents, employees, paid	45.9	71.0
A2	2021	1497		14.2	45.4
C	2020	287	Employees, convenience sample	6.6	33.8
B1	2018	740	Participants who, in the overwhelming majority, combine studies at the Faculty of Management with professional work, rewarded by extra points in their study	1.2	21.9
B3	2021	414		2.1	13.8
B2	2020	341		2.9	26.1
B4	2021	308		3.2	27.6
B5	2021	140		5.0	22.1
B6	2021	497		2.8	25.2
D	2015	1203	EWCS, offline, personal interviews	0.4	8.9
E1	2005	1000	WVS, offline, personal interviews	0.9	15.2
E2	2010	966	WVS, offline, personal interviews	0.0	7.3

***Table 30** Percent of excluded respondents in each data set depending on which criterion was used*

The highest exclusion percentage is for paid panel data, and the lowest is for the offline personal interviews. Comparing it with the data reported in the literature (see Table 2 in the literature review), we can say that the range of FALSE response rates seen in previous research is between 4 and 97.8%, while in my analyses it varied from 7.3 to 71.0%.

Table 31 below shows the percentage of respondents flagged by one or more warning signs.

The results are split by warning sign combinations.

Combination of warning signs	A1	A2	C	B1	B2	B3[a]	B4	B5	B6	D	E1	E2
1 time	**56.4**	**26.1**	8.7[b]	4.1	8.5[b]	6.3[b]	3.2	7.1	10.1	-	-	-
2 test	**31.3**[c]	**2.2**	7.7	-	2.6	2.2	5.8	3.6	0.2	-	-	-
3 rating	**11.0**	**6.0**	0.7	0.8	5.3	1.3	1.0	2.1	0.6	3.7	2.8	3.1
4 logical	**48.1**	**30.3**	25.4	18.2[d]	21.4[e]	2.2[f]	20.8	14.3[e]	17.3	5.7[g]	13.2[h]	5.7[h]
1 & 2	29.9	1.9	2.4	-	1.8	0	0	2.1	0	-	-	-
1 & 3	10.2	3.6	0	0.4	0.6	0	0.3	0	0.2	-	-	-
1 & 4	34.6	11.7	3.5	0.8	4.7	0.4	1.0	2.1	2.4	-	-	-
2 & 3	7.8	0.3	0	-	0.3	0	0.3	0	0	-	-	-
2 & 4	22.0	1.5	4.9	-	1.2	0.4	1.6	0	0.2	-	-	-
3 & 4	6.3	2.6	0	0	0.9	0	0	0.7	0.4	0.4	0.9	0.3
1 & 2 & 3	7.4	0.3	0	-	0.3	0	0	0	0	-	-	-
1 & 2 & 4	21.2	1.3	2.1	-	1.2	0	0	0	0	-	-	-
1 & 3 & 4	5.8	2.1	0	0	0.3	0	0	0	0.2	-	-	-
2 & 3 & 4	4.5	0.3	0	-	0.3	0	0	0	0	-	-	-
1 & 2 & 3 & 4	4.3	0.3	0	-	0.3	0	0	0	0	-	-	-
Completion time (OAT) median	14:06	26:25	38:04	52:39	29:09	26:59	53:33	45:40	17:54	-	-	-

A1 - Panel respondents, paid, 2018
A2 - Panel respondents, paid, 2021
C - Employees, convenience sample
B1 - B6 - Students were rewarded with bonus points
D - EWCS, offline, personal interviews
E1, E2 - WVS, offline, personal interviews

1 – Answering time
2 – Attention check questions
3 – Rating style
4 – Declarative cooperation, logical consistency, open-ended questions

a. analysis based on 223 respondents
b. based only on overall answering time
c. based on 3 attention check questions of instructed response type (“Here choose 3”)
d. based on logical consistency and open-ended questions sub signs
e. based on declarative cooperation and logical consistency sub signs
f. based only on declarative cooperation sub sign
g. based on assessed cooperation – 2 questions
h. based on assessed interest – 1 question

Table 31 *Percent of respondents flagged by the different combinations of the 4 warning signs*

Combination of warning signs	A1	A2
WS1 time	**56.4**	**26.1**
WS2 test	**31.3**	**2.2**
WS3 rating	**11.0**	**6.0**
WS4 logical	**48.1**	**30.3**
WS1(time) + WS2 (test)	57.8	26.4
WS1 & WS2	29.9	1.9
WS1(time) + WS3 (rating)	57.3	28.4
WS1 & WS3	10.2	3.6
WS1(time) + WS4 (logical)	70.2	44.6
WS1 & WS4	34.6	11.7
WS2 + WS3	34.5	7.8
WS2 + WS4	57.8	31.0
WS3 + WS4	52.8	33.7
WS1 + WS2 + WS3	58.3	28.8
WS1 + WS2 + WS4	70.9	44.8
WS1 + WS3 + WS4	70.6	46.5
WS2 + WS3 + WS4	59.2	34.3
WS1 + WS2 + WS3 + WS4 (at least one WS)	**71.0**	**46.6**
OAT median[172]	14:06	26:25
Number of words	3383	3628
Median time without FALSE respondents	27:17	30:17

Table 32 *Comparison of exclusion percentages for different combinations of WS sums and interceptions for datasets A1 and A2*

Let us focus our attention on two paid panel studies. These studies were carried out on users provided by a commercial company that sells its services to researchers, so it is essential to know the quality of the responses provided by their panel members. In those panels, respondents receive a certain amount of points for answering a survey (based on a survey length), and these points can be later exchanged for rewards, chosen by respondents from award options provided by the panel. Panels usually warn their respondents that their answers are subjected to some quality checks, but do not specify the type of these.

[172] for attentive respondents not excluded by WS1

The filter most commonly used by researchers is WS1 (time)[173]. Some studies use the time that the respondent spent on a specific page[174]. However, I have not found any research in which time would be calculated for specific blocks of questions, leading to the local exclusion of respondents instead of global. On the other hand, we have shown that the local exclusion method yields very good results.

The table above also shows that median overall answering time for attentive respondents is longer for the survey that had a bigger total number of words, which is what we would expect to happen – the longer the survey, the more time it takes for respondents to read questions and answer.

Is using only overall answering time to detect FALSE respondents enough? Imagine a respondent who gives random answers to test questions but decides to take a coffee break. In this case the overall answering time will not help to detect such a case despite this respondent having very short answering times in all questions. Using more than one variable considering answering times helps to detect respondents taking breaks, but speeding through the survey anyway.

The second most commonly used filter is WS2 (attention check questions)[175]. Some literature sources found that even a single attention check question can be effective[176], but other sources recommend using more than one attention check question[177]. The data presented in this dissertation supports the conclusion of most literature sources, namely that using more than one attention check question is more efficient. However, we need to remember that the survey should not be longer than it has to be, so it should be carefully considered how many should be used.

I have not found studies using WS3 (response style) in the Polish literature. Parts of WS4 (logical compatibilty and questions about cooperation) were used in Polish General Social Surveys[178], but there were no Polish studies that used all four WS together. However,

[173] i.e. Skarżyńska et al., 2021

[174] Greszki et al., 2015

[175] Kuźmińska & Pazura, 2018; Kuźmińska et al., 2019

[176] Maniaci & Rogge, 2014

[177] Liu & Wronski, 2018; Berinsky et al., 2014

[178] see study documentation for Polish General Social Surveys 1992-2010

some studies that were not conducted in Poland use FALSE responding measures that are the same or close to four WS[179].

Is using only one of the WS as a detection method enough? Although four WS partially detect the same respondents, some respondents are not flagged by, for example, WS2 (attention check questions) but are flagged by WS1 (answering time), which means that each warning sign flags respondents showing different types of inattentive behaviour. Therefore, using only one WS is not enough to detect all possibly FALSE respondents in a data set.

The shocking difference (29 p.p.) in WS2 between A1 and A2 can be explained by the type of attention check questions used. In A1 three instructed response items (i.e., "Please choose <<Rather A>> in this question") were used as attention check, but in A2 five arithmetic questions (i.e., "Choose correct result of this operation 23+5=") were used. In the case of A1, it could be explained by reactance, or negative response to commands, especially if it has not been explained why a respondent should choose that answer and not something else. This interpretation would mean that WS2 should be treated leniently in data set A1.

In the A2 survey, we have shown that the arithmetic attention check questions do not exclude those who answer too quickly. Perhaps they were answering the attention check questions correctly, but it may happen that only the arithmetic ones interested them from the entire survey. We opt to use arithmetic questions because it is easy to explain that those questions serve as breaks from monotony. The low number of errors in arithmetic questions may be caused by the fact that respondents were motivated by a chance of winnig a prize and were informed their answers will be subjected to the procedure for detecting FALSE respondents. It can also be explained by the software change between the two surveys. In A1 respondents could not go back to previous question and change their answer, in A2 respondents were able to change their answer if they noticed that they made a mistake.

The analysis of WS in subsequent studies does not show any general patterns. That means that all WS should be calculated – it is not enough to calculate only one of them.

[179] Ward & Meade, 2018; Brühlmann et al., 2020

2.11.2 Research task #2: Consequences of including FALSE respondents in data analyses

The second research task was to show the consequences of ignoring the problem of FALSE respondents. For this purpose, the reliability of the measurement was compared in groups of excluded respondents (FALSE) and not excluded (attentive) respondents.

I have proposed two procedures to divide survey samples into FALSE and attentive responders groups:

(1) FR procedure based on 4 warning signs (WS)
(2) Flexmix analysis

To show how they work, we need to point out the items of known theoretical correlation. It could be, for example, three items from the METHODICALITY[180] index used in Survey A1.

1. *Person A starts the task only after he has thought out exactly how to perform it. Person B starts the task even when he does not know exactly how to do it and is counting on ideas to come in the process.*
2. *Person A starts writing an essay without having an exact vision of what he will write. Person B first creates a mental vision of what he wants to write, and only then begins to write.*
3. *Person A often starts different tasks thinking that they will do it SOMEHOW. Person B feels bad when s/he does not know HOW to do it.*

The reliability measure of the index called the Cronbach alfa is based on the mean correlation of the items. Therefore, it should be not greater than one, but greater than zero, because it is assumed that the questions included in the indicator should not correlate negatively.
The higher the value of the Cronbach alpha, the higher the reliability of the indicator.

In survey A2 an EXTRAVERSION[181] index consisted of following 4 items was used:

1. *Person A usually directs the course of the conversation with others. Person B is often silent in the company of other people.*

[180] Wieczorkowska, 2022
[181] Wieczorkowska, 2022

2. *Participants in a business dinner or social gathering may believe that A has dominated the conversation. Of person B, they think she said little and that it was the others who had to keep the conversation going.*

3. *Being in a large group of people, person A typically talks to several people, primarily those he knows. Person B talks to many people, including those she did not know before.*

4. *Person A rests best in a place where there is always something going on and among many people. Person B rests best alone or in a small group, in a quiet and peaceful place.*

Analyses were performed according to FR procedures with a lenient criterion used in both survey A1 and A2.

Comparison based on FR procedures (Warning Signs)

Internal reliability operationalised as Cronbach's alfa was tested on 2 subsets of respondents:

(1) ready to be excluded (**flagged "FALSE" respondents**)

(2) **who passed all tests** for all warning signs.

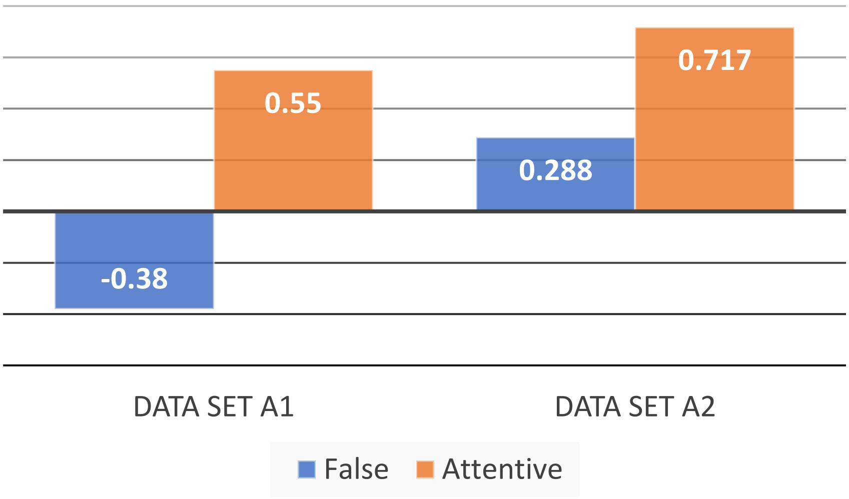


Figure 9 *Cronbach's Alphas' values for the group of FALSE and attentive respondents [based on four WS]. A1: FALSE N=652, attentive N=769; A2: FALSE N=261, attentive N=1233*

As presented in Figure 9 above, the Cronbach Alphas were significantly higher for attentive (**orange bars**) than for FALSE respondents.

In data sets A1, the alpha is much lower in set A2. However, let us look at the **blue bars** representing FALSE respondents who did not pass the Warning Sign tests. We can see that those FALSE respondents did not read the questions because the correlation between some items in the index is negative.

Comparison based on the Flexmix selection

To determine the membership of the FR group, the Flexmix model (a general framework for finite mixtures of regression models) was used. Combining cluster and regression analysis allows us to divide respondents into subgroups based on their fit to different regression lines. Suppose that the theory predicts that the correlation between the answers to the two questions should be positive. In that case, respondents classified by the Flexmix algorithm as the group with a negative correlation are potentially suspected to be not attentive in reading the questions.

The groups of FALSE and attentive respondents were extracted using the Flexmix model, based on two questions taken from personality indices.

