Paradata: An economic source for conscientiousness indicators?

Zur Erlangung des

Mastergrades

(MSc)

an der Naturwissenschaftlichen Fakultät

Der Paris-Lodron Universität Salzburg

Eingereicht von

Gerrit Hendrik Hüller

01522522

Gutachterin: Univ.-Prof. Dr. Tuulia Ortner

Fachbereich: Psychologie

Salzburg, Januar, 2022

Content

Summary

Paradata offer a novel and inexpensive approach in assessing behavior in computerized testing (e.g., online surveys). They are automatically generated and comprise response times, (mouse/touch) click accuracy and the used device among many others. Paradata might also come in handy for personality assessment, and when addressing careless responding. Personality in general is an important predictor for future behavior and the personality trait conscientiousness is important for job performance and academic success. Consequently, the research question arises, if paradata contain economic indicators for conscientiousness. This thesis was based on an online survey, where participants' behavior was tracked via Java Script and saved as paradata. The paradata indicators average response time over several survey pages, the variability in response time across pages, as well as average (mouse) clicking inaccuracy and the variability in clicking inaccuracy across those pages were created. Each of these indicators comprised behavior, which was argued to belong to the domain of conscientiousness. Consequently, their relationship to both a self-report and a behavioral measure for conscientiousness, based on retrieval questions about the informed consent, were analyzed. While the average response time was assumed to show an inverted U-shaped relationship to the conscientiousness measures, all other paradata indicators were assumed to be negatively correlated with both conscientiousness measures. Due to highly skewed data and so-called careless responders, whose insufficient effort in responding is both empirically and theoretically negatively associated with conscientiousness, log-transformation of the response time as well as a conservative outlier analysis was conducted. In addition to that, all analyses were conducted with an outlier free sample including careless responders and a clean sample with neither outliers nor careless responders. The behavioral measure for conscientiousness failed to correlate with the self-report measure, questioning the assessed construct. Only the log-transformed average response time per survey page correlated with r (87) = -.33 with self-reported conscientiousness, when careless responders were excluded. This is contradictory to the assumed curvilinear inverted U-shaped effect but is assumed to be due the conservative outlier analysis. Further, all other paradata indicators failed to explain any conscientiousness measure. The variability in response time offered no additional information to the average response time explaining conscientiousness measures. Both paradata indicators about clicking inaccuracy failed to correlate with conscientiousness, which possibly was caused by the used device (computer mouse vs. touchscreen). Overall, much future research is needed examining paradata as a source for conscientiousness indicators. But with a proper data preparation and outlier analysis the already found relevance of response times is promising.

Introduction

Since the earliest days of psychology there has always been interest in interindividual differences like (cognitive) abilities and personality (e.g., Asendorpf, 2011). As personality describes a person's propensity to show specific behaviors, feelings and thoughts on a regular basis, personality is just by definition a predictor of one's experience and future behavior (e.g., Stanek & Ones, 2018).

There are different approaches to assess personality. One way is the self-report by a questionnaire, where multiple personality dimensions can be assessed simultaneously. Even non-observable aspects can be assessed and often there is no need for supervision. In addition, there are many assessment tools and norms for comparison available (Schmidt-Atzert & Amelang, 2012 p. 240 & 241). Another possible approach is the diagnostic interview, which has the same advantages as a questionnaire but needs supervision (i.e., an interviewer). Though none of those approaches work if the assessed person has limited or no self-awareness. Both approaches are prone to self-deception of the assessed person and deliberate faking. Further, their informative value suffers when the assessed person responds with insufficient effort (Schmidt-Atzert & Amelang, 2012, p. 241; e.g., Arthur, Hagen, & George, 2020). There is research (Arthur et al., 2020), which links deliberate faking to high-stake situations like prehire assessments and careless responding out of boredom to low-stake situations. Both practices are highly problematic as they taint data by adding error variance, which in return alters results in analyses (e.g., Meade & Craig, 2012; Leiner, 2019). Consequently, this makes results less replicable, fuel the replication crisis and compromises a field of study.

Behavioral observations as an assessment approach are also susceptible to faking, when the observed persons know what they are observed for. Further, the development and execution of such assessment is very complex, expensive and it can only be used on a limited number of persons and personality dimensions simultaneously (Schmidt-Atzert & Amelang, 2012 p. 241). Therefore, only few standardized assessments are available, which also means that hardly any norms for comparison are available. One approach to overcome at least the bias through faking is a modified version of behavioral observations, which is called objective personality tests (OPT).

OPTs assess a construct by seemingly offering a cognitive performance test. But instead of assessing the performance itself, observational data are gathered to assess a dimension of personality. By disguising the true goal (i.e., assessing one's personality) a participant can not deliberately or unintentionally bias the results unless he/ she knows this test. But there are also downsides for OPTs: To create comparative norms, the test situation must be highly

standardized (Koch, Ortner, Eid, Caspers & Schmitt, 2014; Schmidt-Atzert & Amelang, 2012), which is why such tests are computerized. In addition, OPTs are difficult and expensive to develop. Further, the content of OPTs needs to be protected, as it is essential, that the participants do not know the test. OPTs are not without critic either, as they seem to assess trait components which are not shared with other measures like self-report questionnaires. Despite this so-called method specificity, criterion validity is nevertheless given (Koch et al., 2014). In conclusion, OPTs offer a good approach to assess personality without worrying about faking. On the downside, OPTs are less economical: They are expensive and complex to develop and norm, their test situations have to be highly standardized, and they need special test protection, which makes their usage especially in (online) research difficult. But in sum, gathering observational data unnoticed seems to be an efficient way to overcome some problems arising from self-reported data. There might be also an approach to overcome the economical hurdles of an OPT, while keeping its advantages: using paradata.

Paradata are automatically generated data in computerized testing and comprise response times, response behavior (e.g., accuracy of mouse clicks on answers) or information about the operating system among other aspects (e.g., Kroehne & Goldhammer, 2018). In contrast to the highly standardized OPTs, paradata can capture participants' natural behavior in a wide range of high- or low-stake test situations. Because paradata are automatically generated and therefore "garbage" output, they might be a very economic source of information.

As noted earlier, personality is an important predictor for one's behavior, thoughts, and feelings. The personality dimension of conscientiousness comprises a broad domain of characteristics like being reliable, planful and prone to delay gratification for a specific goal (Roberts, Jackson, Fayard, Edmonds & Meints, 2009; Friedman, Schuhstack & Rindermann, 2004 in Schreiber & Iller, 2016). Those characteristics are of special importance for academic and professional success (e.g., Asendorpf, 2011). In a meta-analysis, Barrick and Mount (1991) report a correlation of $\rho = .22$ between conscientiousness and job performance across different occupations. In another meta-analysis, Trapman, Hell, Hirn and Schuler (2015) report $\rho = .27$ between conscientiousness and grades as an indicator for academic success. These population corrected effect sizes underline the importance and the predictive power of conscientiousness in general and for study admission tests as well as for prehire assessments in particular.

While personality assessment is either susceptible for faking or is expensive, its relevance for society's aptitude issues in higher education or job personal selection is without question. As paradata offer an approach to address both faking and expenses, the question rises, if paradata contain information about personality dimensions like conscientiousness. In this

thesis, paradata indicators like response time, mouse click accuracy and response consistency are to be linked to the personality dimension conscientiousness.

This leads to the research question:

Can survey paradata be used as economic behavioral observations indicating conscientiousness?

Theoretical Background

The following sections comprise the theoretical background and the current state of research regarding paradata, conscientiousness and the phenomenon of lacking conscientiousness in (online) surveys: careless responding. All those topics are put into context with regard to the research question, from which the hypotheses are derived.

Paradata

Paradata or logging (log) data are a by-product of computer usage, as a user's input information are saved or logged in order to process the input. Regarding computerized surveys, Callegaro (2013) describes paradata as process data, which are generated by the participants themselves and their interaction with the survey. Kroehne and Goldhammer (2018) in contrast propose a more detailed taxonomy of paradata. They suggest three types of paradata categories, which they emphasize are non-exhaustive for the literature.

Taxonomy of paradata

Access-related paradata describe information about the used device, like using a smartphone or a computer for answering an online survey. They also comprise the software environment, like using a Windows or Apple device or which browser software is used. Another aspect of this category is the setting. What is the location of this person? What is the environment (e.g., the volume in a place)? The last aspect within access-related paradata are contact information like ID or the participation status of the survey.

Process-related paradata can be divided into micro and macro process data. In micro process data actions like zooming or scrolling on a single survey page are gathered. Also, the order of responses and the response time for single entities (i.e., survey pages) belong to this subcategory. Macro process data in contrast comprise information about the process of the whole survey i.e., over several survey pages. For example, logging a person's behavior switching back and forth between survey pages belong to macro process data. This also includes the overall time needed to complete the whole survey.

The third category are response-related paradata. These can be divided in answer data and input data. Answer data comprise changes in the answer options after already selecting one answer before, as well as response times on item level. Input data consist of mouse clicks or touch events (e.g., with smartphones) as well as keystrokes or a user's deliberate input as a response on an item.

While those categories refer to the type of a single unit of paradata describing an event, log data normally comprise many different events and therefore contain a mix of those mentioned. For some research questions single events suffice, while for other research questions composition of multiple events are necessary. In this thesis a combination of several events, namely time series, are relevant and are extracted from multiple events (Kroehne & Goldhammer, 2018).

Paradata in current research

Previous research predominantly focused on paradata and its possibilities addressing the nonresponder error (Kreuter & Olson, 2013) and total survey error (Kreuter, 2013), while current research shifts its interests to personality assessment and its opportunities through computational power (Vinciarelli, 2014; Wight, 2014, Bleidorn & Hopwood, 2019). Diedenhofen and Musch (2017) used paradata to detect cheating in an online achievement test. Cheating in tests and academic dishonesty is associated with the personality dimension conscientiousness (Heny & Montargot, 2019; Day, Hudson, Dobies, & Waris, 2011; de Bruin, 2006), which can also be argued by theory. The next section covers the personality dimension conscientiousness and its broad domain of behaviors.

Conscientiousness

Conscientiousness is a personality trait and a dimension of different personality theories like the Big-5 model of personality or the HEXACO model. Overall, conscientiousness expresses itself as a person's propensity to control their impulses relative to social norms, their ability to delay gratification, their planning ability, and their tendency to be goal directed (Roberts, Jackson et al., 2009). Further, conscientious people are careful, reliable, and responsible (Friedman et al., 2004 in Schreiber & Iller, 2016). In sum, conscientiousness sometimes is described as the lack of impulsivity. From these short descriptions alone it gets clear, that conscientiousness is less a stand-alone and clear-cut trait than more an agglomeration of single traits which lie in the spectrum of the conscientiousness trait (e.g., Roberts, Lejuez, Krueger, Richards & Hill, 2012). This family of "sub-traits" is structured hierarchically and can be differentiated between traits and states at the highest level.

An Outline of the Structure of Personality: Trait vs. State

Personality traits are defined as a person's propensity to show similar patterns in behavior, thoughts, and feelings in trait-affording situations in a relatively consistent manner over time (e.g., Roberts et al., 2009; Herzberg & Roth, 2014; Roberts et al., 2012). In contrast to traits, the so-called states are short-term, concrete patterns of acting, feeling, and thinking (Heller, Komar & Lee, 2007). Buss and Craik (1983) proposed that the expression of a personality trait is the frequency of its states. In other words, showing a lot of conscientious state behavior, thoughts and feelings indicates a highly conscientious trait. From this more historical point of view a personality trait is more the quantity than the quality of specific behaviors, thoughts, and feelings. The sociogenomic model of personality traits by Roberts and Jackson (2008) in figure 1, shows the interconnections of the three main aspects of personality, namely thoughts, behavior, and feelings, as well their relationship to states and traits. Further, biological influences on both states and traits, as well as environmental influences on states are depicted. As this thesis focuses on the highest level of conscientiousness (i.e., the overall conscientiousness trait) biological factors will be ignored. For a more detailed description see Roberts and Jackson (2008) or Roberts et al. (2012).

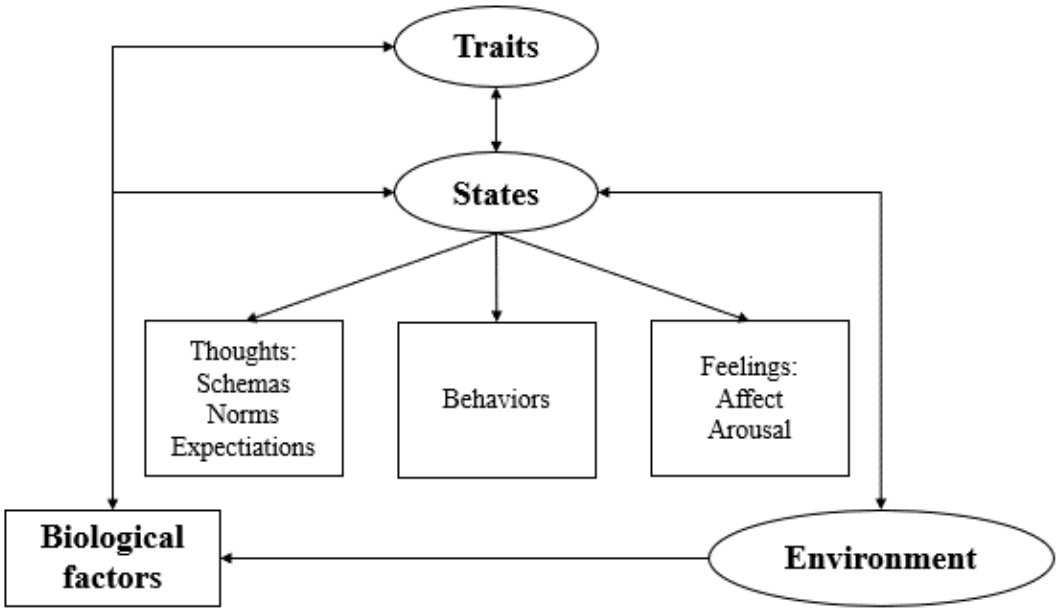

Figure 1 The sociogenomic model of personality trait by Roberts and Jackson (2008).

When looking at the lower order traits in the spectrum of conscientiousness (both state and trait), past research identified some domains of conscientiousness. Roberts et al. (2012) offer a good summary over these domains.

Domains of Conscientiousness

Roberts et al. (2012) calls the most common domains orderliness, industriousness, self-control, and responsibility. Orderliness captures neatness, cleanliness and planfulness, whereas industriousness describes the propensity for hard work, aspiration to excellence and persistency when challenged by something (DeYoung, Quilty, & Peterson, 2007; Roberts, Chernyshenko, Stark, & Goldberg, 2005; Roberts, Bogg, Walton, Chernyshenko, & Stark, 2004; Jackson, Wood, Bogg, Walton, Harms, & Roberts, 2010; Perufini & Gallucci, 1997; Saucier & Ostendorf, 1999; De Raad & Peabody, 2005; MacCann, Duckworth, & Roberts, 2009; Stanek & Ones, 2018). Self-control stands for impulse control and can therefore be described as the lack of impulsivity and recklessness (Roberts et al., 2004, 2005; Jackson et al., 2010, Perugini & Gallucci, 1997; De Raad & Peabody, 2005; MacCann et al., 2009). Responsibility, also called reliability or cautiousness, encompasses the tendency of keeping promises and following rules, that make social groups work more smoothly (Roberts et al. 2004, 2005; Jackson et al., 2010; Perugini & Gallucci, 1997; Saucier & Ostendorf, 1999; De Raad & Peabody, 2005; Stanek & Ones, 2018). Other domains, which have a little less research basis are traditionality (or conventionality; Roberts et al. 2004, 2005), decisiveness (Roberts et al., 2004; Saucier & Ostendorf, 1999), formality and punctuality (Roberts et al., 2004; Jackson et al., 2010), persistence (or grit; De Raad & Peabody, 2005; MacCann et al., 2009; Stanek & Ones, 2018), and virtue (Roberts et al., 2005). This is just a brief outline of the domains of conscientiousness. Different authors cluster those facets differently. As those facets further correlate with each other and therefore could be subsumed, not all those facets can be found in conventional personality inventories simultaneously. Roberts et al. (2005) reason, that either their study samples were biased, and some dimensions can not be found because of that or that some identified facets can be subsumed to a greater facet. Consequently, those conventional questionnaires focus on specific facets of conscientiousness (Roberts et al., 2005).

Conscientious Behavior in (online) surveys

To apply the domains of conscientiousness on (online) surveys, one would expect conscientious participants to answer the survey thoughtfully and accurately by reading the ethical and legal information (i.e., informed consent; Theiss, Hobbs, Giordano, & Brunson, 2014) and instructions carefully. They would control their impulses to do something nicer (e.g., watching funny animal videos) and answer the survey without interruption consistently and consequently. In sum, they would put effort in their response, as they know that carelessness and inattentiveness would harm the study behind the survey. Further, they understand the social convention, that the potential reward of the study (e.g., money or the aimed university degree)

is only righteously achieved, if they put effort in the survey. They trade their effort with (im-) material rewards of the study's authors. In contrast, not doing those things would indicate a low level or even absence of conscientiousness, which is often called careless responding or insufficient effort responding.

Careless Responding

The terms careless responding (Meade & Craig, 2012), insufficient effort responding (Huang, Curran, Keeney, Poposki & DeShon, 2012), content independent responding (Evans & Dinning, 1983), content-nonresponsivity (Nichols, Greene, & Schmolck, 1989), inattentive responding (Berry, Rana, Lockwood, Fletcher & Pratt, 2019), inconsistent responding (Meade & Craig, 2012) and sometimes even random responding (e.g., Credé, 2010) share one definition: participants' answers do not reflect their true attitude, or ability because they were inattentive or did not put any effort into answering (Meade & Craig, 2012; Haerzen & Chill, 1963 in Huang et al. 2012; Arias et al. 2020; Nichols & Edlund, 2020). Although all before mentioned terms share this definition, they often additionally focus on specific behavior, like (the absence of) specific response patterns. The term careless responding subsumes all above mentioned terms and is therefore used to describe this diffuse response behavior. Meade and Craig (2012) define careless responding as responding without regard to the item content. This might be the result from a lack of effort (i.e., insufficient effort responding; Huang et al., 2012) or attention (i.e., inattentive responding; see Berry et al., 2019; Arias et al. 2020), caused by environmental distractions, the survey's length, a lack of social contact (and therefore no enforcement of social norms) and boredom (i.e., lack of interest; Meade & Craig, 2012). This results in responses, which are detached from the item content (i.e., content independent responding or content non-responsivity; Evans & Dinning, 1983; Nichols, Greene, & Schmolck, 1989). Consequently, the mix of "normal" and negatively formulated survey questions lead to inconsistent responses (Meade & Craig, 2012) for careless responders. Sometimes these response patterns are described as random (e.g., Credé, 2010). Arias et al. (2020) in contrast argued, that humans can not respond completely random and always tend to show some systematic response. An extreme form of systematic response is the repeated selection of the same answer option (e.g., "I totally agree") irrespective of the item content. DeSimone, DeSimone, Harms and Wood (2018) call this behavior straightlining and report a high impact on data properties and quality.

Implications on data

Both, deliberate faking and careless responding are a huge problem for research and knowledge generation, because they highly depend on data quality. This thesis wants to shed light on careless responding in a low-stake situation. Basically, careless responding taints data by adding error variance (e.g., Meade & Craig, 2012). This additional variance can add spurious effects in the data and can also weaken true relationships, for which both alpha and beta error of the respective analyses are increased. In other words, analyzing samples including careless responders will reduce statistical power (Maniaci & Rogge, 2013). In both cases, researchers may come to wrong conclusions, fueling the replication crisis (Leiner, 2019) and compromising the validity of a field of study (Nichols & Edlund, 2020; Arias et al., 2020, Goldhammer, Annen, Stöckli & Jonas, 2020).

Detection Methods

Previous research suggested many different measures to detect careless responding. Some researchers have suggested to include special items like instructed response items (e.g., "Select 'I totally agree' in this item."; e.g., Meade & Craig, 2012; Arthur, Hagen, & George, 2020) and nonsensical/ bogus items ("I like to eat concrete."; e.g., Meade & Craig, 2012; Berry, et al., 2019; Arthur et al., 2020, Leiner, 2019). Others have suggested measuring response consistency and using indices like the variance, standard deviation, or their robust equivalent (median and median absolute deviation, MAD) between responses, as well as the within-person correlations across item pairs. Another approach is the detection of response patterns, e.g., straightlining using so called longstrings (i.e., the number of identical consecutive responses). Also, multivariate outlier analyses, like the Mahalanobis distance can be used (e.g., Meade & Craig, 2012; Arthur et al., 2020, McKay et al., 2018; Bowling, Huang, Bragg, Khazon, Liu and Blackmore, 2016; Leiner, 2019). Finally, literature also mentions response time, although there are mixed results leading to disagreement about its legitimacy as an indicator of carelessness (e.g., Meade & Craig, 2012, Arthur et al., 2020; McKay et al. 2018, Berry et al. 2019, Leiner, 2019). Several of these approaches and techniques lie within the taxonomy of paradata by Kroehne and Goldhammer (2018). For instance, the response time for answering a single survey page or even a whole survey can be assigned to process-related paradata. Response consistency, as well as response patterns like straightlining can be assigned to response-related paradata. Consequently, paradata can be used to detect careless responding.

Prevalence

The prevalence of careless responding depends on the method and its detection threshold (Meade & Craig, 2012; Nichols & Edlund, 2020). Those authors report 10-12% prevalence in their undergraduate sample. Arias et al. (2020) reports a prevalence of 5-12%, sampled with Prolific Academic (a data collection tool). Nichols and Edlund (2020) in contrast, found 25%-66% careless responders in their three studies, sampled with Amazon's Mechanical Turk (MTurk). Goldhammer, Annen, Stöckli and Jonas (2020) sampled their candidates within the Swiss Army and found 33% careless responders. Partially based on literature, Arthur, Hagen, and George (2020) assume a prevalence of 10-50% for careless responding in low-stake and consequence free settings, like in research and organizational studies.

Characteristics of Careless Responders

Regarding the personal profile of careless responders, Nichols and Edlund (2020) found men, with the educational level of a bachelor's degree or lower to put insufficient effort in responding. Other factors were age ($r = -.11$), and agreeableness ($r = -.22$), conscientiousness ($r = -.27$), emotional stability ($r = -.10$), locus of control ($r = -.25$), need for consistency ($r = -.20$), need for structure ($r = -.17$), openness ($r = -.24$), private self-conscientiousness ($r = -.14$) and self-esteem ($r = -.12$). Further, Nichols and Edlund (2020) found significant sex differences in self-reported conscientiousness ($d = 0.35$) among other variables in their overall sample. When removing persons, who answered two or more (careless responding) detecting questions incorrectly (i.e., some careless responders were still included in the analysis), this effect still existed. When excluding all potentially careless responders, who answered at least one detecting question incorrectly, this sex difference regarding conscientiousness disappeared. Although other scales showed somewhat differing effects, when completely removing potential careless responders, the overall effect is consistent even when bootstrapping is used: only by removing all potential careless responders, results of (online) surveys might be less biased. Although Nichols and Edlund (2020) emphasize that their results are only preliminary, they still offer some benchmark on the deleterious effect of careless responding.