In data set A1 (N = 1421), the comparison was made on the first items of the METHODICALITY index. In data set A2 (N = 1497), a comparison was made with respect to the first 2 items of the EXTRAVERSION index.

Figure 10 below shows a comparison of Cronbach Alphas values for the two data sets for attentive and FALSE respondents groups.

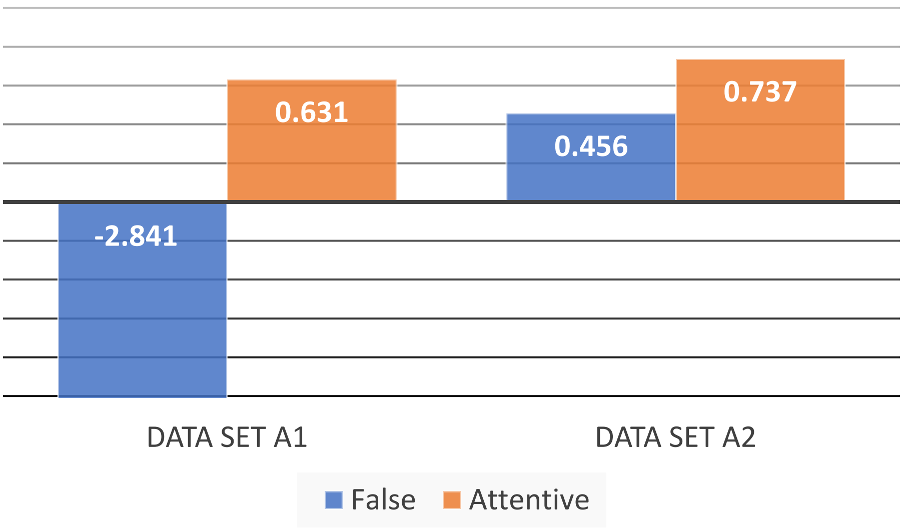


Figure 10 *Cronbach's Alphas' values for two data sets, for FALSE and attentive respondents, based on flexmix. A1: FALSE N=456, attentive N=965; A2 FALSE N=509, attentive N=986*

Similarly to WS, the division of respondents into attentive and FALSE groups based on the Flexmix model showed that the index reliability - Cronbach's Alpha - is much lower (in A1 even below zero) in the FALSE respondents' group than in the group of attentive respondents' group.

2.12 Limitations

The limitations of this research come from the type of research data that were analysed.

Offline data files consisted of publicly available high-budget international surveys (World Values Survey and European Working Conditions Survey) that are carefully designed by international teams of researchers.

Online data files consisted of research conducted by the Managerial Psychology and Sociology Unit in our Faculty, where measurement tools were constructed with great concern about respondent's motivation, encouraged taking breaks, carefully prepared instructions and information about the topic and content of the questions. In our research, the respondent has the freedom not to answer a question or simply say 'Don't know', which means they can choose non-informative answer.

Invitation to take part in the research was sent to selected groups of respondents who were motivated by different means (e.g., paid, getting bonus points for MBA and other students). Invitation links to surveys were anonymous; however, respondents groups to which those links were sent were known to the researchers. Even such efforts were not enough, and not all respondents passed all warning signs tests.

An invitation to take part in a typical Internet survey is posted on the Internet where everyone has access to it, so we can predict that the number of FALSE respondents will be much bigger.

The main limitation of this dissertation is the lack of experimental studies. We have just started the research program in which the dependent variable is the frequency of Warning Signs in experimental groups which differ in the values of independent variables, e.g. type of feedback.

The first experiment has been already conducted and is described in the Attachment 11. Respondents were **randomly** divided into 2 groups that differed in the type of feedback in the test questions (arithmetic questions). In group E1 (N = 191) the respondent chooses the wrong answer, eg, '25' in the question '18 + 4 =' Got the signal 'incorrect' and was forced to choose again, in group E2 (N = 223) the wrong answer was accepted. There were paradoxically and significantly more errors (operationalized as more than two clicks on the arithmetic question) in group group E1 than in E2. Both groups did not differ with respect to other warning signs. Contrary to the hypothesis, forcing respondents to correct wrong answer **did not improve** their attention.

Another limitation of my research is the restricted education level of the respondents - all of them graduated from highschool, which means that the studies on the group of less educated respondents are needed.

The consequences of ignoring the presence of FALSE respondents were shown only in the case of 2 data sets, since they had the highest rates of FALSE respondents, so the comparison of two groups (FALSE vs. attentive) were statistically valid.

2.13 Directions for future research

I can see 5 possible directions for future research.

First of all, automatization of the process - the FR procedure proposed in this dissertation has to be executed mostly manually, with the researcher making decisions about which thresholds are suitable for a particular dataset at hand.

Second, the proposed procedure should be compared with the results of machine learning algorithms[182].

Third, it would be interesting to check whether the **FR procedure** could be used to detect bots[183] (machines that fill questionnaires without human intervention), and if it could, how efficient it is in doing so.

Fourth, it would be interesting to test the impact of immediate feedback and feedback in general, which seems to be a way of motivating respondents to give more thought out responses.

Fifth, it would be interesting to further study the relationship between respondent's age and the number of warning signs they were flagged by. The negative correlation we found in A2 is consistent with previous research[184] indicating, that older respondents are more attentive than younger respondents.

[182] Schroeders, et al., 2022; Gogami et al., 2021
[183] Dennis et al., 2018; Buchanan & Scofield, 2018
[184] Maniaci & Rogge, 2014

2.14 Impact of the dissertation

The doctoral dissertation has a cognitive, methodological, and application contribution. It tries to estimate the scale of the occurrence of FALSE respondents in a well-prepared survey – it was shown that the presence of FALSE respondents drastically reduced reliability of the measurement. Unreliable data coming from FALSE respondents may change correlations[185], make the analysis and evaluation of the results of research difficult[186], decrease statistical power[187] and effect size[188], and lower internal consistency[189]. HRM theories confirmed by biased (not reliable) data are not valid so FALSE respondents detection is an important pre-analysis task to do.

The application contribution consists of developing a procedure for detecting FALSE respondents in HRM studies that could be used by other researchers.

The original methodological contribution is FR detection procedure and the empirically tested proposal of using the FLEXMIX procedure (finite mixtures of generalized regression models) for detecting FALSE respondents.

[185] Huang et al., 2015a; McGrath et al., 2010
[186] Maniaci & Rogge, 2014
[187] Maniaci & Rogge, 2014
[188] Brühlmann et al., 2020
[189] Huang et al., 2012

2.15 Procedure for detecting FALSE respondents

The general form of the FR detection procedure is described in the following. It does not include standard procedures for preparing data for analysis (checking if the data are complete, duplicate answers, program errors, description of variables) or analyses not directly related to the topic of FALSE respondents.

Computation of WS1: Too short answering TIME.

Step 1. Find the maximum reading speed (enabling comprehension) for your sample (e.g., students read faster, etc.). For our analysis of the well-educated samples, a maximum reading speed of **300 words per minute** was assumed for all datasets.

Step 2. Calculate OAT - Overall Answering Time of the study and the partial AT for blocks of your survey (if applicable).

Step 3. Count the total number of words in the survey, considering that the repeated rating scale should be included in this number only once.

Step 4. Divide the number of words by the maximum speed – this is a general estimate of the minimum time needed to read the questions.

Step 5. Flag those respondents whose time was below the minimum time threshold. Extremely long times are not a problem because we do not ask people to hurry in surveys.

Step 6. If the survey had optional elements that could be omitted, but the omission does not affect the main objectives, repeat steps 2-5 for the version that does not include these elements.

Step 7. If PAT is available for individual question blocks / single pages, repeat steps 1-5 for each question/question block.

Step 8. Check the response time globally and locally - respondents may be 'FALSE' only in some parts of the survey. Set the percentage threshold for the number of blocks/questions answered too fast, above which a respondent should be excluded from the survey globally.

Computation of WS2. Incorrect answers to Attention Check Questions.

We have discussed earlier what conditions the attention check questions should meet.

Step 1. Count the number of incorrect responses to attention check questions for each respondent.

Step 2. Decide whether a strict (no errors) or lenient (1 error allowed) criterion will be used.

Computation of WS3. Low Differentiation Rating Style and Non-Informative (DK) Answers.

Find out which items can be used to analyse DK answers and variance - e.g., with the same rating scale, on the same topic, matrix questions (multiple statements on the same page).

Step 1. Count the number of DK answers for each respondent locally and globally.

Step 2. Calculate rating style indicators, e.g. the standard deviation or variance, for each series.

Step 3. Decide on local or global thresholds for the number of DK answers and rating style indicator.

Step 4. Compare the number of DK and rating style indicators with the set thresholds.

Step 5. Flag respondents above previously set thresholds.

Computation of WS4. Low declarative cooperation level, logical inconsistency, odd answers to open-ended questions.

Applicable only if the following types of items are included in your survey:
(1) respondents' self-evaluation of their engagement
(2) open-ended questions

(3) pairs of items that allow answers to them to be tested for consistency (e.g., answer NONE to the question about the number of children & answer LOW, HIGH, or anything meaningful to the question about satisfaction with the relationship with children)**.**

Step 1. Decide on the threshold on the **respondents' engagement self-evaluation scale** (e.g., on a 6-point scale, from 1 indicating a complete lack of commitment to 6 indicating a very high commitment).

Step 2. Check the logical consistency in the pairs of items. Flag inconsistent respondents.

Step 3. Code answers to open-ended questions into 5 categories: (1) no answer, (2) answer not connected with the topic of the question, (3) too short answer, (4) informative answer, (5) refusal. For obligatory questions, count the answers from the first three categories. For the facultative questions, count the answers from categories (2) and (3). Flag respondents that have been counted.

Filtering out FALSE respondents.

- Decide on the number of WS flags threshold – none allowed (strict) or 1 flag allowed (lenient).
- Sum up the number of WS for each respondent
- If the threshold is equal to ZERO and SUM(WS) > 0, flag the respondents as FALSE. If threshold is equal to ONE and SUM(WS) > 1, flag the respondents as FALSE.

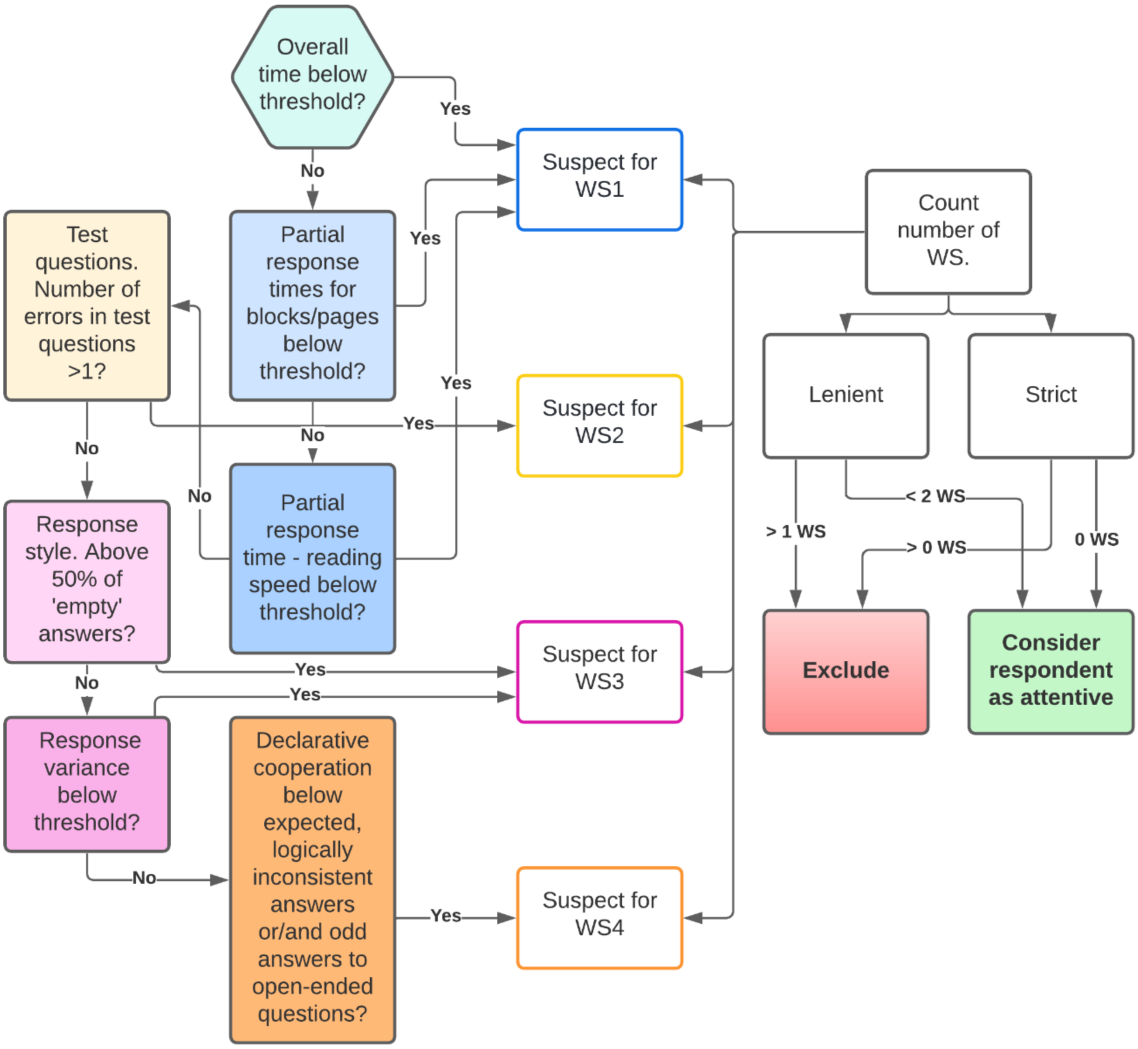

Overall time below threshold?
Yes
No
Suspect for WS1
Partial response times for blocks/pages below threshold?
Yes
Test questions. Number of errors in test questions >1?
Yes
Yes
Suspect for WS2
No
No
No
Partial response time - reading speed below threshold?
Response style. Above 50% of 'empty' answers?
No
Yes
Yes
Suspect for WS3
Response variance below threshold?
Declarative cooperation below expected, logically inconsistent answers or/and odd answers to open-ended questions?
No
Yes
Suspect for WS4
Count number of WS.
Lenient
Strict
< 2 WS
> 1 WS
> 0 WS
0 WS
Exclude
Consider respondent as attentive