McKay et al. (2018) reports correlations of $r = -.18$ between the HEXACO conscientiousness scale and answering instructed response items and a $r = -.02$ between the survey time (in minutes) and the respective scale. Other literature (Maniaci & Rogge, 2013) also reports $r = .01$ between survey time and the Big 5 conscientiousness. For the latter results, the true relationship between response time and conscientiousness is unclear. Ward and Pond (2015) and Bowling, Huang, Brower, and Bragg (2021) connected a shorter response time to more careless responding. Clicking through a survey as fast as possible indicates that the

participants' effort is put into the process of answering and not the survey itself. This careless response behavior automatically results in less time for accurate reading and item dependent responding (Bowling et al., 2021). On the other hand, McKay et al. (2018) reported a big discrepancy between the arithmetic mean of the survey's response time (81.63 minutes) and the median (39.86) with a standard deviation of SD = 360.13, indicating a very big variance and influences of heavy outliers. The domain of conscientiousness subsumes impulse and self-control, as well as acting firmly and consequently. Thus, longer response times may be reached by breaks and distractions. As breaks and embraced distractions are contrary to conscientiousness' impulse control and consistent acting, one could argue that a far above average response time also indicates a low conscientiousness. Putting both together, one expects curvilinear inverted U-shape relationship between conscientiousness and response time (Maniaci & Rogge, 2012). This implies that both very short and very long response times can be seen on participants with a low conscientiousness. Furthermore, conscientious persons would be expected to show average response times reflecting the time needed to respond adequately on the items. But while very short and very long response times can be a result of e.g., straightlining or breaks, average response times can be a result of both and are therefore not clearly assignable to any conscientiousness level (McKay et al., 2018). Consequently, average response times do not prove careful or thoughtful responding. Irrespective of the average responders, we expect more low conscientious persons on the lower and higher tail of the response time to cause this curvilinear inverted U-shaped effect over all data points. This non-linear relationship would lead to weak or even non-existing Pearson's or Spearman's correlations as they test only linear and monotonous relationships. These assumptions are strengthened by the sometimes found approximately complete absence of a correlation between response time and conscientiousness (McKay et al., 2018, Maniaci & Rogge, 2012). Ultimately, as the inverted U-shape relationship is expected to be a result of careless responding, this curvilinear effect should vanish when excluding careless responders from the analysis.

Hypotheses

In summary, paradata potentially offer an economic approach to measure behavior like careless responding, which is ultimately related to one's personality. Careless responding indicates solely by definition and theory a lack of or the absence of conscientiousness. It expresses itself in insufficient reading accuracy (for e.g., the informed consent; Theiss et al., 2014, Bowling et al., 2021), content independent response patterns (e.g., straighlining; DeSimone et al., 2018)

resulting in extremely fast responses, or breaks and distractions which lead to very long responses, Bowling et al. (2016), McKay et al. (2018) and Berry et al. (2019) further support the theoretical implication, that the lack of conscientiousness partly causes careless responding. In return, measures and detection approaches for careless responding which partially lie within the paradata framework (Kroehne & Goldhammer, 2018) can reflect a person's conscientiousness. One controversial measure for careless responding is a survey's response time. McKay et al. (2018), as well as Maniaci and Rogge (2012) report near zero correlations between response time and conscientiousness. While McKay et al. (2018) argues that there is no relationship between response time and careless responding, which in return effects its relationship with conscientiousness, Maniaci and Rogge (2012) suggest a nonlinear relationship as a reason for the undetectable (linear) correlation. To shed light onto this matter, this thesis examines if a curvilinear inverted U-shaped relationship exists between conscientiousness measures and response time. Also, Ranger and Ortner (2011) examined this potential inverted U-shaped relationship. In contrast to this thesis, they argued that more extreme (both high and low) personality trait levels lead to faster responses because items are either easily identified as well-fitting or completely misfitting. Consequently, individuals have a high response probability on agreeing well-fitting items and a high rejection probability in bad fitting items. In sum, a persons' response probability (either endorsing or rejecting an item) is related to their response time. Ranger and Ortner's (2011) analyses showed that the response probability is negatively associated with response time. They examined this relationship in a military sample in which unusually fast responders were excluded as they were assumed to be not motivated enough. In contrast to their study design, where honesty and motivation were ensured, this thesis aims to examine the relationship between conscientiousness and response time under the scope of a more natural test situation, where participants might be unmotivated or carelessly responding even though they voluntarily participate and potentially get rewarded. Consequently, the diagnosis of a curvilinear inverted U-shape relationship between response time and conscientiousness will be controlled for careless responders by using two samples. One sample including careless responders and one clean sample without careless responders.

H1.1: There is a curvilinear inverted U-shaped relationship between self-reported conscientiousness and the averaged response time per page

H1.2: There is a curvilinear inverted U-shaped relationship between the number of correct retrieval questions about the informed consent and the averaged response time per page

We would further expect conscientious persons to work persistently and control their impulses to do more pleasant things (e.g., “offline” coffee break, searching for cat videos on YouTube, visiting social media platforms, etc.) instead or while responding to the online survey (e.g., Jackson et al., 2012). As the used online survey comprised multiple pages, the variability in response times per page can be calculated as a measure for consistent behavior. A higher intra-person variability in response time per page indicates an inconsistent response mode, which is more likely occurring within less conscientious individuals. Therefore, we expect a higher variability in response time per page to be negatively associated with our conscientiousness measures.

H 2.1.1 Variability in response time per page correlates significantly negative with self-report conscientiousness

H 2.1.2 Variability in response time per page correlates significantly negative with the number of correct retrieval questions about the informed consent

Further, as conscientious persons tend to work persistently accurate over time it is expected, that conscientious persons will also answer survey items accurately. Examining mouse click or touch accuracy, the paradata indicator ‘click inaccuracy’ is used to explore the relationship between click-inaccuracy and conscientiousness measures exploratory. Conscientious persons are expected to have a higher accuracy, which therefore shows itself in a lower click-inaccuracy than not-conscientious persons.

H 2.2.1 Click-inaccuracy of a person shows a significant negative correlation with self-reported conscientiousness

H 2.2.Click-inaccuracy of a person shows a significant negative correlation with the number of correct retrieval questions about the informed consent

Again, we also expect conscientious persons to show a persistent behavior and therefore the variability of click-inaccuracy across the survey pages is expected to be negatively associated with conscientiousness.

H 2.3.1 The variability in click-inaccuracy of a person shows a significant negative correlation with self-reported conscientiousness

H 2.3.2 The variability in click-inaccuracy of a person shows a significant negative correlation with the number of correct retrieval questions about the informed consent

Finally, the above-mentioned predictors are entered into one hierarchical regression model to examine their unique explained variance in conscientiousness measures. In a first hierarchical step, all predictors get included. In the second step the global response time will be squared if a U-shaped effect was detected in H1. Although using a quadratic regression approach is unsuitable for detecting U-shaped effects (see the respective section in Methods), it can be used to examine the change in model fit.

H 3.1 Response time per page and its variability, as well as averaged click-inaccuracy and its variability are significant unique predictors for self-reported conscientiousness

H 3.2 Response time per page and its variability, as well as averaged click-inaccuracy and its variability are significant unique predictors for the number of correct retrieval questions about the informed consent

To sum it up: In order to find an economic alternative for OPTs this thesis examines paradata, which are basically behavioral observations in computerized tests. Some types of paradata are already used to detect careless responding, which is empirically and theoretically associated with the personality dimension conscientiousness. Consequently, we expect that paradata indicators for careless responding like the response time and behavioral consistency measures are also indicators for conscientiousness.

Methods

Sample

In the Winter Semester 2019/2020 this sample was conducted as part of an empirical seminar for undergraduate Psychology students at the University of Salzburg. The seminar participants created this online survey with the program Limesurvey and advertised it in their circle of acquaintances. As a reward for completing the online survey, Psychology students at the University of Salzburg were promised 0.5 curricular credits. Additional to that, one 20€ Amazon gift card was randomly given away among 50 survey participants. A total of 194 persons completed the online survey. After data preprocessing and the outlier analysis, two samples were used for the analyses in order to control for careless responding.

Measures

Demographic Variables

The participants were asked about their sex, age, and mother tongue. Additional to that, they had to indicate if they study at a university and if so, which subject they study. Furthermore, the participants had to report, which device (e.g., Notebook, Smartphone, etc.) was used for answering the survey.

IPIP-240

To assess the self-reported personality trait conscientiousness, the IPIP-240 (Schreiber & Iller, 2016) likewise named scale was used. IPIP means International Personality Item Pool and the respective IPIP-240 is a free personality questionnaire in German language with 240 items on a 5-Point Likert-Scale. It is based on the Big-5 factor model by McCrae and Costa (1985) and assess the same six facets of each dimension like the NEO-PI-R (Costa & McCrae, 1992). For conscientiousness the facets are self-efficacy, orderliness, dutifulness, achievement striving, self-discipline and cautiousness. In contrast to the original questionnaire IPIP-300 with 300 items in English language, the German version has fewer items, and its length is therefore equal to the NEO-PI-R. As the mentioned original scale is in English language, the German translation was used for this survey. Treiber (2013) did the validation of the IPIP-240 with the German version of the NEO-PI-R (Ostendorf & Angleitner, 2004). The validation study showed satisfactory results.

Filler Questionnaire about Social Desirability

The filler questionnaire consisted of questions about general socially desirable behaviors and skills which make one suitable for studying at a university. Exemplary statements were “I’m able to present topics in a creative way.” and “I can cope with bad grades.”. Overall, 54 such statements were presented across six survey pages. On each page three item blocks with each three statements had to be rated on a 5-point Likert-Scale with (1) “Not desired at all” to (5) “Very desired” (see figure 4). Participants were not able to progress the survey unless all nine statements were rated. Ultimately it was irrelevant how the statements were rated as the only goal was to collect data about participants’ response behavior using a Java Script.

Operationalization

Paradata Indicators

A Java Script of Scherndl (2019) collected paradata for each survey page. Each of the six relevant survey pages looked like in figure 4. Users had to rate each statement on a 5-point Likert-Scale. They could rate a statement by either selecting the small inner circle of the box itself or by clicking on the rectangular box around the smaller circle. Every action of a person got logged by the paradata Java Script. More specific, the Java Script logged the ID of the respective block of statements (Question ID), which belonged to the survey program Limesurvey. Followed by the timestamp of the action (i.e., selecting an answer) as well as the answer with its statement number (0-2) starting from zero and the chosen rating (1-5). Figure 2 shows the raw paradata string of one page. The timestamps show the passed seconds since a specific date, which is irrelevant for further calculations, as we use the timestamps per survey page as reference.

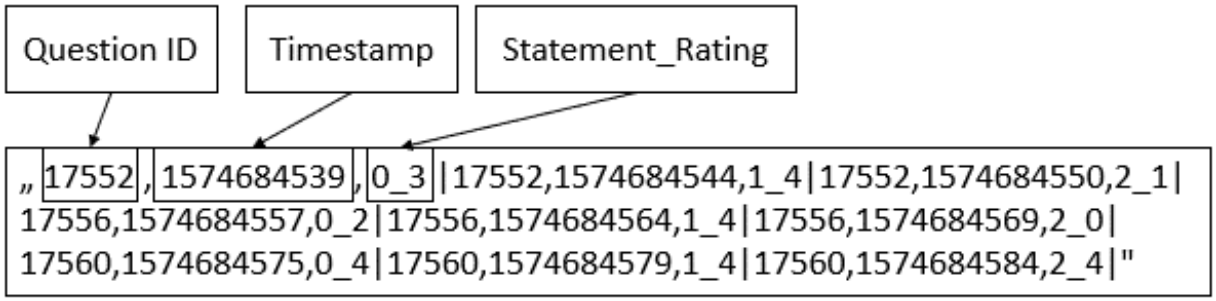

Figure 2 An example of the raw paradata character string for one of the six survey pages of the filler questionnaire like in figure 4. All events are separated by a vertical line "|". Each event comprises the question ID, the timestamp and the statement with its Likert rating separated with an underscore "_". Each question consisted of three statements, which had to be rated. The statements are zero-index (starting with 0) while the rating ranged from 1 to 5.