Figure 11 *Visual scheme of the prodcedure for detecting FALSE respondents*

4 Attachments

Attachment 1. Summary of descriptive statistics of all data sets

Data set	N	Year of the study	% of female participants	Mean Age	SD Age
A1	1421	2018	54	34.6	3.55
A2	1497	2021	55	42.6	10.69
C	287	2020	56	36	12.04
B1	740	2018	67	24	5.75
B2	341	2020	73.2	22.6	3.02
B3	414	2021	65.7	22	2.61
B4	308	2021	70	22	3.08
B5	140	2021	37.9	21.7	1.90
B6	497	2021	69.6	23.4	3.98

Table 33 *Summary of descriptive statistics of online data sets*

Data set	N	Year of the study	% of female participants	Mean Age	SD Age
D	43850	2015	49.6	43.4	12.7
E1	83975	2005-2007	50.9	41.3	16.5
E2	89565	2010-2014	51.1	41.7	16.5

Table 34 *Summary of descriptive statistics of offline data sets*

Attachment 2. How changing an answer to just one question changes variance

Let us assume that there are three respondents answering a five-question, a ten-question, and a fifteen-question survey. Respondent 1 (R1) is genuinely answering questions, Respondent 2 (R2) is just speeding through the survey, choosing one answer option, and Respondent 3 (R3) doing the same as the second, but they randomly decide to click any other option just once.

Below, a setup and an outcome of this simple simulation are presented.

'pa' is a shortcut for 'points away' – how far the one differing answer is located from the original string of values.

Respondent	**R1**			**R2**	**R3**					
					1pa	**2pa**	**3pa**	**4pa**	**5pa**	**6pa**
Scale size	*3*	*5*	*7*	*3/5/7*	*3/5/7*	*3/5/7*	*5/7*	*5/7*	*7*	*7*
5 questions	1	1	1	1	1	1	1	1	1	1
	3	4	4	1	1	1	1	1	1	1
	3	3	4	1	1	1	1	1	1	1
	2	3	6	1	1	1	1	1	1	1
	3	5	5	1	2	3	4	5	6	7
Variance	**0.8**	**2.7**	**3.5**	**0**	**0.2**	**0.8**	**1.8**	**3.2**	**5**	**7.2**
10 questions	1	1	1	1	1	1	1	1	1	1
	3	2	4	1	1	1	1	1	1	1
	3	5	4	1	1	1	1	1	1	1
	2	2	6	1	1	1	1	1	1	1
	3	4	1	1	1	1	1	1	1	1
	1	3	3	1	1	1	1	1	1	1
	2	3	2	1	1	1	1	1	1	1
	1	2	3	1	1	1	1	1	1	1
	1	5	4	1	1	1	1	1	1	1
	3	4	7	1	2	3	4	5	6	7
Variance	**0.89**	**1.88**	**3.83**	**0.00**	**0.10**	**0.40**	**0.90**	**1.60**	**2.50**	**3.50**
15 questions	1	1	1	1	1	1	1	1	1	1
	3	2	4	1	1	1	1	1	1	1
	3	5	4	1	1	1	1	1	1	1
	2	2	6	1	1	1	1	1	1	1
	3	4	5	1	1	1	1	1	1	1
	1	3	3	1	1	1	1	1	1	1
	2	3	2	1	1	1	1	1	1	1
	1	2	3	1	1	1	1	1	1	1
	1	5	4	1	1	1	1	1	1	1
	3	4	7	1	1	1	1	1	1	1
	1	5	7	1	1	1	1	1	1	1
	3	5	1	1	1	1	1	1	1	1
	1	2	6	1	1	1	1	1	1	1
	1	1	3	1	1	1	1	1	1	1
	3	4	2	1	2	3	4	5	6	7
Variance	**0.92**	**2.17**	**3.98**	**0**	**0.07**	**0.27**	**0.60**	**1.07**	**1.67**	**2.4**

Table 35 *Simulation of variance for different question numbers and rating scale sizes*

The above table shows that choosing just one different option in an extreme case may make the variance of FALSE respondents similar to the variance of genuine respondents. The low variance sign flags only the most common occurrence – usually the first case for the three respondents, as higher values could be considered acceptable.

Attachment 3. The original content of the questions about the involvement

Below contents refer to data sets A1 to C.

Jak oceniają Państwo stopień swojego zaangażowania w to zadanie?
1 – bardzo niski, 2, 3, 4, 5, 6 – bardzo wysoki

Na ile to zadanie było dla Państwa męczące?
1 – bardzo męczące, 2, 3, 4, 5, 6 – w ogóle nie męczące

Gdyby mieli Państwo powtórnie (np. jutro) uczestniczyć w ankiecie, to czy Państwa odpowiedzi byłyby:

1. identyczne
2. mogłyby się różnić nieznacznie
3. mogłyby się różnić diamteralnie

Attachment 4. Distributions of overall answering times for data sets C and B2

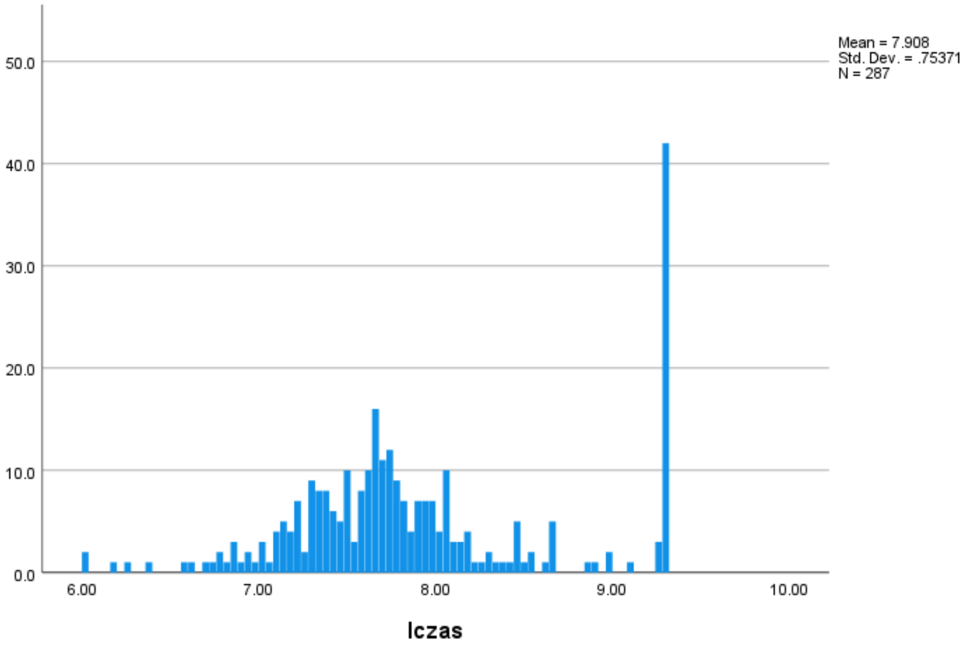


Figure 12 *Distribution of natural logarithm of overall answering time, data set C*

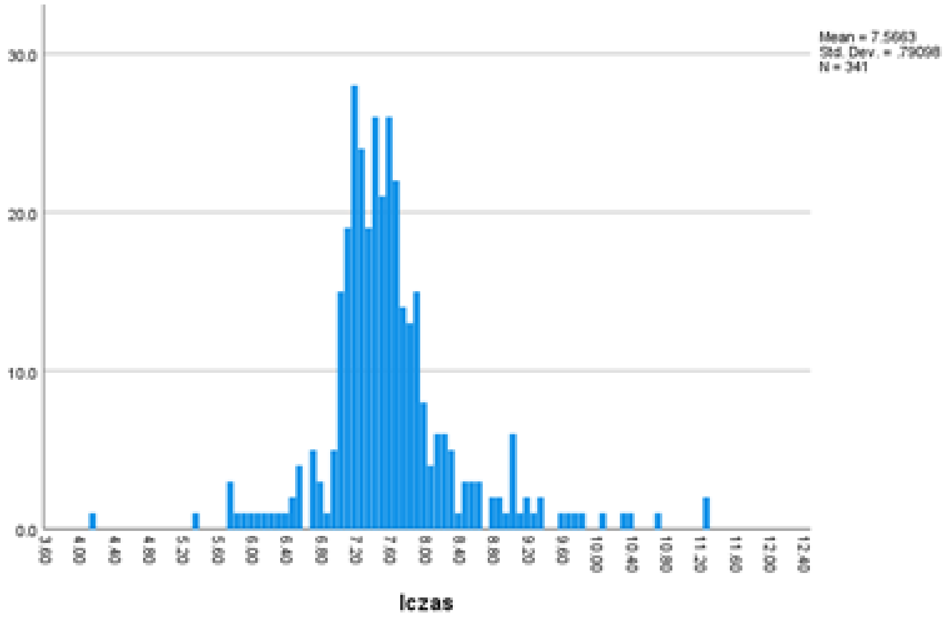


Figure 13 *Distribution of natural logarithm of overall answering time, data set B2*

Attachment 5. Distributions of partial answering times for data set B3

Distribution has been truncated to values at the ends of the main part of the tail.

Blocks presented to group E1 (N=191)

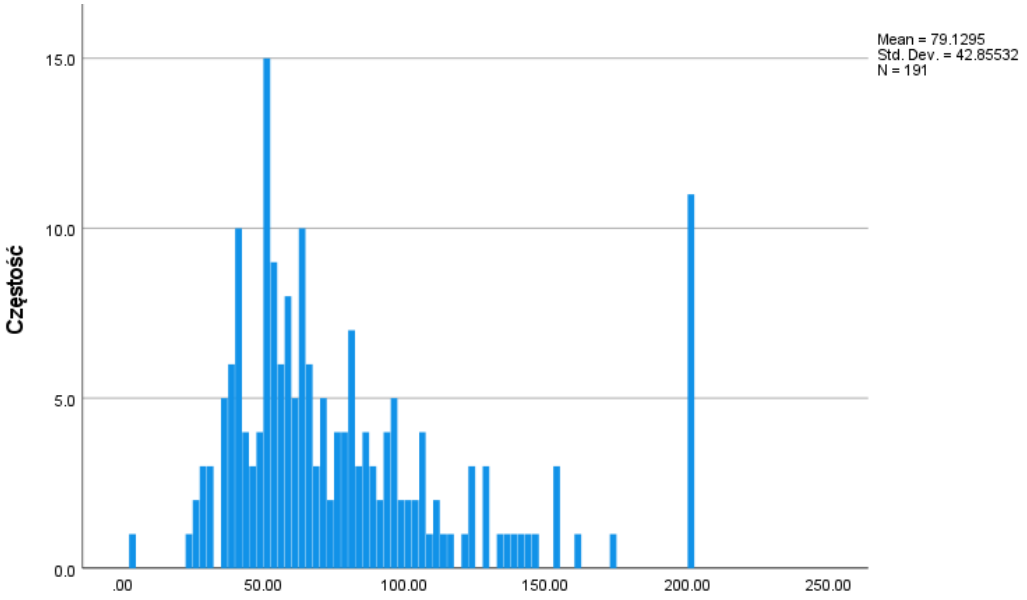


Figure 14 *Distribution of partial answering time for Block 1, group E1*

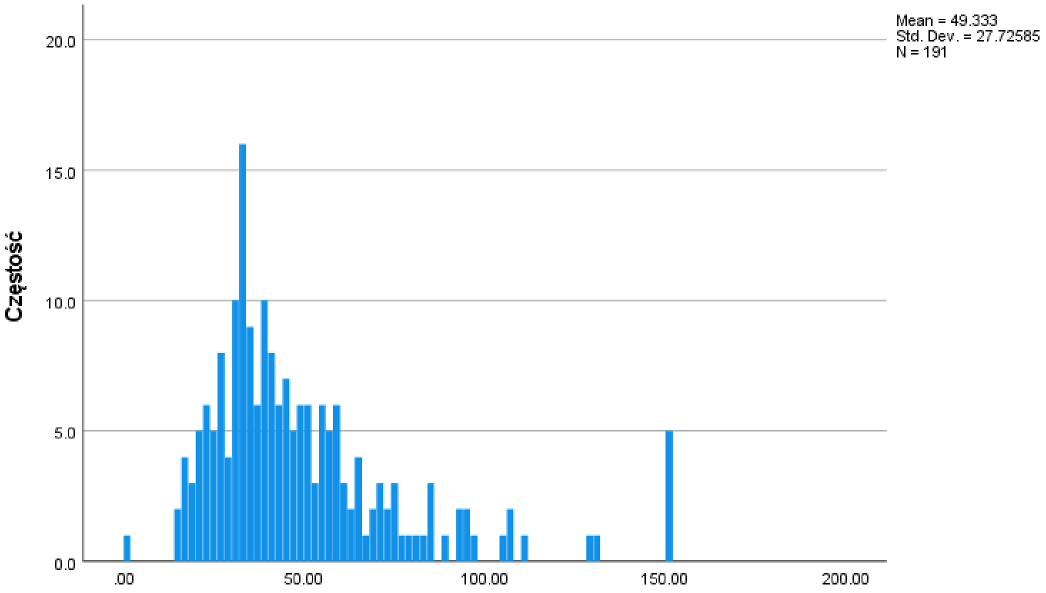


***Figure 15** Distribution of partial answering time for Block 2, group E1*

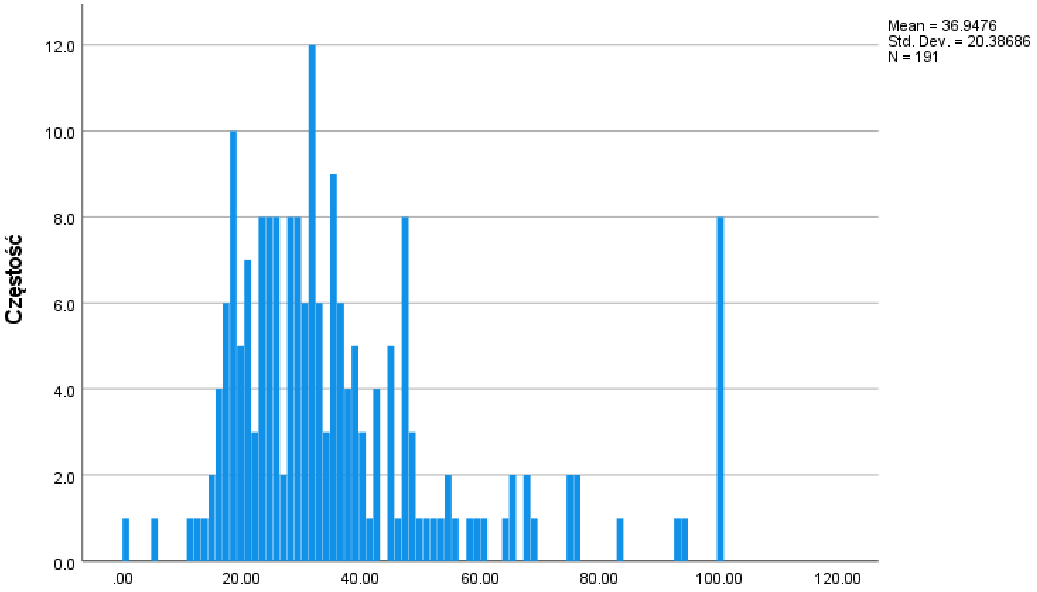


***Figure 16** Distribution of partial answering time for Block 3, group E1*

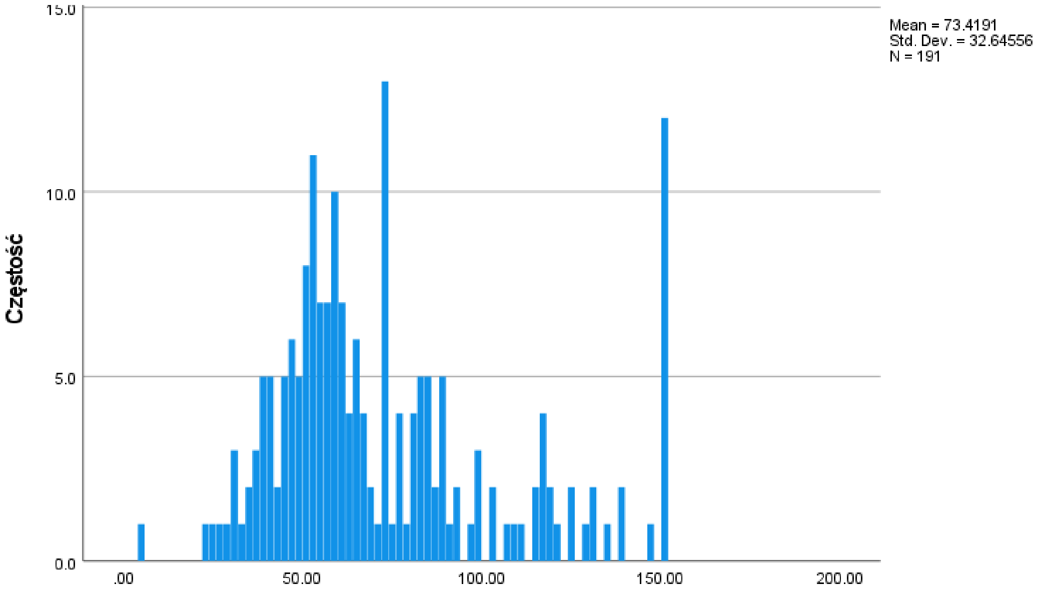


__Figure 17__ Distribution of partial answering time for Block 4, group E1

Blocks presented to group E2 (N=223)