Response time per page

As noted above, several timestamps for each of the six pages were available. Although it seemed reasonable to subtract the first timestamp from the last timestamp to get the passed time between the first and the last response, it creates an information gap. A participant still can either make a pause before responding to any item or making a pause between the last response and pressing the "next page" button to progress the survey. Therefore, the response time per Limesurvey question group from the survey program itself was used as the time per page indicator. Limesurvey's response time per page starts when the survey page finished loading and ends with the click on the next page button. Consequently, it offers additional information to the

paradata strings created with Java Script as the time before answering the first item and after answering the last item also is measured.

Participants had to answer six survey pages and the response time of each survey page was tracked. Overall, response times were extremely skewed reaching from a minimum of 9 seconds to a maximum of 54 679 seconds (is equal to over 15 hours). Due to this skewness and the influence of extreme outliers, the median response time over all six survey pages was used as a person's individual response time. Additionally, as a measure for a person's consistency in response time behavior the median absolute deviation (MAD) was calculated. To address the still skewed response times (natural) logarithmic (log-) transformation was applied (Ranger & Ortner, 2011; Ranger, 2013; Höhne & Schlosser, 2018). As log-transformed values are unintuitive to interpret, the average and variability in response times are z-standardized. Both the average response time as well as the variability in response time can be classified as macro process-related paradata (Kroehne & Goldhammer, 2018).

Average (Avg) response time *:= the z-standardized and log-transformed median of all 6 response times of each person*

Variability in response time *:= the z-standardized and log-transformed median absolute deviation (MAD) of all 6 response times of each person*

Click-Inaccuracy

Click-inaccuracy indicates, where a person clicks on the Likert-Scale to respond to an item. As mentioned earlier, participants can either click on the smaller circle, where the answer dot is visible afterwards, or click on the rectangular box surrounding this smaller circle (see figure 4). Limesurvey has the feature (or maybe bug), that clicks on the rectangular surrounding box is logged two times in the paradata string. An explanation would be, that the click on the box is one event which the paradata script logs. But as the answer dot has to appear within the inner circle, a second click automatically is done by Limesurvey. This in return also gets logged as an event in the paradata string, resulting in two exactly equivalent paradata events. Equivalence means, that question ID, timestamp, and answer are identical in both events. Therefore, we can count identical events per page to create an indicator of click-inaccuracy.

Like in the section for the response time per page, click-inaccuracy also gets averaged over the six pages. The more conservative approach using median, and MAD is also used here. In contrast to the response time, variables were not logarithmically transformed. One reason is, that the distribution is less skewed, and variance is not that extreme high. Another reason is the

missing literature. And although the maximum to minimum ratio is also over 10 (Sabin & Stafford, 1990) it does not seem to be that necessary like with response time. Further, as click-inaccuracy is not log-transformed and therefore interpretable, it was not standardized. Based on Kroehne and Goldhammer's (2018) taxonomy, both indicators can be classified as answer response-related paradata.

Average (Avg) click-inaccuracy *:= the median in clicks, which were administered in the rectangular box surrounding the inner circle*

Variability in click-inaccuracy *:= the MAD in clicks, which were administered in the rectangular box surrounding the inner circle*

Conscientiousness measures

Self-reported conscientiousness

The IPIP-240 (Schreiber & Iller, 2016) was used to assess the self-reported Big-5 trait dimension conscientiousness. The self-reported trait conscientiousness measure comprised the six facets of conscientiousness: self-efficacy, orderliness, dutifulness, achievement striving, self-discipline and cautiousness. The items for each facet, as well as all facets together were aggregated to an overall self-report trait conscientiousness score like instructed in the manual (Schreiber & Iller, 2016). For simplicity only the overall conscientiousness score was used as a dependent variable, as it covers a broad spectrum of the domain of conscientiousness with its facets.

Self-reported conscientiousness *:= mean of all six IPIP-240 facet test scores*

*Bitte bewerten Sie die folgenden Aussagen auf einer Skala von 1 – 5, wie sozial erwünscht diese für die Studieneignung sind.

	gar nicht erwünscht 1	2	neutral 3	4	sehr erwünscht 5
Ich weiß, wie ich durch mein Verhalten im Unterricht, meine Mitschüler*innen/Mitstudierenden motivieren kann.	○	○	○	○	○
Versäumte Informationen gebe ich an nicht anwesende Mitschüler*innen/Mitstudierende weiter.	○	○	○	○	○
Ich erweitere mein Netzwerk mit Schüler*innen von anderen Schulen/Studierenden aus anderen Studienrichtungen.	○	○	○	○	○

*Bitte bewerten Sie die folgenden Aussagen auf einer Skala von 1 – 5, wie sozial erwünscht diese für die Studieneignung sind.

	gar nicht erwünscht 1	2	neutral 3	4	sehr erwünscht 5
Mir wichtige Anliegen bringe ich in Diskussionen im Unterricht/in Lehrveranstaltungen ein.	○	○	○	○	○
Auch wenn ich mit anderen als gewünscht in einer Gruppe bin, kann ich mit diesen gut zusammenarbeiten.	○	○	○	○	○
Ich engagiere mich in einer Vertretung von Schüler*innen/Studierenden oder Ähnlichem.	○	○	○	○	○

*Bitte bewerten Sie die folgenden Aussagen auf einer Skala von 1 – 5, wie sozial erwünscht diese für die Studieneignung sind.

	gar nicht erwünscht 1	2	neutral 3	4	sehr erwünscht 5
Damit ich ein Thema umfassend begreife, stelle ich weiterführende Fragen.	○	○	○	○	○
Um ein Lernziel zu erreichen, plane ich die nötigen Schritte im Voraus.	○	○	○	○	○
Wenn ich unverschuldet zu spät in den Unterricht/eine Vorlesung komme, beschäftigt mich das nicht lange.	○	○	○	○	○

Figure 4 One exemplary survey page for the filler questionnaire, where paradata were tracked

"Objective" Conscientiousness: Informed consent (IC) retrieval score

Conscientious persons are by theory more careful, reliable, and work more accurately which express itself by double-checking work, reading instructions, and proofreading writing (Jackson et al., 2010, Bowling et al., 2021). Theiss et al. (2014) therefore concluded that conscientious persons read the informed consent (IC) more carefully and thoroughly, which they also showed in their paper. To have another measure for conscientiousness beside self-report, the percentage of correctly remembered information about the informed consent is used as a behavioral observation of conscientiousness. By masking this measure for conscientiousness as a performance task, it resembles an Objective Personality Test (OPT). In this survey, participants had to answer five questions about information which were presented in the informed consent. Those questions comprised the minimum age to participate in the survey, what and how many rewards are available for finishing the survey and who the contact persons for additional questions were.

***"Objective" Conscientiousness**:= percentage of correctly retrieved answers about the informed consent*

Indicator of Careless Responding: Failed Attention Checks (FAC)

As proposed in the literature (e.g., Meade & Craig, 2012; Arthur, Hagen, & George, 2020), instructed response items were used as an empirically established indicator for careless responding. The first attention check was offered directly after the experimental instruction for the filler questionnaire (B) in figure 5. The participants were instructed to respond with "no answer", when they were asked, if they understood the instructions. "Yes" was the predefined answer. Consequently, all participants who failed to actively select "no answer" failed the first attention check and their score for careless responding was increased by one. The second attention check was presented after the filler questionnaire (B). They were asked how they were instructed to answer the filler questionnaire (B). If participants answered "I don't know" or failed to choose the correct instruction, their careless responding score once again increased by 1. Summarized, the careless responding score is the sum of failed attention checks.

***Careless responding score** := sum of failed attention checks (FAC score)*

Experimental Design and Procedure

An online survey in German language, was conducted for acquiring the data. The survey consisted of several parts of which only some were used in this thesis (bold face text, see figure 5). The informed consent (IC) contained a welcome text and organizational, legal, and ethical information for the participants, which were accessible by clicking on a "show more" button under the welcome text. Throughout the whole survey paradata were collected. The paradata relevant to the research question (i.e., response times and click-inaccuracy) were collected on survey pages one to six which presented the filler questionnaire (A). The following motivational questions, as well as the experimentally manipulated filler questionnaire (B) were irrelevant for this thesis. In order to check if the instructions for filler questionnaire (B) were read carefully, the first instructed response item was presented (Attention Check 1). Filler questionnaire (B) was followed by the second Attention Check (2) where participants had to retrieve the earlier presented (experimental) instruction. After that, the IPIP-240 (Schreiber & Iller, 2016) questionnaire subscale for conscientiousness, which assesses the self-reported likewise named Big-5 personality trait was presented. Next, demographic and access-related information (e.g., used device) were asked. On the next page the participants had to answer five questions about

the informed consent, which was presented at the beginning of the survey. The last page again was irrelevant for this survey.

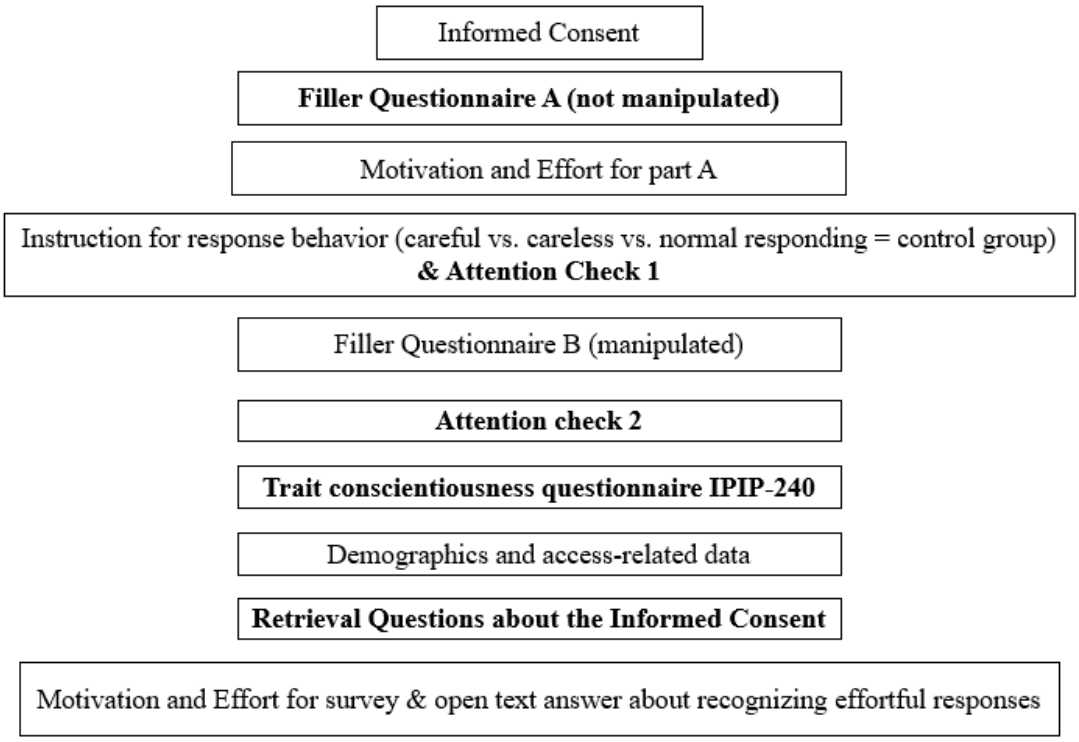

Figure 5 The structure of the survey. Parts written in bold face were used and therefore important for this thesis.

Using Interrupted Regression to diagnose U-shaped relationships

In this thesis a so called curvilinear inverted U-shaped effect is hypothesized in the relationship between the response time per survey page and the conscientiousness measures. Based on Simonsohn (2018), a U-shaped relationship comprises a sign flip from the left to the right side of a specific breakpoint. So, like the letter U, the first half of the data implicate a significant negative relationship, while the second half of the data after the breakpoint implicate a significant positive relationship. It is vice versa for the inverted U-shape relationship. Originally this type of relationship was tested with a quadratic regression. In addition to the "normal" linear predictor, an interaction term of the predictor with itself (i.e., a squared predictor) is included in the regression model. Consequently, the fit of a quadratic parabola is tested with this model. The main problem is that the quadratic regression analysis is simultaneously oversensitive and undersensitive in the detection of (inverted) U-shaped effects. Quadratic regression diagnoses an exponential or logarithmic relationship as quadratic as the curvature within the data has a better fit with a quadratic predictor than with a linear predictor alone (Simonsohn, 2018). Both exponential and logarithmic relationships are far from being a U-

shape, which makes the quadratic regression oversensitive to detect a U-shape. In contrast, if the data implicate a very wide parabola as a result of an even degree polynomial (i.e., with an even exponent like x^4 or x^8 in the leading term), quadratic regression is not sensitive enough to detect this truly U-shaped effect. Those parabolas are too wide to fit in a quadratic term and are therefore not significant, even if all parabolas have a U-shape.

With regards to those problems of the quadratic regression analysis, Simonsohn (2018) proposes another technique to diagnose (inverted) U-shape relationships. Based on Simonsohn's (2018) definition of U-shape, he proposes to test it by using an interrupted regression. Basically, a breakpoint (e.g., the tip of the U) gets calculated. If a linear regression on the left and the right side of the breakpoint are significant and have different signs, then a (inverted) U-shape is detected. To have a higher statistical power Simonsohn (2018) proposes a specific algorithm to calculate the breakpoint. For a little deeper and technical understanding see appendix BI and for the detailed original paper see Simonsohn (2018).

To sum it up: The Two-Lines analysis is a more robust measure for U-shaped effects than quadratic regression. It does not try to model the true relationship between two variables, but it tests the assumed sign flip in U-shaped effects. Simonsohn (2018) proposes an algorithm he calls "Robin Hood" to find the optimal breakpoint for the analysis, which has the highest statistical power compared to other approaches (e.g., quadratic regression). As the Two-Lines test is based on two linear regressions, the assumptions are equal to the "normal" linear regression. Consequently, heteroscedasticity also increases the false-positive rate in the Two-Lines analysis (Simonsohn, 2017) and therefore robust standard errors should be used if heteroscedasticity is found. One limitation neither quadratic regression, nor the Two-Lines test overcome, is the question about modelling the true relationship between two variables.

Outlier Analysis & Data Preprocessing

The exclusion of outliers was conducted after log-transforming the response time variables and before standardizing them, to prevent the outliers from influencing the standardization process. The whole data preparation pipeline is shown in figure 6.Outlier analysis in general is important as outliers have high statistical leverage and can therefore bias analyses. Although one could argue that outliers in response time are of special interest as we expect extremely fast and extremely slow responders to have low conscientiousness, they still bias analyses leading to unstable and potentially not replicable results. Outlier behavior may in fact be a result of low conscientiousness, but also can be the result of randomness or interference. A person answering

the online survey at home might get interrupted by a ringing telephone or when the doorbell rings. In order to overcome such random outliers and this error variance averages out, a big sample size is needed. As the sample used for this thesis is not big enough, a literature based conservative outlier analysis is conducted in order to get robust and replicable results from the used analyses. Nevertheless, outliers might be an interesting subgroup and should be studied for thoroughly in future research.

Although literature shows different approaches to deal with outliers especially in response times, we use the non-symmetric outlier analysis by Hoaglin, Mosteller and Tukey (2000; in Höhne & Schlosser, 2018). In our case, 3 times the quartile range below and above the median will be used as the lower and upper threshold to determine outliers. For a more thorough discussion of different outlier analyses see appendix BII.

Ultimately, the data get filtered based on Hoaglin et al.'s (2000 in Höhn, 2018) approach regarding self-report conscientiousness, log-transformed median and MAD time per page as well as median and MAD click-inaccuracy. For the sake of simplicity, the Hoaglin filter was also applied on non-response time variables. Due to severe skewness the IC retrieval score as a behavioral indicator for conscientiousness was not included in this outlier analysis, as it would have led to a large amount (N = 38) of exclusion solely by this single variable.

Ultimately, two data sets were used for the analyses in order to control for the influence of careless responders. Regarding the bigger data set (N = 154) only outliers (Hoaglin et al., 2000 in Höhn, 2018) were excluded. Of those 40 identified and excluded outliers 21 did not fail any attention check. Consequently, 47.50% of the excluded outliers were potential careless responders. Next to the outliers, persons who failed any attention check (FAC score ≥ 1) were excluded in sample 2 (N = 87). Consequently, the smaller sample is believed to be free of careless responders. See appendix A for more information about the characteristics of the excluded cases.

Results

All data preparation and analyses were conducted with the statistical software program R (v4.1.1; R Core Team, 2021). The R package *pacman* (v.0.5.1; Rinker, Kurkiewicz, Highitt, Wang, Aden-Buie, & Burk, 2019) was used as a package manager to install and load other packages. The packages *data.table* (v1.14.0; Dowle, Srinivasan, Gorecki, Chirico, Stetsenko, Short,… Schwen, 2021), *stringr* (v1.4.0; Wickham, RStudio, 2019), *stringi* (v1.7.3; Gagolewski, Tartanus, and others, 2021), *tidyr* (v.1.1.3; Wickham, RStudio, 2021) and *dplyr*

(v1.0.7; Wickham, Francois, Henry, Müller, RStudio, 2021) were used for data preparation and manipulation. The analyses and plots were conducted with the following packages: *apaTables* (v2.0.8; Stanley, 2021), *car* (v.3.0-11; Fox, Weisberg, Price, Adler, Bates, Baud-Bovy,…R Core Team, 2021), *corrplot* (v0.90; Wei, Simko, Levy, Xie, Jin, Zemla,…Protovinsky, 2021), *ggplot2* (v3.3.5; Wickham, Chang, Herny, Pedersen, Takahashi, Wilke,…RStudio, 2021), *gridExtra* (v2.3; Auguie & Antonov, 2017), *jtools* (v2.1.3; Long, 2021), *lmtest* (v0.9-38; Hothorn, Zeileis, Farerbother, Cummins, Millo, &Mitchell, 2020), *mgcv* (v1.8-36; Wood, 2021), *nlme* (v3.1-152; Pinheiro, Bates, DebRoy, Sarkar, EISPACK authors, Heisterkamp,…R Core Team, 2021), *PerformanceAnalytics* (v2.0.4; Peterson, Carl, Boudt, Bennett, Ulrich, Zivot,…Shea, 2020), *psych* (v2.1.6; Revelle, 2021) and *sandwich* (3.0-1; Zeileis, Lumley, Graham, & Koell, 2021).

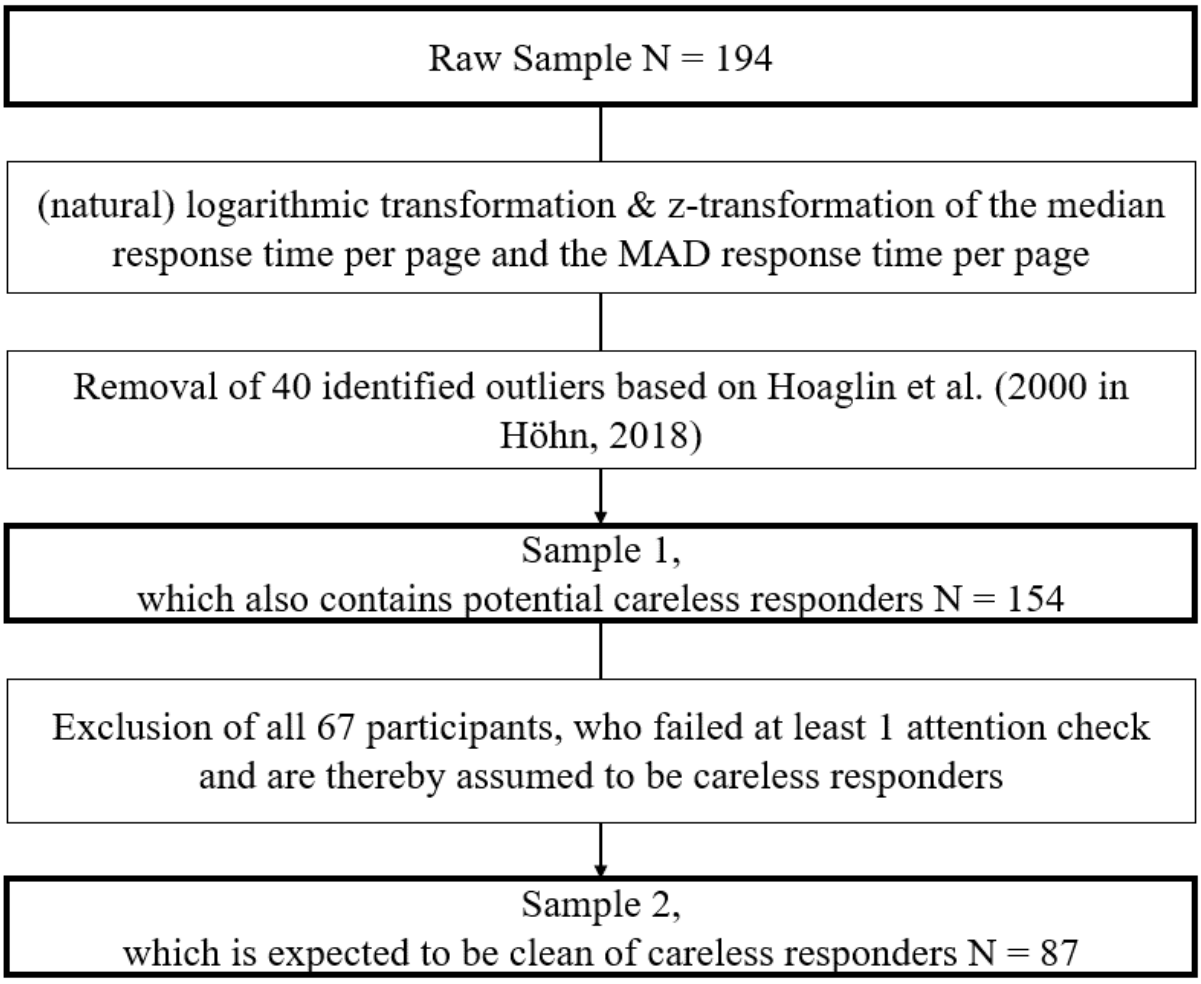

Figure 6 Visualizing the data processing pipeline, comprising the data preparation and the exclusion of outliers and assumed careless responders.

Sociodemographic Descriptives

The bigger data set (N = 154) solely excluded outliers. Consequently, this data set included the 18 persons who failed both attention checks and the 49 persons who failed one attention check. As attention checks are an indicator of careless responding, these 67 (43.50%) participants who failed at least one attention check (FAC ≥ 1) are assumed to be careless responders. Ultimately, this data set contained 117 female (75.97%) and 37 male (24.03%) participants. Their age ranged from 16 years to 67 years with a median of 22 years (1st quartile = 20, 3rd quartile = 23.25). Most of them reported to study (N = 142), whereas two people refused or failed to answer. Based on paradata about the operating system (e.g., Android, iOS, Windows, etc.), 45 participants (29.22%) used a device with presumably a touchscreen (i.e., Android & iOS users).