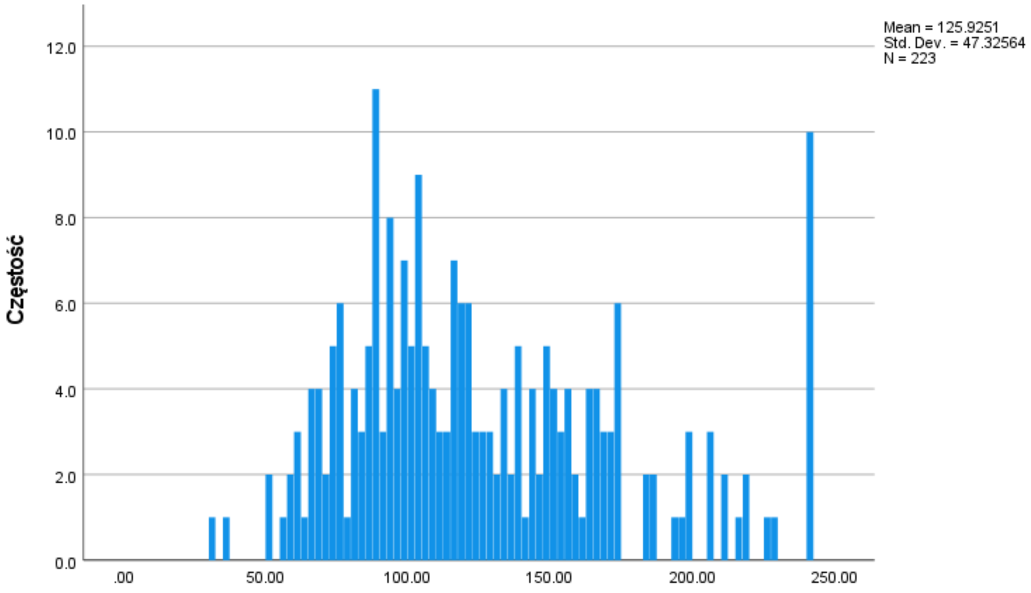


__Figure 18__ Distribution of partial answering time for Block 5, group E2

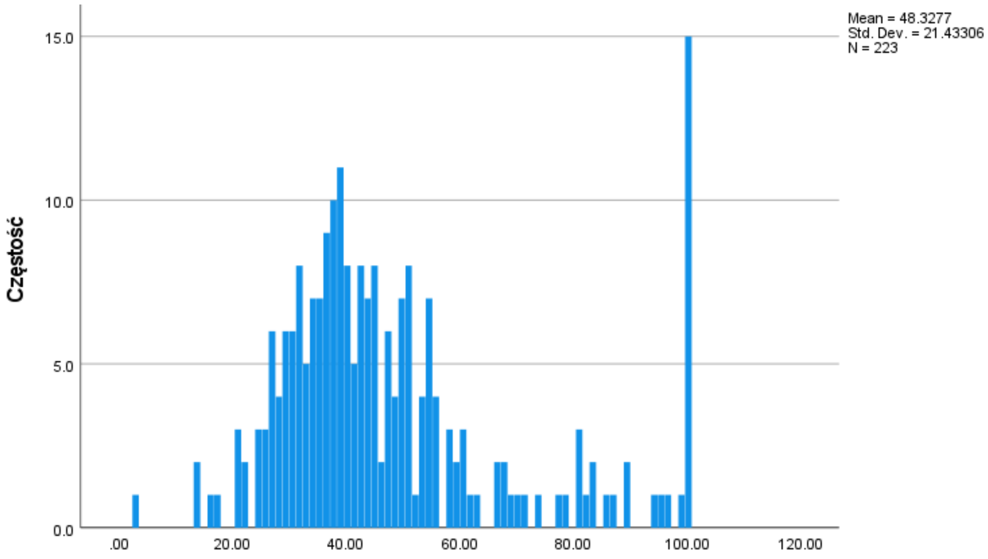


Figure 19 *Distribution of partial answering time for Block 6, group E2*

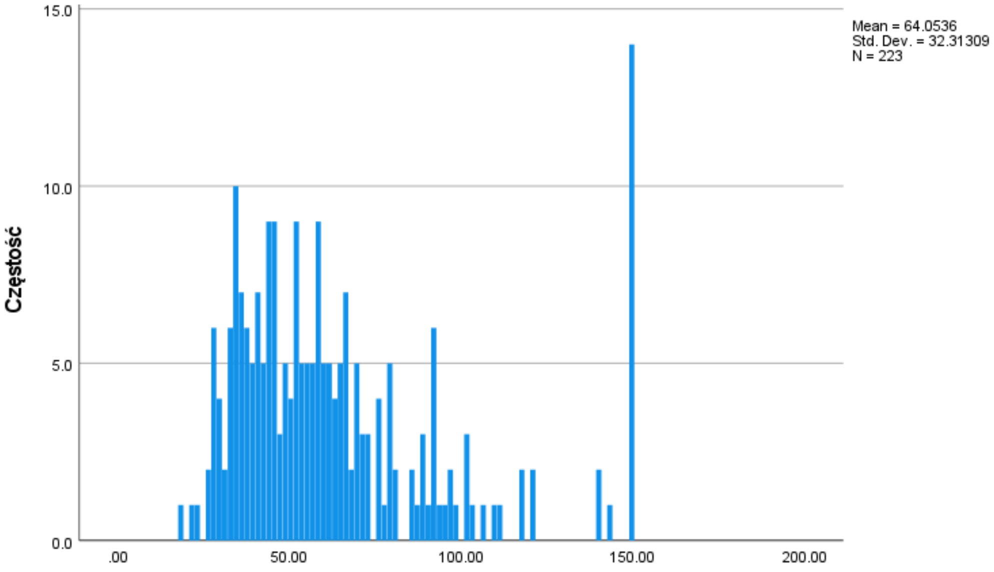


Figure 20 *Distribution of partial answering time for Block 7, group E2*

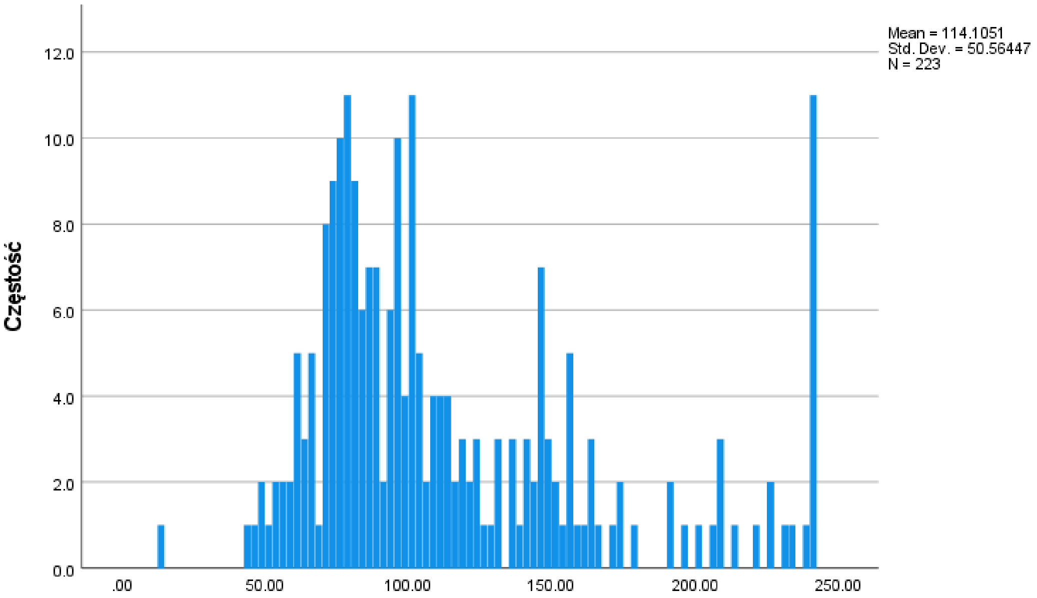


Figure 21 *Distribution of partial answering time for Block 8, group E2*

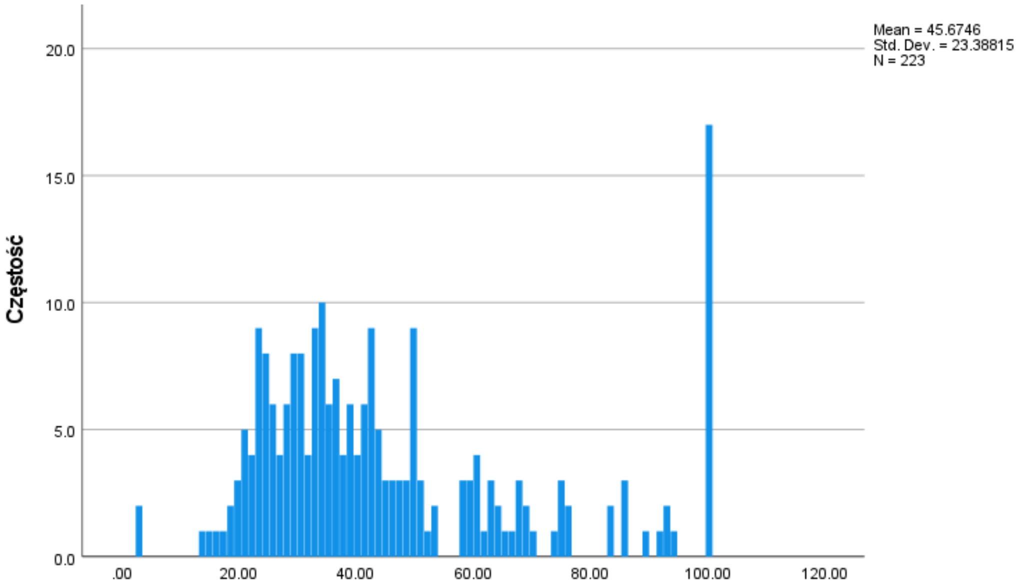


Figure 22 *Distribution of partial answering time for Block 9, group E2*

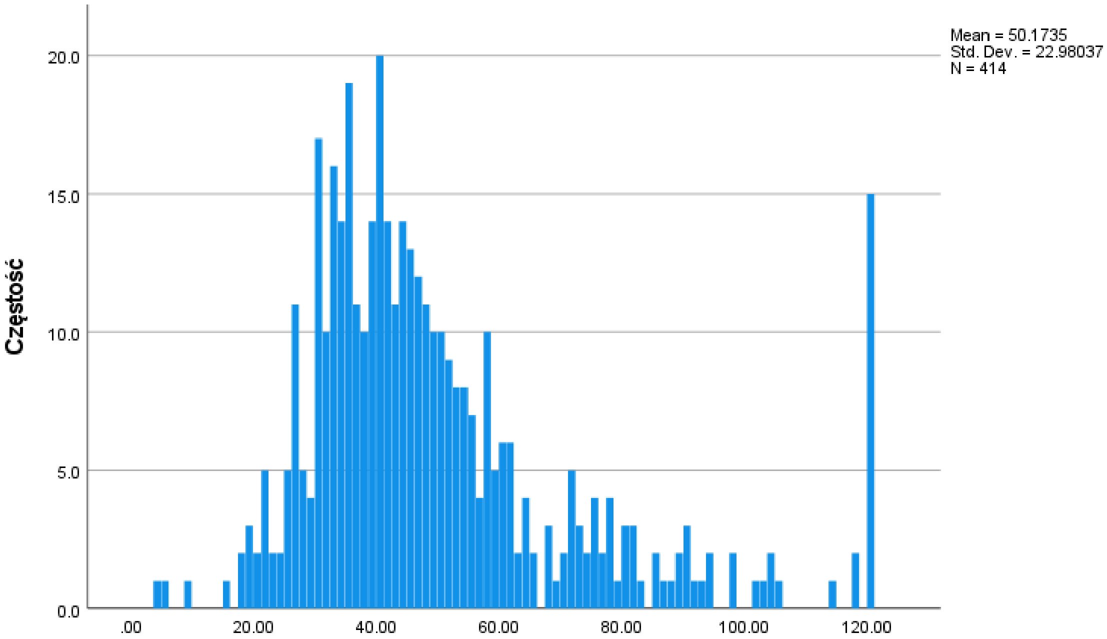


Figure 23 *Distribution of partial answering time for Block 10, group E2*

Attachment 6. Arithmetic questions examples [data set B3]

Arithmetic attention check question used (original Polish version).

1. Wybierz właściwy wynik działania 50 – 15 = [answers:] 30, 35, 40, 45, 50
2. Wybierz właściwy wynik działania 15 + 8 = [answers:] 23, 24, 25, 26 , 27
3. Wybierz właściwy wynik działania 23 – 4 = [answers:] 15, 16, 17, 18, 19
4. Wybierz właściwy wynik działania 30 – 5 = [answers:] 10, 15, 20, 25, 30
5. Wybierz właściwy wynik działania 10 – 5 = [answers:] 3, 4, 5, 6, 7

Attachment 7. Additional information about data set D

Data set D.

Series of questions analysed in Warning Sign #3.

Q29a - Vibrations from hand tools, machinery etc. [Are you exposed at work to…?]

Q29b - Noise so loud that you would have to raise your voice to talk to people [Are you exposed at work to…?]

Q29c - High temperatures which make you perspire even when not working [Are you exposed at work to…?]

Q29d - Low temperatures whether indoors or outdoors [Are you exposed at work to…?]

Q29e - Breathing in smoke, fumes (such as welding or exhaust fumes), powder or dust etc. [Are you exposed at work to…?]

Q29f - Breathing in vapours such as solvents and thinners [Are you exposed at work to…?]

Q29g - Handling or being in skin contact with chemical products or substances [Are you exposed at work to…?]

Q29h - Tobacco smoke from other people [Are you exposed at work to…?]

Q29i - Handling or being in direct contact with materials which can be infectious [Are you exposed at work to…?]

Q30a - Tiring or painful positions [Does your main paid job involve…?]

Q30b - Lifting or moving people [Does your main paid job involve…?]

Q30c - Carrying or moving heavy loads [Does your main paid job involve…?]

Q30d - Sitting [Does your main paid job involve…?]

Q30e - Repetitive hand or arm movements [Does your main paid job involve…?]

Q30f - Dealing directly with people who are not employees at your workplace [Does your main paid job involve…?]

Q30g - Handling angry clients, customers, patients, pupils etc. [Does your main paid job involve…?]

Q30h - Being in situations that are emotionally disturbing for you [Does your main paid job involve…?]

Q30i - Working with computers, laptops, smartphones etc [Does your main paid job involve…?]

Rating scale to Block #1.

1 ‘All of the time’, 2 ‘Almost all of the time’, 3 ‘Around ¾ of the time’, 4 ‘Around half of the time’, 5 ‘Around ¼ of the time’, 6 ‘Almost never’, 7 ‘Never’, 8 ‘DK’, 9 ‘Refusal’

An explanation for excluding ‘Never’ from the analysis

It is possible, although highly unlikely that the respondent was indeed never subjected to any of the conditions described by the questions of this series. Despite the fact of this being unlikely, it cannot be determined for certain which for which respondent this is actually true, so the author of this dissertation decided to exclude this value from the analysis altogether as a ‘default’ answer. To put that into a slightly different perspective – respondents had an option to choose ‘Don’t know’ as their answer, and for some reason, many of them decided to choose ‘Never’. That decision may come from the answers actually reflecting the true state of things, or it may just be a way of avoiding answering ‘Don’t know’ (seen as an uncooperative answer) and choosing ‘second best’ option that does not require a lot of cognitive effort (recalling from the memory) but can be considered as valid – which is actually ‘Never’ in this case. To stay on the safe side, the decision of excluding the answer seems to enable getting around the problem of interpretation of this behaviour.

Attachment 8. Additional information on Data sets E1 and E2

Questions used in the analysis of warning sign #3.

Schwartz: Important to this person to think up new ideas
Schwartz: Important to this person to be rich
Schwartz: Important to this person living in secure surroundings
Schwartz: Important to this person to have a good time
Schwartz: Important to this person to help the people
Schwartz: Important to this person being very successful
Schwartz: Important to this person adventure and taking risks
Schwartz: Important to this person to always behave properly
Schwartz: Important to this person looking after environment
Schwartz: Important to this person tradition
Rating scale: Very much like me, Like me, Somewhat like me, A little like me, Not like me, Not at all like me

Attachment 9. Analysis of warning signs #3 and #4 for data sets D. E1, E2

Data set D, warning sign #3

Block 1. Exposure to stressing/harmful conditions at work

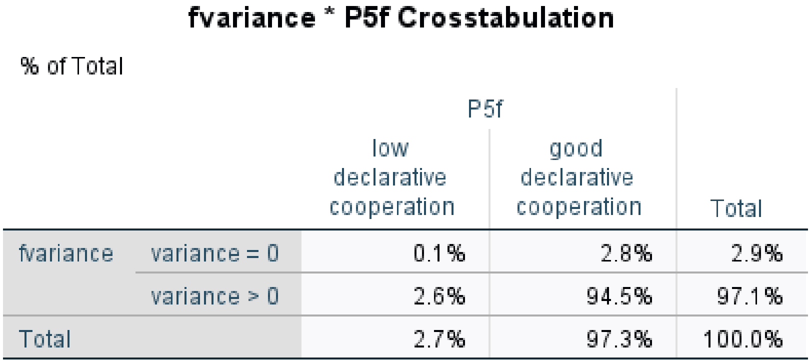
For the whole EWCS 2015 sample, 1264 respondents had variance equal to 0, which amounts to 2.9%.

Data set D, warning sign #4

In **Table 36** below, two shaded rows amount to a total of 436 respondents who had poor or very poor cooperation levels. Those respondents were about 3.5 years older and had about two years of education less than respondents having fair, good or very good cooperation level. 63.3% of those respondents were male (difference statistically significant).

For the whole sample, there were a total of 1190 people who had very poor or poor cooperation, Always or most of the time asked for clarification or had difficulty answering questions. 56% of those were male (difference statistically significant), had about two years of education less, and were about 2.5 years older.

As is shown in **Table 36** below, only 0.1% of respondents were flagged by both signs.



fvariance * P5f Crosstabulation

% of Total

		P5f		
		low declarative cooperation	good declarative cooperation	Total
fvariance	variance = 0	0.1%	2.8%	2.9%
	variance > 0	2.6%	94.5%	97.1%
Total		2.7%	97.3%	100.0%

***Table 36** Comparison of warning signs #3 and #4 results for the whole EWCS sample*

Overall, the analysis of the whole sample excludes 5.5% - which amounts to 2382 respondents in total.

Data set E1, Warning sign #3

The outcome for all countries of WVS 2005 in terms of variance for ten questions about Schwartz values is presented in **Table 37**.

Variance	Percent	Valid percent
Equal to 0	1.6	1.8
Any other value	85.6	98.2
Missing values	12.9	-

Table 37 *Percent and valid (excluding missing values) percent of variance equal to 0 in data set E1*

In terms of these 10 questions, 1326 (1.8%) of valid answers (empty, refusals treated as missing values) had a standard deviation equal to 0, and should be excluded.

Data set E1, Warning sign #4

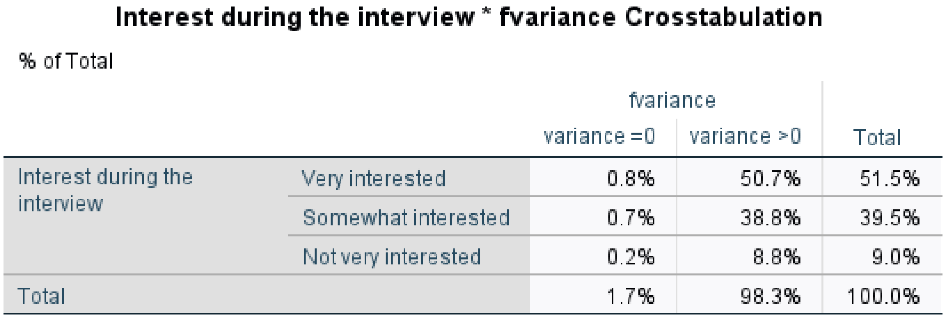
The distribution of all answers is shown below, in **Table 38** below.

Answer	Percent of all answers	Percent of valid answers
Respondent was very interested	46.3	51.5
Respondent was somewhat interested	35.5	39.5
Respondent was not interested	8.1	9.0
Sub-total	**89.9**	**100.0**
Not asked	9.2	-
No answer	0.8	-
Don’t know	0.1	-
Missing: Not asked by the interviewer	0.0	-

Table 38 *Distributions of answers to respondent's interest assessment – all answers and valid answers, all countries, the E1 data set*

Although the respondent’s interest during the interview was assessed by the interviewer (and introduced interviewer’s bias into the picture), there is still missing data to this question, but excluding these cases leads to the final number of 9% of the sample that was not interested in the interview.

Summary of analysis for data set E1



Interest during the interview * fvariance Crosstabulation

% of Total

		fvariance		
		variance =0	variance >0	Total
Interest during the interview	Very interested	0.8%	50.7%	51.5%
	Somewhat interested	0.7%	38.8%	39.5%
	Not very interested	0.2%	8.8%	9.0%
Total		1.7%	98.3%	100.0%

Table 39 *Crosstabulation of variance groups and respondent interest assessment*

As is shown in **Table 39** above, 0.2% (172 respondents) of the whole sample was flagged by both warning signs.

Overall, a total of 7896 respondents have been excluded, which amounts to about 10.5%[190] of the sample.

Data set E2, Warning sign #3

In the E2 data set, also ten questions about Schwartz values were used.

Results of the analysis are presented in Table 40 below.

Variance	Percent	Valid percent
Equal to 0	2.4	2.5
Any other value	96.1	97.5
Missing values	1.4	-

Table 40 *Percent and valid (excluding missing values) percent of variance equal to 0 in data set E2*

In terms of these 10 questions, 2183 (2.5%) of valid answers (empty, refusals treated as missing values) had a standard deviation equal to 0, and should be excluded.

Data set E2, Warning sign #4

Distributions of percentages for all answers and valid answers are shown below in Table 41.