In contrast to the first data set which only excluded outliers, the second, clean data set (N = 87) additionally excluded persons who failed at least one attention check (N = 67). This data set is therefore assumed to be free of careless responders. The clean sample comprised 25 men (28.74%) and 62 women (71.26%). Their median age is 21 years (1st quartile = 20, 3rd quartile = 23). 89.66% of those reported to study, while 2 persons failed or refused to answer this question. Comparable to the bigger sample, 26 persons (29.89%) presumably used a device with a touchscreen.

Table 1 An overview over the number of potential careless responders (CR, number of failed attention checks FAC ≥ 1) in different samples and the item difficulties of both attention checks based on the classical test theory.

	N	Percentage of participants with FAC ≥ 1	Percentage of participants with FAC = 2	Difficulty Attention Check 1	Difficulty Attention Check 2
Raw Sample	194	44.33%	11.34%	.79	.88
Outliers	40	47.50%	10.00%	.78	.85
Sample 1: no outliers	154	43.50%	11.69%	.79	.89
Sample 2: no outliers, no CR	87	0%	-	-	-

Self-report Conscientiousness: IPIP-240

The global score of the IPIP-240 was used as a self-report indicator for trait conscientiousness on a 5-Point-Likert-Scale. Its internal consistency with α = .85 in sample 1 (N = 154; see table 2) and α = .84 in sample 2 (N = 87) is a little bit smaller than the reported reliability in the IPIP-240 manual (Schreiber & Iller, 2016). Participants rated their trait conscientiousness with a mean of 3.05 (SD = 0.35) in sample 1 (N = 154) and 3.01 (SD = 0.33) in sample 2 (N = 87), while are approximately normally distributed. Nevertheless, the full range of the 5-Point Likert-Scale is not used, as no observations below 1 and substantially above 4 exist.

"Objective" Conscientiousness: IC retrieval questions

Based on Theiss et al. (2014), conscientious persons were expected to read the informed consent (IC) more thoroughly than less conscientious persons. Further, Bowling et al. (2021) found a higher reading accuracy in conscientious persons. By using five retrieval questions about the IC, we masked this behavioral indicator for conscientiousness as a performance task similar to OPTs. As can be seen in table 2, the internal consistency of the "objective" conscientiousness measure is low with $\alpha = .53$ for sample 1 (N = 154) and $\alpha = .38$ for sample 2 (N = 87), even when accounting for the low number of items and that heterogeneous retrieval questions usually have lower internal consistencies than e.g., homogeneous reasoning items. Additional psychometric analyses showed that only one of the five items had a moderate difficulty ($p = .66$ for N = 154; $p = .72$ for N = 87) based on classical test theory. Another item was barely solved ($p = .03$ for N = 154; $p = .02$ for N = 87) and the remaining three items were easy ($.79 \leq p \leq .85$ for N = 154; $p = .80 – p = .91$ for N = 87). This led to a negatively skewed (-0.92 for sample 1 and -0.93 for sample 2) distribution of the percentage of correct IC retrieval items. On average, sample 1 (N = 154) showed a mean of 62% and sample 2 (N = 87) a mean of 66% correctly solved IC retrieval items with a respective standard deviation of 23% and 18%.

Paradata Indicators

The average and variability in response times per page, as well as the average and variability in click-inaccuracy are paradata indicators, which are assumed to be related to the personality trait conscientious. Cronbach's α as a measure for internal consistency was calculated with the data on which the aggregated median and MAD score is based on. Consequently, α is identical for the average and the variability in response time per page as well for the average and the variability in click-inaccuracy. Therefore, α is solely reported for the average scores in table 2.

Response times in seconds were recorded over 6 survey pages. When looking at the raw data (N = 194) of the 6 survey pages, where no exclusion criteria or log-transformation were applied on, extreme low (min = 9.02 sec) as well as extreme long response times (max = 15.19 hours) could be found. After calculating the median response time per page over the 6 survey pages and performing the outlier analysis, it took participants (N = 154) on average M = 65.99 (SD = 16.22) seconds to answer all items on a survey page. Examining solely persons who passed both attention checks (N = 87) and therefore excluding assumed careless responders, one finds a raw (i.e., not yet log-transformed) average response time per page of M = 67.26 (SD = 16.89) seconds. The variability of the raw response time per page, as well as both log-

transformed average and variability in response time per page can be found in table 2. When looking at the internal consistency of the response times for all 6 survey pages, raw response times (i.e., no log-transformation) in sample 1 (N = 154) with assumed careless responders lead to a very low α = .13. In contrast to sample 2 (N = 87) where α = .74 is substantially higher, despite also not being log-transformed. After log-transforming the 6 response times for each participant, the internal consistency in the careless responder sample (N = 154) is with α = .68 still lower than in the clean sample 2 (N = 87) with α =.84. But just by the logarithmic transformation Cronbach's α was increased.

The average click-inaccuracy as well as the variability in click-inaccuracy showed a high internal consistency with α = .95. On average in sample 1 (N = 154), the participants clicked M = 4.40 (SD = 3.60) times in the surrounding rectangular box instead of the center. When excluding persons who failed at least one attention check (sample 2, N = 87), the remaining participants clicked on average M = 4.67 (SD = 3.44) times in the surrounding rectangular box instead of the center. The variability in click-inaccuracy is depicted in table 2.

Additional to those paradata indicators, the number of failed attention checks (FAC score) was used to detect and filter out careless responders. While half of the participants failed no attention check at all (FAC = 0), the mean score of failed attention checks was M = 0.55 (SD = 0.70) in sample 1 (N = 154). The internal consistency of the measure was low with α = .36.

Correlation between conscientiousness and paradata indicators (H2)

For the H2 hypotheses, different linear relationships (correlations) were expected. A high average clicking inaccuracy, as well as a high variability in clicking inaccuracy and variability in response time were hypothesized to be negatively associated with both measures of consciousness. The respective correlation tables can be found in the appendix B. In table B1, assumed careless responders were included in the analysis (N = 154). Whereas in table B2, the correlations were calculated without careless responders (N = 87).

Unexpectedly, there was a significant negative correlation ($r(152) = -.21$, $p < .01$) between the self-reported conscientiousness and the log-transformed average response time per page. The effect got stronger, when careless responders were excluded ($r(85) = -.33$, $p < .01$). This relationship was hypothesized to be inverted U-shaped. Additional analyses regarding the linear relationship showed homoscedasticity for both samples. Also, the red dashed local regression (with loess smoothing) in figure 7 strongly indicates a rectilinear relationship.

Table 2 Descriptive statistics stratified for assumed careless responders (CR). The bigger sample 1 with potential careless responders consisted of N = 154 participants and the clean sample 2 without assumed careless responders (FAC ≥ 1) comprised N = 87 participants.

Variable category	Variables	number of items	Mean (Median) With CR	Mean (Median) Without CR	SD (MAD) With CR	SD (MAD) Without CR	α With CR	α Without CR
Conscientiousness measures	self-report conscientiousness	48	3.05 (3.05)	3.01 (3.00)	0.35 (0.36)	0.33 (0.37)	.85	.84
	"objective" conscientiousness (percentage of correctly solved items about the IC)	5	0.62 (0.60)	0.66 (0.80)	0.23 (0.30)	0.18 (0.00)	.53	.38
Careless responding	Failed Attention Check (FAC) score	2	0.55 (0.00)	-	0.70 (0.00)	-	.36 (2)	-
Response time variables	average response time on page (in sec)	6	65.99 (65.95)	67.26 (66.50)	16.22 (16.91)	16.89 (16.61)	.13	.74
	log-transformed average response time on page	6	4.16 (4.19)	4.18 (4.20)	0.25 (0.26)	0.26 (0.23)	.68	.84
	variability in response time on page (in sec)		15.57 (12.42)	16.94 (14.41)	9.79 (6.87)	10.27 (7.55)		
	log-transformed variability in response time on page		4.16 (4.19)	4.18 (4.20)	0.25 (0.26)	0.26 (0.23)		
Clicking Inaccuracy Variables	average click-inaccuracy	6	4.40 (4.75)	4.67 (5.00)	3.60 (5.56)	3.44 (5.19)	.95	.95
	variability in click-inaccuracy		1.29 (1.48)	1.35 (1.48)	1.03 (1.10)	0.99 (1.10)		

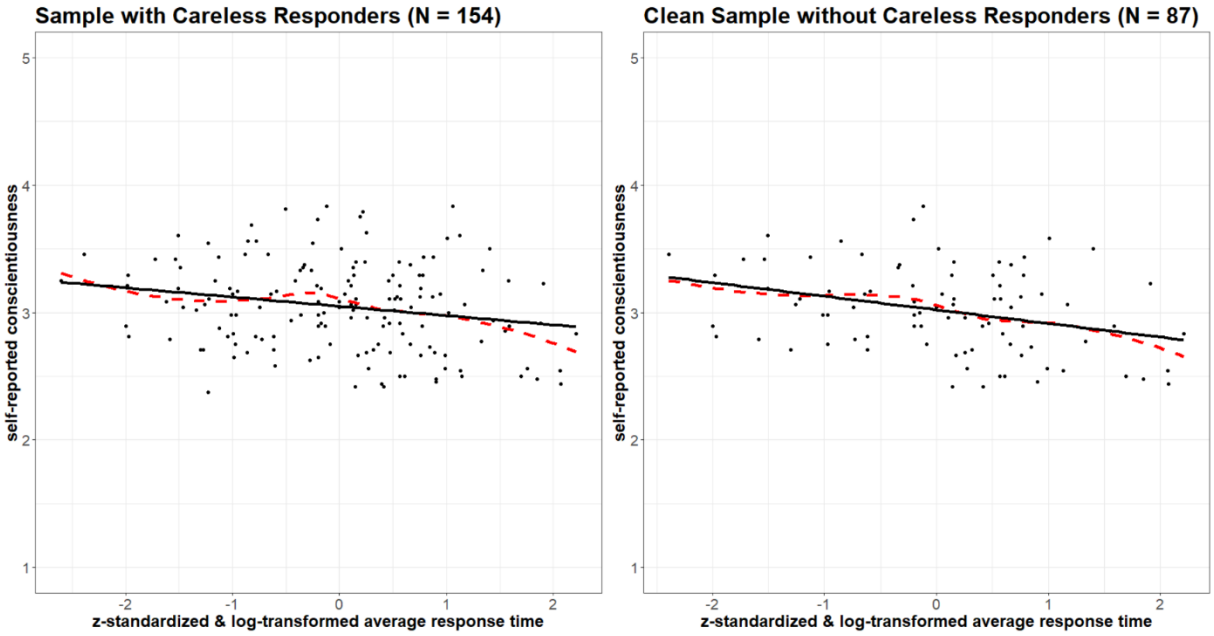

Figure 7 Scatterplots with the z-standardized and log transformed average response time on the x-axis and self-reported conscientiousness as the criterium. The black line is a linear regression line, while the red dashed line is a local regression line based on the loess formula.

In contrast, no significant correlation was found between the log-transformed median time per page and the measure for "objective" conscientiousness (IC retrieval score). The log-transformed variability of response times over survey pages showed no correlation to neither self-reported nor "objective" conscientiousness in the overall sample. However, when careless responders were excluded the variability in response time correlated significantly and negatively ($r(85) = -.22$, $p = .045$) with self-reported trait conscientiousness.

No relationship was found between any conscientiousness measure and the average clicking accuracy, regardless the in- or exclusion or careless responders. In contrast, a person's variability in their click-inaccuracy over several pages showed a significant positive correlation ($r(152) = .16$, $p = .042$) to the "objective" conscientiousness, when careless responders are included in the analysis. In the clean sample 2, a positive trend could be observed with $r(85) = .21$, $p = .054$.

Additionally, the sum of failed attention checks (FAC score) showed a significant negative correlation ($r(152) = -.19$, $p = .016$) to the IC retrieval score. Persons who read the survey carelessly and therefore failed attention checks, also showed a low performance when they were asked to answer questions about the informed consent. This analysis can not be found in table 4, because the FAC score was used to filter careless responders out. The average and variability of both time per page and click-inaccuracy highly correlated with each other ($r > .50$, $p < .01$). This indicates multicollinearity and should be accounted for in the multiple regression analysis.