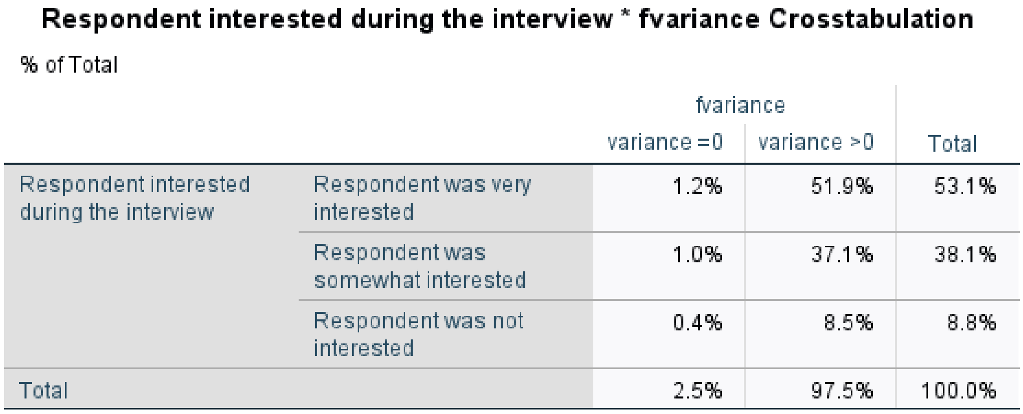
[190] Due to missing data, sample size was reduced to 75488 respondents – the percent of excluded respondents was calculated for reduced sample size.

Answer	Percent of all answers	Percent of valid answers
Respondent was very interested	49.4	53.1
Respondent was somewhat interested	35.4	38.1
Respondent was not interested	8.2	8.8
Sub-total	93.0	100.0
Not asked	6.7	-
No answer	0.1	-
Don't know	0.0	-
Missing: Unknown	0.2	-

***Table 41** Distributions of answers to respondent's interest assessment – all answers and valid answers, all countries, the E2 data set*

There was less missing data in the E2 data set, although still a considerable amount. In 8.8% of valid cases, interviewers assessed that respondent was not interested.

Summary of analysis for data set E2



Respondent interested during the interview * fvariance Crosstabulation

% of Total

		fvariance		
		variance =0	variance >0	Total
Respondent interested during the interview	Respondent was very interested	1.2%	51.9%	53.1%
	Respondent was somewhat interested	1.0%	37.1%	38.1%
	Respondent was not interested	0.4%	8.5%	8.8%
Total		2.5%	97.5%	100.0%

***Table 42** Crosstabulation of variance groups and respondent interest assessment*

As is shown in **Table 42** above, 0.4% of the whole sample was flagged by both warning signs.

Overall, a total of 9127 respondents have been excluded, which amounts to about 11%[191] of the sample.

It is worth noting that for data set E2 actual description of the method of data collection is missing for most countries. If the method was known, it was still face-to-face interviews in most known cases (15 countries), so the author of this dissertation assumes that actual

[191] In some countries one of the questions asked was different, and therefore it created missing data, leaving sample size of 83302 – the percent of excluded respondents was calculated for reduced sample size.

methods were not changed, and they were still face-to-face interviews in most countries participating.

Attachment 10. Detailed statistics for open-ended questions coding

Data set	Question number	No answer		Non-informative answer		Too short answer		Informative answer		Refusal/ No opinion/Not applicable	
		c.	[%]	c.	[%]	c.	[%]	c.	[%]	c.	[%]
A1	Q1	310	21.8	79	5.6	132	9.3	890	62.6	10	0.7
	Q2	8	0.6	38	2.7	5	0.4	1370	96.4	0	0.0
	Q3	289	20.3	56	2.9	0	0.0	1066	75.0	10	0.7
A2	Q1	899	60.1	36	2.4	2	0.1	537	35.9	23	1.5
	Q2	1058	70.7	31	2.1	0	0.0	379	25.3	29	1.9
	Q3	829	55.4	30	2.0	0	0.0	635	42.4	3	0.2
B1	Q1	185	38.5	4	0.5	0	0.0	451	60.9	0	0.0
	Q2	305	41.2	4	0.5	0	0.0	431	58.2	0	0.0
B4	Q1	32	10.4	0	0.0	0	0.0	272	88.3	4	1.3
	Q2	54	17.5	0	0.0	0	0.0	254	82.5	0	0.0
B6	Q1	45	9.1	8	1.6	0	0.0	438	88.1	6	1.2
	Q2	98	19.7	7	1.4	0	0.0	392	78.9	0	0.0

Table 43 *Results of open-ended questions analysis across five data sets [A1, A2, B1, B4, B6]*

Attachment 11. Influence of preventing errors in attention check questions on FALSE responding rate

The hypothesis was tested on data set B3, N=414.

H1: Feedback in case of incorrect answer will increase respondent's attention.

There were two experimental groups, conditions in each groups are presented in **Table 44** below.

<table>
<tr><th>Group E1 (prevented error)</th><th>Group E2 (no error prevention)</th></tr>
<tr><td colspan="2">Respondents were presented five attention check questions (arithmetic type)[192] with five answer options scattered across the survey – experimental groups had different major questions sets, but were made to be approximately similar in terms of the length and time needed to complete the survey.</td></tr>
<tr><td>When question was answered correctly, nothing happened – for respondents, who did not make any errors at first attempt of answering, this survey behaved exactly like survey in group E2.
When question was answered incorrectly, a warning was presented to respondent, and they were not allowed to proceed to next question until they gave a correct answer (which means that all respondends were eventually forced to give a correct answers). Number of attempts was not restricted.</td><td>Regardless of how the question was answered, survey allowed the respondent to proceed to next question. Wheter they wanted to return and correct their mistake (assuming they noticed that they made an error at all), was the respondent's decision. As the survey software did not allow for recording both attempts, only final answer was recorded.</td></tr>
</table>

***Table 44** Description of experimental conditions for testing the influence of disciplining reminder*

Independent variable in this study was the presence of disciplining reminder – one groups was forced to answer correctly, the other was not.

Dependent variable was respondent's attention. Respondent's attentions' operationalization is usually a number of errors in attention check questions (if only attention check questions were used), but in this case group E1 was forced to eventually answer correctly, so there were no errors. For this reason, the operationalization was changed to number of clicks on the page with questions – value greater than 2 clicks[193] meant that the respondent likely made an error.

[192] Contents of questions is presented in Attachment 6.

[193] Survey had auto-advance mode turned on, which means that only 1 click was needed to answer the question, no confirmation required.

Percent of respondents flagged by strict (1 click allowed) and lenient (2 clicks allowed) criterion based on number of clicks is presented in **Table 45** below.

Group	N	Number of attention check questions in data set	Respondents flagged by strict criterion		Respondents flagged by lenient criterion	
			[count]	[%]	[count]	[%]
E1	191	5	65	34.0	20	10.5
E2	223	5	61	27.4	9	4.0
Total	414	-	126	30.4	29	7.0

Table 45 *Percent of respondents flagged by number of clicks in division by experimental group, data set B3*

Hypothesis was tested using Chi-squared test (for lenient criterion numbers), results are presented in **Table 46**.

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.541[a]	1	.011		
Continuity Correction[b]	5.590	1	.018		
Likelihood Ratio	6.615	1	.010		
Fisher's Exact Test				.012	.009
Linear-by-Linear Association	6.525	1	.011		
N of Valid Cases	414				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13.38.

b. Computed only for 2x2 table

Table 46 *Results of Chi-square test of H1*

The difference between E1 and E2 groups was statistically significant. As can be seen in **Figure 24** below, the biggest difference occurred for Q2 – detailed analyses of differences for each question separately showed a significant difference[194] for Q2 and Q3, in which case more respondents answered without an error in group E2 (without preventing error). Differences in those two questions are also possibly responsible for difference on all 5 questions being significant.

Overall, **H1 can be considered not confirmed.** Respondents answer more attentively when making error in attention check question was not prevented, but this difference is

[194] χ^2=4.12, p=0.047

only true for the attention check questions. For other warning signs testing differences was not possible due to low counts of respondents in subgroups.

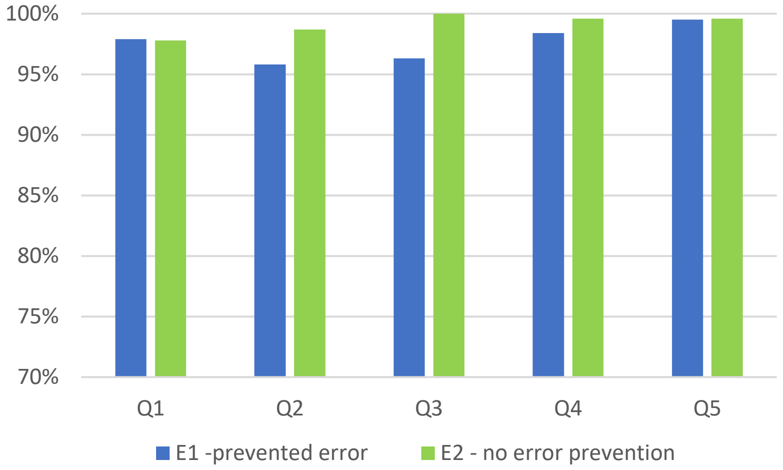
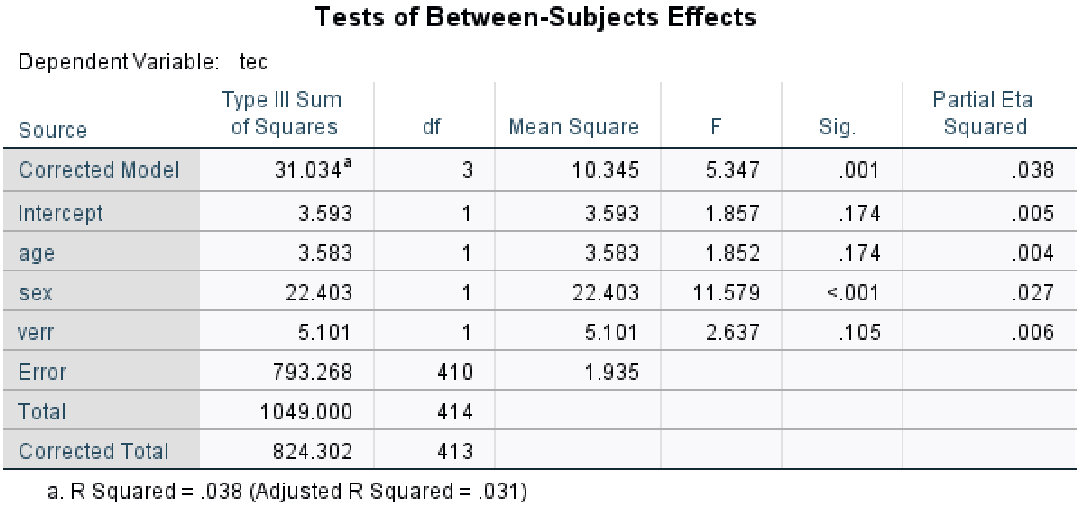


***Figure 24** Comparison of percent of correct answers to five arithmetic questions for two experimental groups*

Additional analysis of variance (to control for gender and age) showed that gender was an unexpected significant covariate, so gender differences were further investigated. Results of these analyses are presented in Attachment 11 – in short, women made significantly more errors than men in E2 group, the difference was not significant in E1 group.

Analysis of gender difference for disciplining reminder.

Analysis of covariance, presented in **Table 47** and illustrated in **Figure 25** (below) show, that there were gender differences between groups in the number of errors.



Tests of Between-Subjects Effects

Dependent Variable: tec

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	31.034[a]	3	10.345	5.347	.001	.038
Intercept	3.593	1	3.593	1.857	.174	.005
age	3.583	1	3.583	1.852	.174	.004
sex	22.403	1	22.403	11.579	<.001	.027
verr	5.101	1	5.101	2.637	.105	.006
Error	793.268	410	1.935			
Total	1049.000	414				
Corrected Total	824.302	413				

a. R Squared = .038 (Adjusted R Squared = .031)

Table 47 *Number of errors in attention check questions dependent on the presence of disciplining reminder [verr] adjusted for gender [sex] and age, data set B3*

Women made, on average, more errors than men in both groups (interaction between gender and experimental group was not significant).

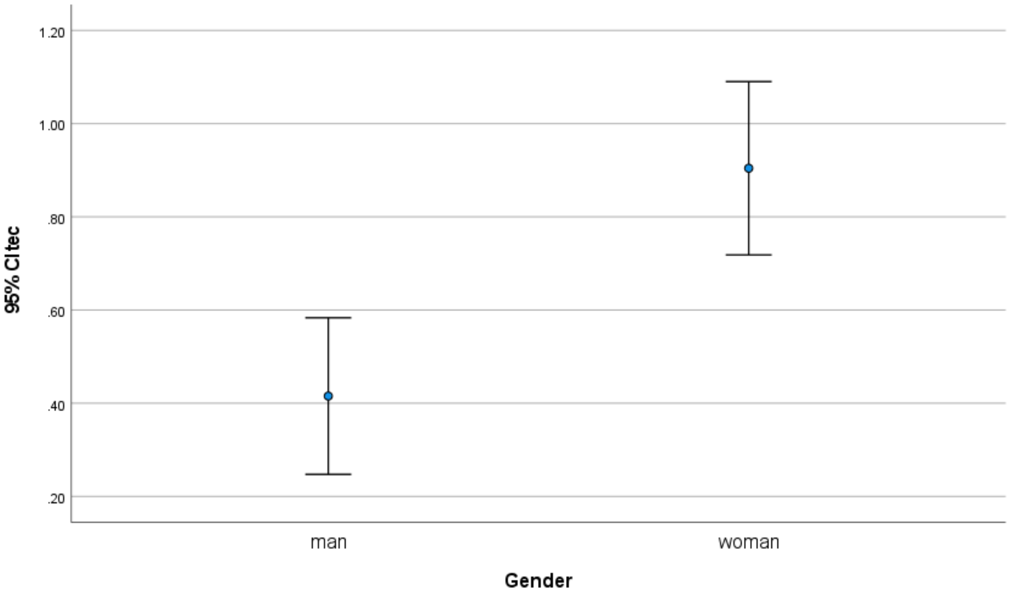


Figure 25 *Errorbar graph showing the difference between men and women number of errors in data set B2*

Another interesting aspect of this analysis is that further inquisition on the matter of gender differences has shown, after analysis of warning sign #4, that women are declared to be more attentive than men, despite making more errors. This surprising result was not present in the previous study, used for ad hoc comparison retrospectively, so it may be accidental. Nonetheless, gender differences are beyond the scope of this dissertation and will not be discussed further for this reason.

References

Albaum, G., Wiley, J., Roster, C., & Smith, S. M. (2011). Visiting Item Non-responses in Internet Survey Data Collection. *International Journal of Market Research*, 53(5), 687–703. https://doi.org/10.2501/IJMR-53-5-687-703

Allen, M. (2017). The sage encyclopedia of communication research methods (Vols. 1-4). Thousand Oaks, CA: SAGE Publications. https://doi.org/10.4135/9781483381411

Alvarez, M. R., 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

Anseel, F., Lievens, F., Schollaert, E. & Choragwicka, B. (2010). Response Rates in Organizational Science, 1995–2008: A Meta-analytic Review and Guidelines for Survey Researchers. *Journal of Business and Psychology, 25*(3), 335-349. http://dx.doi.org/10.1007/s10869-010-9157-6

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*, 527–535. https://doi.org/10.3758/s13428-012-0265-2

Baer, R. A., Ballenger, J., Berry, D. T. R. & Wetter, M. W. (1997) Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*, 139-151.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., & Zahs, D. (2010). Research synthesis: AAPOR report on Online Panels. Public Opinion Quarterly, 74(4), 711–781. https://doi.org/10.1093/poq/nfq048

Barakat, L. L., Lorenz, M. P., Ramsey, J. R., & Cretoiu, S. L. (2015). Global managers: An analysis of the impact of cultural intelligence on job satisfaction and performance. *International Journal of Emerging Markets*, *10*(4), 781–800. https://doi.org/10.1108/IJOEM-01-2014-0011

Bassett, J., Cleveland, A., Acorn, D., Nix, M., & Snyder, T. (2017). Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations. *Assessment & Evaluation in Higher Education*, *42*(3), 431–442. https://doi.org/10.1080/02602938.2015.1119801

Batorski, D. & Olcoń-Kubicka, M. (2006). Prowadzenie badań przez Internet – podstawowe zagadnienia metodologiczne. *Studia Socjologiczne, 3*(182), 99-132.

Beatty, P. & Herrmann, D. (2002). To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse. In Groves, R. M., Dillman, D. A., Eltinge, J. L. & Little R. J. A. (Eds.), Survey Nonresponse (pp. 71-85). New York: Wiley.

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-Fit as an Index of Inattentive Responding: A Comparison of Methods Using Polytomous Survey Data. *APPLIED PSYCHOLOGICAL MEASUREMENT, 43*(5), 374–387. https://doi.org/10.1177/0146621618798666

Bell, D. E., Raiffa, H. & Tversky, A. (1988). Descriptive, Normative, and Prescriptive Interactions in Decision Making. In D. E. Bell, H. Raiffa, and A. Tversky (eds.), Decision Making — Descriptive, Normative, and Prescriptive Interactions (pp. 9–30). New York: Cambridge University Press.

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*, 739–573. https://doi.org/10.1111/ajps.12081

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

Bowling, N. A., & Huang, J. L. (2018). Your Attention Please! Toward a Better Understanding of Research Participant Carelessness. *Applied Psychology: An International Review, 67*(2), 227–230. https://doi.org/10.1111/apps.12143

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the Questions Ever End? Person-Level Increases in Careless Responding During

Questionnaire Completion. *Organizational Research Methods, 24*(4), 718–738. https://doi.org/10.1177/1094428120947794

Brown, R. V. (1968). Evaluation of total survey error. *Statistician, 17*(4), 335-343.

Brown, R. V. (1969). *Research and the credibility of estimates*. Boston, MA: Harvard University, Graduate School of Business Administration, Division of Research.

Brown, R. V. (1992). The state of the art of decision analysis: A personal perspective. *Interfaces, 22*, 5-14.

Brown, R. V. & Vari, A. (1992). Towards an agenda for prescriptive decision research: The normative tempered by the descriptive. *Acta Psychologica, 80*, 33-47.

Buchanan, E. A., & Hvizdak, E. E. (2009). Online Survey Tools: Ethical and Methodological Concerns of Human Research Ethics Committees. Journal of Empirical Research on *Human Research Ethics, 4*(2), 37–48

Buchanan, E., & Scofield, J. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods, 50*(6), 2586–2596. https://doi.org/10.3758/s13428-018-1035-6

Carver, R.P. (1992), Reading rate: Theory, research and practical implications. *Journal of Reading, 36*, 84–95.

Christensen, A. I., Ekholm, O., Glümer, C., & Juel, K. (2014). Effect of survey mode on response patterns: comparison of face-to-face and self-administered modes in health surveys. *European Journal of Public Health, 24*(2), 327–332.

Cichomski, B., & Morawski, P. (1996). *Polski Generalny Sondaż Społeczny: Struktura skumulowanych danych, 1992-1995*. Warszawa: Instytut Studiów Społecznych UW.

Conijn, J. M., van der Ark, L. A., & Spinhoven, P. (2020). Satisficing in Mental Health Care Patients: The Effect of Cognitive Symptoms on Self-Report Data Quality. *Assessment, 27*(1), 178–193. https://doi.org/10.1177/1073191117714557

Conrad, F. G., Tourangeau, R., Couper, M. P., Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods, 11*(1), 45-61.

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (206-261). New York: Free Press

Couper, M. P., Tourangeau, R., Conrad, F. G., Crawford, S. D. (2004). What They See Is What We Get: Response Options for Web Surveys. *Social Science Computer Review, 22*(1), 111-127.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*, 596–612.

Curran, P. G., Kotrba, L., Denison, D. (2010) Careless responding in surveys: applying traditional techniques to organizational settings. 25th annual conference of Society for Industrial and Organizational Psychology, Atlanta, GA.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66,* 4-19.

Dennis, S., Goodson, B., & Pearson, C. (2019). Online Worker Fraud and Evolving Threats to the Integrity of MTurk Data: A Discussion of Virtual Private Servers and the Limitations of IP-Based Screening Procedures. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3233954

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*(5), 559-577.

Dewberry, C., Davies-Muir, A., & Newell, S. (2013). Impact and Causes of Rater Severity/Leniency in Appraisals Without Postevaluation Communication between Raters and Ratees. *International Journal of Selection and Assessment*, *21*(3), 286-293.

Dodou, D. & de Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior, 36*, 487-495. https://doi.org/10.1016/j.chb.2014.04.005

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology, 33*(1), 105-121.

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*. https://doi.org/10.1177/1073191120957102

European Society for Opinion and Market Research (ESOMAR). (2013). Global Market Research 2013 [online]. Available at: https://www.esomar.org/uploads/industry/reports/global-market-research-2013/ESOMAR-GMR2013-Preview.pdf

European Society for Opinion and Market Research (ESOMAR). (2014). Global Marketing Research 2014: an ESOMAR industry report [online]. Available at: https://www.esomar.org/uploads/industry/reports/global-market-research-2014/ESOMAR-GMR2014-Preview.pdf [Accessed: September 5, 2018]

Eysenbach, G. & Wyatt, J. (2002). Using the Internet for Surveys and Health Research. J Med Internet Res., 4(2), e13. Published online 2002 Nov 2 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1761932/ [Accessed: September 5, 2018].

Fazio, R. H. (1986) How do attitudes guide behavior? In Sorrentino & Higgins (Eds.) *Handbook of motivation and cognition* (pp. 204-243). New York: Guilford.

Fiske, S. T. & Kinder, D. R. (1981). Involvement, expertise, and schema use: evidence from political cognition. In: N. Cantor & J. Kihlstrom (Eds.), *Personality, cognition, and social interaction* (pp. 171-190). Hilsdale, NJ: Erlbaum.

Forgas, J.P., Vargas, P.T. (2005). Wpływ nastroju na społeczne oceny i rozumowanie. In: M. Lewis, J.M. Haviland-Jones (eds.). *Psychologia emocji*. Gdańsk: GWP.

Fronczyk, K. (2014) The identification of random or careless responding in questionnaires: The example of the NEO-FFI [Identyfikacja odpowiadania losowego lub nieuważnego w kwestionariuszu na przykładzie NEO-FFI]. *Roczniki Psychologiczne, 17*(2), 439-473.

Galesic, M., Tourangeau, R., Couper, M. P. & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72*(5), 892-913.

Gibson, A. M. & Bowling, N. A. (2020). The Effects of Questionnaire Length and Behavioral Consequences on Careless Responding. European Journal of Psychological Assessment, 36(2), 410-420. https://doi.org/10.1027/1015-5759/a000526

Gigliotti, L. & Dietsch, A. (2014). Does Age Matter? The Influence of Age on Response Rates in a Mixed-Mode Survey. *Human Dimensions of Wildlife, 19*(3), 280-287. https://doi.org/10.1080/10871209.2014.880137

Gittelman, S., & Trimarchi, E. (2009). Variance between purchasing behavior profiles in a wide spectrum of online sample sources. White Paper, Marketing. Inc.

Gogami, M., Matsuda, Y., Arakawa, Y. & Yasumoto, K. (2021) Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone. IEEE Access(99), 1-1. https://doi.org/10.1109/ACCESS.2021.3069049

Goldammer, P., Annen, H., Stöckli, P. L. & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly, 31*(4). https://doi.org/10.1016/j.leaqua.2020.101384

Górecki, M. A. (2011). Electoral Salience and Vote Overreporting: Another Look at the Problem of Validity in Voter Turnout Studies. *International Journal of Public Opinion Research, 23*(4), Winter 2011, 544–557. https://doi.org/10.1093/ijpor/edr023

Göritz, A. S., Reinhold, N., & Batinic, B. (2002). Online panels. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online Social Sciences* (pp. 27–47).

Göritz, A. S. (2010). Using online panels in psychological research. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), Oxford Handbook of Internet Psychology.

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the Effects of Removing 'Too Fast' Responses and Respondents from Web Surveys. *Public Opinion Quarterly, 79*(2), 471–503. https://doi.org/10.1093/poq/nfu058

Grice, P. H. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, Vol. III, Speech acts* (pp. 41–58). Academic Press.

Gummer, T., Roßmann, J., & Silber, H. (2018). Using Instructed Response Items as Attention Checks in Web Surveys: Properties and Implementation. *Sociological*

Methods and Research, *50*(1), 238–264. https://doi.org/10.1177/0049124118769083

Hauser, D. J., Sunderrajan, A., Natarajan, M., & Schwarz, N. (2017). Prior Exposure to Instructional Manipulation Checks does not Attenuate Survey Context Effects Driven by Satisficing or Gricean Norms. *Methods, Data, Analyses*, *10*(2), 195–220. https://doi.org/10.12758/MDA.2016.008

Hensel, P. G. (2021). Reproducibility and replicability crisis: How management compares to psychology and economics – A systematic review of literature. European Management Journal, article in press. https://doi.org/10.1016/j.emj.2021.01.002

Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in non-probability online panels. Online panel research: A data quality perspective, 1, 219-237.

Holland, J. L., & Christian, L. M. (2009). The Influence of Topic Interest and Interactive Probing on Responses to Open-Ended Questions in Web Surveys. *Social Science Computer Review, 27*(2), 196–212. https://doi.org/10.1177/0894439308327481

Hoyt, W.T. (2000). Rater bias in psychological research: when is it a problem and what can we do about it? *Psychological Methods, 5*, 64-86.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015a). Detecting Insufficient Effort Responding with an Infrequency Scale: Evaluating Validity and Participant Reactions. Journal of Business and Psychology, 30(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

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.

Huang, J. L., & DeSimone, J. A. (2021). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology, 36*(5), 807-828.

Huang, J. L., Liu, M., Bowling, N. A. (2015b) Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828-845.

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Five - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV5.jsp. Madrid: JD Systems Institute.

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Six - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Madrid: JD Systems Institute.

Ioannidis J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jerzyński, T. (2009). Wybrane korelaty liczby odpowiedzi beztreściowych w badaniach sondażowych [Unpublished doctoral dissertation]. University of Warsaw.

Johnson, J. A. (2005) Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*(1), 103-129.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kam, C. C. S., Meyer, J.P. (2015) How Careless Responding and Acquiescence Response Bias Can Influence Construct Dimensionality: The Case of Job Satisfaction. *Organizational Research Methods, 18*(3), 512-541.

Kane, J. V., & Barabas, J. (2019). No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments. *American Journal of Political Science, 63*(1), 234–249. https://doi.org/https://doi.org/10.1111/ajps.12396

Kasvi, A. (2017). *Employee satisfaction survey: Reippailuhalli Huimala* [Haaga-Helia ammattikorkeakoulu]. http://www.theseus.fi/handle/10024/128824

Keeley, J. W., Webb, C., Peterson, D., Roussin, L., & Flanagan, E. H. (2016). Development of a Response Inconsistency Scale for the Personality Inventory for

DSM – 5 Development of a Response Inconsistency Scale for the Personality Inventory for. *Journal of Personality Assessment*, *98*(4), 351–359. https://doi.org/10.1080/00223891.2016.1158719

Kiesler, S., Sproull, L. S. (1986) Response Effects in the Electronic Survey. *Public Opinion Quarterly 50*(3), 402–413

Kordos, J. (1988). Jakość danych statystycznych. Warszawa: PWE.