Table 3 Pearson's correlations with 95% confidence intervals for the sample without outliers (N = 154) including careless responders.

Variable	1	2	3	4	5	6
1. self-reported conscientiousness						
2. „objective" conscientiousness (IC retrieval score)	-.10 [-.25, .06]					
3. careless responding score (FAC score)	.12 [-.04, .27]	-.19* [-.34, -.04]				
4. average response time (log-transformed)	-.21** [-.36, -.05]	.04 [-.12, .20]	.00 [-.16, .16]			
5. variability in response times (log-transformed)	-.14 [-.29, .02]	-.06 [-.22, .10]	-.12 [-.27, .04]	.58** [.46, .68]		
6. average click-inaccuracy	-.03 [-.19, .13]	.15 [-.01, .30]	-.05 [-.20, .11]	-.17* [-.32, -.01]	.02 [-.14, .18]	
7. variability in click-inaccuracy	-.05 [-.20, .11]	.16* [.01, .31]	-.05 [-.20, .11]	-.08 [-.23, .08]	-.08 [-.23, .08]	.57** [.46, .67]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 4 Pearson's correlations with 95% confidence intervals for the clean sample without outliers and careless responder (N = 87).

Variable	1	2	3	4	5
1. self-reported conscientiousness					
2. „objective" conscientiousness (IC retrieval score)	-.09 [-.30, .12]				
3. average response time (log-transformed)	-.33** [-.50, -.13]	-.01 [-.22, .20]			
4. variability in response time (log-transformed)	-.22* [-.41, -.01]	-.10 [-.31, .11]	.62** [.48, .74]		
5. average click-inaccuracy	-.07 [-.27, .15]	.12 [-.09, .32]	-.16 [-.36, .05]	.03 [-.18, .24]	
6. variability in click-inaccuracy	-.05 [-.25, .17]	.21 [-.00, .40]	-.09 [-.30, .12]	-.06 [-.27, .15]	.50** [.32, .64]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Diagnosing the inverted U-shape effect (H1)

The two H1 hypotheses assumed an inverted U relationship between conscientiousness measures and the average response time per survey page. A significant negative linear relationship between self-report conscientiousness and the average response time in both samples (with and without careless responders) with given homoscedasticity was found. The additionally plotted local regression with loess smoothing (see figure 7) does also support a linear relationship. Consequently, the U-shape diagnostic approach using the two-lines analysis is redundant for the self-report measure in this thesis.

Regarding the approximately zero correlation between "objective" conscientiousness and the response time per page, the diagnostics for an inverted U-shape effect were conducted. First a linear and a quadratic regression were applied, followed by the two-lines analysis. For the linear and quadratic regression, the assumption of homoscedasticity might be injured. The usage of robust standard errors with the correction factor of "HC3" did not change the p-values. Therefore, the non-corrected p-values and confidence intervals are reported in table 5. Neither quadratic nor linear models irrespective of the in- or exclusion of careless responders performed better than the mean model. Both model fits showed $F \leq 0.28$ with $p \geq .601$. Consequently, none of the response time regression coefficients was significant, as all $p \geq .601$.

Also, the two-lines analysis neither found a U-shaped effect, nor any linear relationships in both halves of the data, irrespective of the in- or exclusion of careless responders. Within the sample with careless responders (N = 154) the breakpoint of the Robin-Hood algorithm (Simonsohn, 2018) was at the z-standardized response time of $z = -0.20$. The slope left to the breakpoint was not significant with $b = -0.01$, $p = .883$ and the slope right side was also insignificant with $b = 0.02$, $p = .674$. Within the clean sample without careless responders (N = 87), similar results were found with the same breakpoint as in the bigger sample: the slopes of both sides of the breakpoint were insignificant with $b = -0.03$, $p = .565$ on the left side and $b = 0.02$, $p = .723$ on the right side.

Out of the author's exploratory interest, the two-lines analysis was additionally conducted with the overall sample (N = 194) which included all outliers and careless responders and where the average response time was not log-transformed (see figure 8). Although there was still no significant inverted U-shape effect, one could see the curvature within the data based on the grey dashed local regression line slightly resembling a 3rd degree polynomial. The breakpoint of the interrupted regression is located at the z-standardized response time of $z = 1.13$. There is no relationship between the response time per page and self-report conscientiousness in the below average to average range. Although a slight curvature can be seen there, the residuals are too high to make this curvature interpretable. In the above average range right of the breakpoint a positive trend emerges indicating that longer response times per page cooccurs with higher self-reported conscientiousness. In other

words, it seems that response time outliers tend to assign themselves a higher conscientiousness. These findings are based on outliers and should therefore be interpreted very carefully. Ultimately this shows that outliers themselves might be an interesting subgroup of people.

In order to replicate previous zero-correlations a simple linear regression was additionally calculated on the raw data set (N = 194) as can be seen as a scatter plot in figure 8. The simple linear regression model was not substantially better than the mean model with $F(2, 192) = 0.76$, $p = .38$ and 0.4% explained variance. It showed a slope of $b = 0.00$, $p = .385$, which is equivalent to a Pearson's $r(192) = -0.06$.

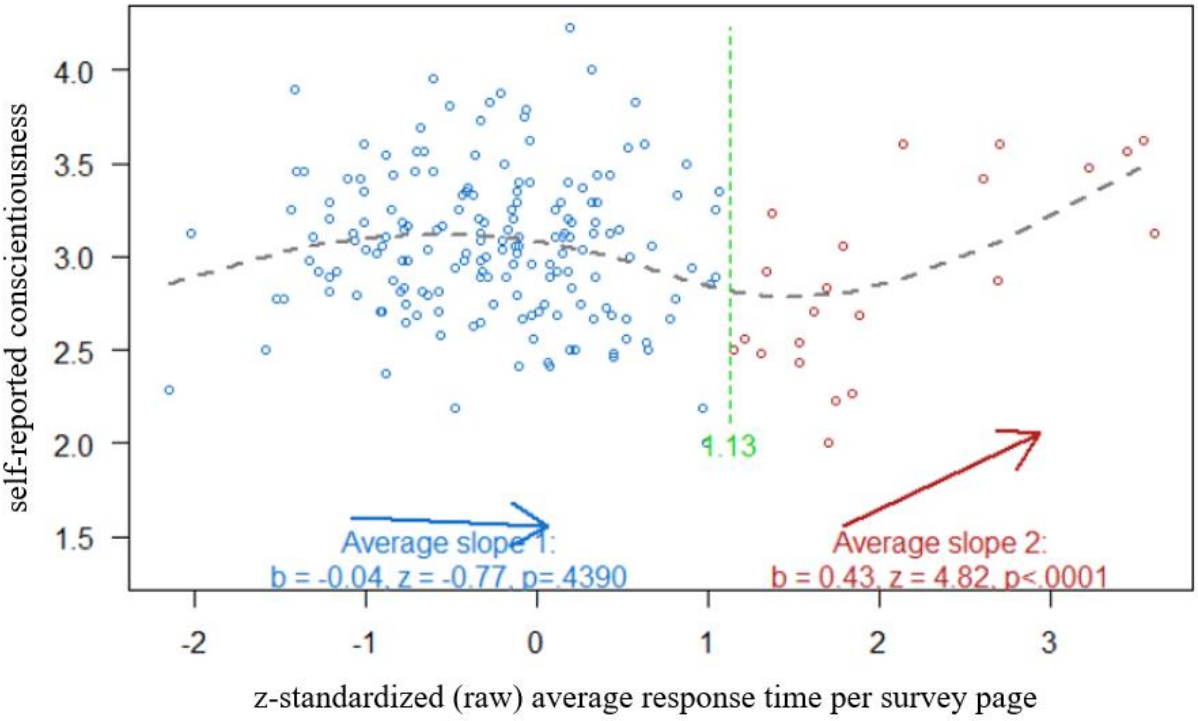

Figure 8 The two-lines analysis (Simonsohn, 2018) testing if a U-shaped relationship is found between the z-standardized (raw) average response time and the self-reported conscientiousness is found in the full sample (N = 194) including all outliers and careless responders. The plot is automatically generated by the analysis script. The green dashed line with the z = 1.13 indicates the breakpoint calculated by the Robin-Hood algorithm.

Overall Regression Model (H3)

In a final regression model, the importance of paradata indicators as predictors for conscientiousness measures was tested simultaneously (Table 6). As no curvilinear relationship was found, only a forced entry linear regression was conducted. The average and variability in response time per page, as well as the average and variability of click accuracy were used as predictors for the conscientiousness measures. Homoscedasticity was given in both samples and with both conscientiousness measures. The earlier found potential multicollinearity was not problematic, as all VIF ≤ 1.73.

Table 5 The linear and quadratic regression predicting observed conscientiousness with the z-standardized and log-transformed average response time per survey page stratified for both samples.

	Sample with Careless Responders (N = 154)		Clean Sample without Careless Responders (N = 87)	
Linear Model on	$R^2 = 0$		$R^2 = 0$	
Observed Conscientiousness	F(1, 152) = 0.28, p = .601		F(1, 85) = 0.01, P 0.921	
	b [95%CI]	Beta [95%CI]	b [95%CI]	Beta [95%CI]
Intercept	0.62 **	-	0.66 **	-
	[0.59, 0.66]		[0.62, 0.70]	
Response time	0.01	.04	-0.00	-.01
	[-0.03, 0.05]	[-.12, .20]	[-0.04, 0.04]	[-.23, .20]
Quadratic Model on	$R^2 = 0$		$R^2 = 0$	
Observed Conscientiousness	F(2, 151) = 0.14, p = .872		F(2, 84) = 0.09, P = .918	
	b [95%CI]	Beta [95%CI]	b [95%CI]	Beta [95%CI]
Intercept	0.62 **	-	0.66 **	-
	[0.58, 0.67]		[0.61, 0.71]	
Response time	0.01	0.04	-0.00	-.01
	[-0.03, 0.05]	[-0.12, 0.20]	[-0.04, 0.04]	[-.23, .21]
Response time^2	0.00	0	0.01	.04
	[-0.03, 0.03]	[-0.16, 0.16]	[-0.02, 0.04]	[-.17, .26]

Table 6 The overall regression models predicting self-reported and observed conscientiousness with all paradata indicators stratified for both samples in- or excluding careless responders.