Kountur, R. (2016) Detecting careless responses to self-reported questionnaires. *Egitim Arastirmalari - Eurasian Journal of Educational Research, (64)*, 307-318.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. (2002). The Causes of No-Opinion Responses to Attitude Measures in Surveys: They are Rarely What They Appear to be. In Groves, Dillman, Eltinge and Little (Eds.) *Survey Nonresponse* (pp. 87-100). New York: Wiley.

Krosnick, J.A., Holbrook, A.L., Berent, M.K., Carson, R.T., Hanemann, W.M., Kopp, R.J., Mitchell, R.C., Presser, S., Ruud, P.A., Smith, V.K., Moody, W.R., Green, M.C., Conaway, M. (2002). The impact of „no opinion" response options on data quality: non–attitude reduction or an invitation to satisfice? *The Public Opinion Quarterly, 66*(3), 371–403.

Krosnick, J. A. & Alwin, D. F. (1989). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52*, 526-538.

Krosnick, J. A. & Presser, S. (2010). Questionnaire design. In: J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). West Yorkshire, England: Emerald Group.

Król, G., & Kowalczyk, K. (2014). Ewaluacja projektów i abstraktów - wpływ indywidualnego stylu ewaluacji na oceny. *Problemy Zarządzania, 1/2014*(45), 137-155.

Kuha, J., Butt, S., Katsikatsou, M. & Skinner, C. J. (2017). The Effect of Probing "Don't Know" Responses on Measurement Quality and Nonresponse in Surveys. *Journal of the American Statistical Association*, *113*(521), 26-40. https://doi.org/10.1080/01621459.2017.1323640

Kung, F. Y. H., 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

Kurtz, J. E., Parrish, C. L. (2001) Semantic response consistency and protocol validity in structured personality assessment: the case of the NEO-PI-R. *Journal of Personality Assessment, 76*(2), 315-32.

Kuźmińska, A. O. & Pazura, D. (2018). The impact of Control Preferences Fit Between Employees and Their Supervisors on Employee Job Satisfaction. *Studia i Materiały, 29*(2), 18-32.

Kuźmińska, A. O., Schulze, D. & Koval, A. (2019). Who Doesn't Want to Share Leadership? The Role of Personality, Control Preferences, and Political Orientation in Preferences for Shared vs. Focused Leadership in Teams. In Kuźmińska, A. (ed.), Management Challenges in the Era of Globalization (pp. 11–27). Warsaw: University of Warsaw Faculty of Management Press.

Kwak, N. & Radler, B. (2002). A Comparison Between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality. *Journal of Official Statistics, 18*(2), 257-273.

Lake, C. J., Withrow, S., Zickar, M. J., Wood, N. L., Dalal, D. K., Bochinski, J. (2013) Understanding the Relation Between Attitude Involvement and Response Latitude Using Item Response Theory. *Educational and Psychological Measurement, 73*(4), 690-712.

Levi, R., Ridberg, R., Akers, M., & Seligman, H. (2021). Survey Fraud and the Integrity of Web-Based Survey Research. *American Journal of Health Promotion*. Advance online publication. https://doi.org/10.1177/08901171211037531

Lu, J., Wang, F., Wang, X., Lin, L., Wang, W., Li, L., & Zhou, X. (2019). Inequalities in the health survey using validation question to filter insufficient effort responding: Reducing overestimated effects or creating selection bias? *International Journal for Equity in Health, 18*(1). https://doi.org/10.1186/s12939-019-1030-2

McGrath, R. E., Mitchell, M., Kim, B. H., Hough, L. (2010) Evidence for Response Bias as a Source of Error Variance in Applied Assessment. *Psychological Bulletin, 136*(3), 450-470.

McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior, 84*, 295–303. https://doi.org/10.1016/J.CHB.2018.03.007

McKibben, W. B., & Silvia, P. J. (2017). Evaluating the Distorting Effects of Inattentive Responding and Social Desirability on Self-Report Scales in Creativity and the Arts. *The Journal of Creative Behavior, 51*(1), 57–69. https://doi.org/10.1002/jocb.86

Meade, A. W., Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455.

Merckelbach, H., Giesbrecht, T., Jelicic, M., Smeets, T. (2010) The problem of careless respondents in surveys [Kretenzers in surveys: Het probleem van onzorgvuldige respondenten]. *Tijdschrift voor Psychiatrie, 52*(9), 663-669.

Merritt, S. M. (2012) The Two-Factor Solution to Allen and Meyer's (1990) Affective Commitment Scale: Effects of Negatively Worded Items. *Journal of Business and Psychology, 27*(4), 421-436.

Meyer, J. F., Faust, K. A., Faust, D., Baker, A. M., Cook, N. E. (2013) Careless and Random Responding on Clinical and Research Measures in the Addictions: A Concerning Problem and Investigation of their Detection. *International Journal of Mental Health and Addiction, 11*(3), 292-306.

Michałowicz, B. (2016) Ankiety ewaluacyjne w szkolnictwie wyższym: wpływ wyboru ewaluatorów (Doctoral dissertation). [Access on the author's request] https://depotuw.ceon.pl/handle/item/1532

Mitchell, A. L., Hegedüs, L., Žarković, M., Hickey, J. L., & Perros, P. (2021). Patient satisfaction and quality of life in hypothyroidism: An online survey by the British thyroid foundation. *Clinical Endocrinology*, *94*(3), 513–520. https://doi.org/10.1111/cen.14340

Moxey, L. M. & Sanford, A. J. (1986). Quantifiers and Focus. *Journal of Semantics, 5*(3), 189-206. https://doi.org/10.1093/jos/5.3.189

Moxey, L. M. & Sanford, A. J. (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology 14*(3),237–55.

Moździerz, T. (2019). Tempo czytania rodzimych i nierodzimych użytkowników polszczyzny. *Języki Obce W Szkole, 63*(4), 79–85. http://jows.pl/sites/default/files/wydania/jows_4_2019_online-calosc_0.pdf

Murphy, F., Macpherson, K., Jeyabalasingham, T., Manly, T., & Dunn, B. (2013). Modulating mind-wandering in dysphoria. *Frontiers in Psychology, 4*(NOV). https://doi.org/10.3389/fpsyg.2013.00888

Nancarrow, C., & Cartwright, T. (2007). Online access panels and tracking research: The conditioning issue. *International Journal of Market Research, 49*, 573–594.

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. PLoS One, 11(6), e0157732. https://doi.org/10.1371/journal.pone.0157732.

Nichols, A. L., & Edlund, J. E. (2020). Why don't we care more about carelessness? Understanding the causes and consequences of careless participants. *International Journal of Social Research Methodology, 23*(6), 625–638. https://doi.org/10.1080/13645579.2020.1719618

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Pratt, J. W., Raiffa, H., and Schlaifer, R. (1995). *Introduction to statistical decision theory*. Cambridge, MA: MIT Press.

Revilla, M. (2012). Impact of the Mode of Data Collection on the Quality of Answers to Survey Questions Depending on Respondent Characteristics. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 116*(1), 44–60. https://doi.org/10.1177/0759106312456510

Robert Zajonc Institute for Social Studies. (2013). Polish General Social Surveys 1992-2010. Study documentation. Retrieved July 23, 2021, from http://www.ads.org.pl/pobieranie-zbioru-danych.php?id=91

Roivainen, E., Veijola, J. & Miettunen, J. (2016). Careless responses in survey data and the validity of a screening instrument. *Nordic Psychology, 68*(2), 114-123. https://doi.org/10.1080/19012276.2015.1071202

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology, 6*(3), 376–400. https://doi.org/10.1093/JSSAM/SMX020

Queloz, S., & Etter, J.-F. (2019). An online survey of users of tobacco vaporizers, reasons and modes of utilization, perceived advantages and perceived risks. *BMC Public Health, 19*(1), 1–11. https://doi.org/10.1186/S12889-019-6957-0

Sammelman, K., Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior research methods, 50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Schlaifer, R. (1969). *Analysis of decisions under uncertainty*. New York: McGraw-Hill.

Schneider, S., May, M., & Stone, A. A. (2018). Careless responding in internet-based quality of life assessments. *Quality of Life Research, 27*(4), 1077–1088. https://doi.org/10.1007/s11136-017-1767-2

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

Schuman, H. & Presser, S. (1981). Questions and answers in attitude surveys. New York: Academic Press

Shin, E., Johnson, T. P., & Rao, K. (2012). Survey Mode Effects on Data Quality: Comparison of Web and Mail Modes in a U.S. National Panel Survey. *Social Science Computer Review, 30*(2), 212–228. https://doi.org/10.1177/0894439311404508

Silber, H., Danner, D., & Rammstedt, B. (2019). The impact of respondent attentiveness on reliability and validity. *International Journal of Social Research Methodology, 22*(2), 153–164. https://doi.org/10.1080/13645579.2018.1507378

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632*

Skarżyńska, K., Urbańska, B., & Radkiewicz, P. (2021). Under or Out of Government Control? The Effects of Individual Mental Health and Political Views on the Attribution of Responsibility for COVID-19 Incidence Rates. *Social Psychological Bulletin, 16*(1), e4395. https://doi.org/10.32872/spb.4395

Smith, R., & Brown, H. H. (2006). Data and panel quality: Comparing metrics and assessing claims. In: Proceedings of the ESOMAR Panel Research Conference.

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/https://doi.org/10.1111/emip.12256

Sułek, A. (2001). *Sondaż polski: Przygarść rozpraw o badaniach ankietowych*. Warszawa: Wydawnictwo Instytutu Filozofii i Socjologii PAN.

Sztabiński, F., Sztabiński, P. B. & Sawiński, Z. (2005). *Fieldwork jest sztuką : jak dobrać respondenta, skłonić do udziału w wywiadzie, rzetelnie i sprawnie zrealizować badanie: praca zbiorowa*. Warszawa: Wydawnictwo Instytutu Filozofii i Socjologii PAN.

Toepoel, V., Das, M., & Van Soest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly, 72*, 985–1007.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). The Science of Web Surveys. In *The Science of Web Surveys*. Oxford University Press.

Tourangeau, R., Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin, 103*(3), 299-314. https://doi.org/10.1093/acprof:oso/9780199747047.001.0001

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Ulrich, D. (1998) Measuring human resources: An overview of practice and a prescription for results. *Human Resource Management, 36*(3), 303-320.

Vehovar, V., Lozar Manfreda, K. (2008). Overview: Online Surveys. In Fielding, N.; Lee, R. M.; Blank, G. The SAGE Handbook of Online Research Methods. London: SAGE. pp. 177–194. https://doi.org/10.3102/0013189X211040054

Verbree, A.-R., Toepoel, V., & Perada, D. (2020). The Effect of Seriousness and Device Use on Data Quality. Social Science Computer Review, 38(6), 720–738. https://doi.org/10.1177/0894439319841027

Weathers, D., Bardakci, A. (2015) Can response variance effectively identify careless respondents to multi-item, unidimensional scales? *Journal of Marketing Analytics, 3*(2), 96-107.

Weijters, B., Baumgartner, H., Schillewaert, N. (2013) Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320-334.

Wieczorkowska, G (1993). Pułapki statystyczne. W: M.Z. Smoleńska (red.). Badania nad rozwojem w okresie dorastania. Warszawa: Instytut Psychologii PAN.

Wieczorkowska-Nejtardt, G. (1998). Inteligencja motywacyjna: mądre strategie wyboru celu i sposobu działania. Warszawa: Wydawnictwa Instytutu Studiów Społecznych UW.

Wieczorkowska-Wierzbińska, G. (2011) Psychologiczne ograniczenia. WN WZ UW, Warszawa.

Wieczorkowska-Wierzbińska, G. (2014). Diagnoza psychologiczna predyspozycji pracowników. *Problemy Zarządzania, 12*(1), 81-98.

Wieczorkowska-Wierzbińska, G. (2022). Zarządzanie ludźmi – z psychologicznego i metodologicznego punktu widzenia. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego

Wieczorkowska, G., & Kowalczyk, K. (2021). Ensuring Sustainable Evaluation: How to Improve Quality of Evaluating Grant Proposals? *Sustainability, 13*, 2842.

Wieczorkowska, G. & Król, G. (2016). Ten pitfalls of research practices in management science. *Problemy Zarządzania, 2*(2), 173-187.

Wieczorkowska, G., Król, G. Wierzbiński, J. (2015). Metody Ilościowe. W: Kostera, M. (red.) Metody badawcze w zarządzaniu humanistycznym, WA Sedno, Warszawa.

Wieczorkowska, G., Król, G., Wierzbiński, J. (2016). Cztery metodologiczne zagrożenia w naukach o zarządzaniu. *Studia i Materiały, 2*(2), 146-156.

Wieczorkowska, G., & Wierzbiński, J. (2005). *Badania sondażowe i eksperymentalne. Wybrane zagadnienia.* Warszawa: Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego.

Wieczorkowska, G. & Wierzbiński, J. (2011). Statystyka. Od teorii do praktyki. Warszawa: SCHOLAR.

Wieczorkowska-Wierzbińska, G., Wierzbiński, J., Kuźmińska, A. (2014). Porównywalność danych zebranych w różnych krajach. *Psychologia Społeczna 9*(2), 128-143.

Wood, D., Harms, P. D., Lowman, G. H., DeSimone, J. A. (2017) Response Speed and Response Consistency as Mutually Validating Indicators of Data Quality in Online Samples. *Social Psychological and Personality Science, 8*(4), 454-464.

Yan, T., Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology, 22*(1), 51-68.

Zhang, C., & Conrad, F. G. (2014). Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods, 8*(2), 127–135. https://doi.org/https://doi.org/10.18148/srm/2014.v8i2.5453

Zieliński, M. (2009). Analiza hierarchiczna w szacowaniu wpływu ankietera na odpowiedzi respondenta. *Problemy Zarządzania, 4*, 238–253.

List of figures

List of tabels