	Self-reported Conscientiousness				Observed Conscientiousness			
	With CR (N = 154)		Without CR (N = 87)		With CR (N = 154)		Without CR (N = 87)	
	R^2_{adj} = .02 F(4, 149) = 1.94, p = .107		R^2_{adj} = .08 F(4, 82) = 2.84, p = .029		R^2_{adj} = .02 F(4, 149) = 1.93, p = .109		R^2_{adj} = .01 F(4, 82) = 1.30, p = .278	
	b [95%CI]	Beta [95%CI]	b [95%CI]	Beta [95%CI]	b [95%CI]	Beta [95%CI]	b [95%CI]	Beta [95%CI]
Intercept	3.08** [2.99, 3.18]	-	3.08** [2.95, 3.21]	-	0.56** [0.50, 0.62]	-	0.61** [0.53, 0.68]	-
Avg. Response time (log-transformed & z-standardized)	-0.07* [-0.14, -0.00]	-0.21 [-0.41, -0.01]	-0.11* [-0.20, -0.03]	-0.35 [-0.62, -0.08]	0.03 [-0.01, 0.08]	0.15 [-0.04, 0.35]	0.02 [-0.03, 0.07]	0.12 [-0.16, 0.40]
Variability in Response time (log-transformed & z-standardized)	-0.01 [-0.08, 0.06]	-0.02 [-0.22, 0.18]	0.00 [-0.09, 0.09]	0.00 [-0.26, 0.27]	-0.03 [-0.08, 0.01]	-0.15 [-0.34, 0.05]	-0.03 [-0.08, 0.02]	-0.17 [-0.44, 0.11]
Avg. Click Inaccuracy	-0.00 [-0.2, 0.2]	-0.04 [-0.24, 0.16]	-0.01 [-0.03, 0.01]	-0.11 [-0.35, 0.13]	0.01 [-0.00, 0.02]	0.12 [-0.08, 0.32]	0.00 [-0.01, 0.02]	0.05 [-0.20, 0.30]
Variability in Click Inaccuracy	-0.01 [-0.08, 0.05]	-0.04 [-0.24, 0.15]	-0.01 [-0.09, 0.07]	-0.02 [-0.26, 0.22]	0.02 [-0.02, 0.06]	0.10 [-0.10, 0.29]	0.03 [-0.01, 0.08]	0.18 [-0.06, 0.43]

There was no significant predictor for the observed conscientiousness in neither the sample with careless responders ($F(4, 149) = 1.93$, $p = .109$, $R^2_{adj} = .02$) nor without those ($F(4, 82) = 1.30$, $p = .278$, , $R^2_{adj} = .01$). Regarding self-reported conscientiousness, the log-transformed and scaled average response time was a significant predictor ($b = -0.11$ [-0.20, -0.03], $p = .012$, $r_{semi\text{-}part} = .07$) only in the clean sample without careless responders ($F(4, 82) = 2.84$, $p = .029$, $R^2_{adj} = .08$). The higher the response time per page, the lower the self-reported conscientiousness. When including careless responders, the regression model showed no improvement compared to the mean ($F(4, 149) = 1.94$, $p = .107$, $R^2_{adj} = .02$) but the log-transformed and scaled average time per page was a barely significant predictor ($b = -0.07$ [-0.14, -0.00], $p = .042$, $r_{semi\text{-}part} = .03$) with the upper border of the confidence interval deviating from zero at the 3rd decimal place behind zero. Here again, a higher response time per page cooccurs with a lower score of self-reported conscientiousness.

Discussion

This thesis covered the question if a selection of paradata indicators as measures for behavior are associated with the personality trait conscientiousness. Two different measures for conscientiousness were used: the IPIP-240 (Schreiber & Iller, 2016) subscale for conscientiousness was used as a self-report measure. The percentage of correct answers in five retrieval questions about information presented in the informed consent at the beginning of the survey was used as an 'objective' behavioral indicator of the personality trait (Theiss et al., 2014). Further, the average response time and the variability in response time across six survey pages were calculated and used as paradata indicators. The same applies to the average clicking inaccuracy and the variability in clicking inaccuracy.

Altogether, the inverted U-shaped relationship between a person's response time per page and their conscientiousness as well as most of the hypothesized correlations and therefore also the assumed significant predictors in multiple regression did not emerge. The low-stake situation in the online survey, imperfect measures, extreme outliers, and a high rate (~ 40%) of careless responders indicate a poor data quality. In fact, this rate of careless responders is about four times higher than expected in an undergraduate sample and approximates findings from studies using Amazon's Mechanical Turk (Meade & Craig, 2012; Nichols & Edlund, 2020). Despite those caveats this thesis still offers some conclusions and new starting points for future research, which are presented later. Before we discuss the results and hypotheses in more detail, first the hypotheses are repeated.

The variability in response time per page and the variability in click-inaccuracy were expected to be negatively correlated to both conscientiousness measures, as a persistent and consistent behavior (i.e., low variability) is attributed to (higher) conscientiousness (De Raad & Peabody, 2005; MacCann

et al., 2009; Stanek & Ones, 2018). The average click-inaccuracy was also hypothesized to be negatively correlated to any conscientiousness measure, because conscientious people were expected to work more accurately and therefore have a lower inaccuracy (e.g. Jackson et al., 2010). The average response time per page was expected to show an inverted U-relationship indicating extreme fast and extreme slow response times to be associated with low conscientiousness (Maniaci & Rogge, 2012). Ultimately, all beforementioned variables should be tested simultaneously as predictors for the conscientiousness measures in a multiple regression. Within this scope of these hypotheses the term careless responding plays an essential role, as careless responding is not alone by empirical evidence (e.g., Nichols & Edlund, 2020) but also by definition the absence of conscientious responding. As careless responders further endanger the validity of scientific examination by adding error variance and therefore increasing both alpha and beta error (Nichols & Edlund, 2020; Arias et al., 2020, Goldhammer et al., 2020; Leiner, 2019), a conservative outlier analysis was conducted to provide more stable effects, which do not rely on outliers (e.g., Höhne & Schlosser, 2018). In order to allocate the influence of careless responders, two samples, which both excluded outliers defined by Hoaglin et al. (2000, in Höhne & Schlosser, 2018), were used for the analyses in this thesis. The bigger sample (N = 154) included careless responders, while the latter were excluded in the clean smaller sample (N = 87).

The influence of careless responders

The scores of self-reported conscientiousness were similar in both samples. In contrast, the clean sample without careless responders showed a higher average in “objective” conscientiousness than the sample including careless responders. More specifically, the latter sample had on average (median) one informed consent retrieval item less correct than the clean sample. Regarding internal consistencies, careless responders had no influence on the self-report but on the “objective” conscientiousness, as their exclusion led to a drop in Cronbach’s alpha. One possible reason could be, that careless responders performed consistently bad on those retrieval questions, as careless reading the survey content showed itself in a low retrieval score about the respective content. Further, careless responders who showed a consistently worse performance in the IC retrieval items ultimately increased the item correlation. By removing those careless responders, the source for the item correlation also got removed. The resulting “average” sample without assumed careless responders had a higher average IC retrieval score and a lower variance but simultaneously a lower inter item correlation (i.e., Cronbach’s α), as the consistently bad performers were excluded. The exclusion of careless responders led to an increase in the internal consistency within the response time per page.

In sum, careless responders showed their influence predominantly in the IC retrieval questions assessing “objective” conscientiousness and Cronbach’s α in the response time across pages. Except

for the response time variables, the exclusion of careless responders additionally led to a decrease within the dispersion statistics (i.e., SD and MAD). The sometimes found increase of SD within the response time variables after exclusion of careless responders might be the effect of careless responders also showing average response times. Not all careless responders showed above or below average response times. And by excluding observations around the mean while keeping the dispersion relatively similar results in a higher SD. Summa summarum, careless responders are a diffuse subgroup who typically increase the variance within the data. Regarding response times, careless responders exist in the whole bandwidth: below average, average and above average. These results are consistent with the idea, that careless responders are either pretty fast in working through the survey as they respond carelessly, are pretty slow due to breaks and distractions or show average response times by careless responding with additional breaks.

Linear relationships between conscientiousness measures and paradata indicators (H2)

Before discussing the hypothesized linear relationships between paradata indicators and conscientiousness measures, the unrelatedness of both conscientiousness measures must be discussed. No correlation was found between self-reported conscientiousness and the percentage of correct retrieval items about the informed consent. A small correlation could be explained by the trait method specificity (Koch et al., 2014). Basically, the same measurement methods (e.g., self-report) share variance due to assessing the same personality trait components, which leads to higher correlations when identical measurement methods are used. However, when different measurement methods are used like a self-report measure and an “objective” measurement, different trait components are measured leading to smaller correlations. In this case there was an absolute absence of any correlation between self-report conscientiousness and the IC retrieval score. This might be the result of an imperfect measure. The IC retrieval questions were generally speaking either too hard or too easy. There were a high number of careless responders, the IC retrieval score was heavily skewed, and its internal consistency was low regardless of careless responders. Another explanation approach would be, that a person’s tendency to read accurately and carefully the informed consent might reflect more state than trait conscientiousness. This is supported by the negative correlation between the IC retrieval score and the number of failed attention checks. Although, the level of a personality trait can be defined as the frequency of state behavior (Buss & Craik, 1983), the usage of the highly unspecific global score of a self-report measure might further reduced the covariance. Nevertheless, future research could use established objective personality tests (OPTs) for conscientiousness and

additionally use subfacets of self-report conscientiousness scales to circumvent any absence between the personality trait measures.

Regarding linear relationships, both paradata indicators comprising a person's variability in either their response time (H2.1) or clicking accuracy (H2.3) should reflect a person's impulse control by measuring how persistent and consistent they answered the survey. A low variability indicates a consistent behavior, which is expected in conscientious persons (e.g., Jackson et al., 2012). Consequently, both variability measures were expected to show a negative linear relationship with conscientiousness measures. *When using the bigger sample with careless responders, no correlation between self-reported conscientiousness and the variability in response time across pages could be detected. But when using the clean sample, a significant negative correlation occurred, which supports H2.1.1. Consequently, a more consistent behavior was found in participants who reported a higher conscientiousness. The insignificant negative correlation within the bigger sample which contains careless responders, can be assigned to the increased error variance introduced by careless responders (*Nichols & Edlund, 2020; Arias et al., 2020, Goldhammer et al., 2020; Leiner, 2019). Irrespective of the sample, no correlation emerged between the variability in response time and "objective" conscientiousness. This might be due to either the imperfect psychometric characteristics of this scale, the trait method specificity (e.g., Koch et al., 2014) or the assessment of a different facet of conscientiousness, as already discussed above. *Ultimately, hypothesis H2.1.2 is not supported.*

No linear relationship between self-reported conscientiousness and the variability in a person's clicking accuracy was found in neither of the two samples, for which *H2.3.1 is not supported*. A significant positive correlation between the "objective" conscientiousness (i.e., IC retrieval score) and a person's variability in their clicking inaccuracy was found only in the bigger sample including careless responders. This result should be interpreted very carefully. First, the measure of "objective" conscientiousness might be flawed, which consequently questions the construct indicated by the retrieval questions. Therefore, no statement can be made about its relationship with conscientiousness. It seems that persons, who show inconsistent clicking accuracy tend to have higher scores in the retrieval questions. As this correlation only occurs, when careless responders are included, they might be the cause of this correlation. Careless responders are known for adding error variance, which can create spurious correlations (Nichols & Edlund, 2020; Arias et al., 2020; Goldhammer et al., 2020, Leiner, 2019). Another explanation can be found in the device used. Participants who use devices which rely on touch screens might have a lower accuracy simply by the fact, that (most) fingers are thicker and therefore more inaccurate than the mouse curser of a computer. Consequently, users with touch devices have automatically a higher clicking inaccuracy and a higher variability in clicking inaccuracy as they can't control their accuracy in a way a mouse-users do. In both samples used for the analyses almost 30% of the participants used a device, where a touchscreen

is presumed based on their paradata. The final possible explanation for the result is based on statistical power. As the correlation reached significance only in the bigger sample, the loss of statistical power by having a smaller sample might also be the reason for this finding. By using the program G*Power (v3.1.9.7; Faul, Erdfelder, Buchner, & Lang, 2009) a post-hoc power of 50.83% can be calculated for the smaller, clean sample, where no effect was found. Therefore, the analysis' ability to detect such an effect is just chance. But in contrast, the achieved power of the significant correlation when careless responders are included is still only at 51.55%. *In sum, the analyses are underpowered, the used device might be a moderator and careless responders are known add error variance, which in return results in large confidence intervals (Crutzen & Peters, 2017) like in our case. Consequently, this significant positive correlation is likely to be spurious, for which H2.3.2 is not supported.*

Based on the assumption that conscientious persons tend to work accurately, they are also expected to have a higher average clicking accuracy answering the survey questions. In the H2.2 hypotheses negative correlations were assumed between the conscientiousness measures and the average clicking inaccuracy. *As no correlations emerged between any conscientiousness measure and the average clicking inaccuracy irrespective of careless responders, H2.2.1 and H2.2.2 are not supported.* Like noted above, a possible explanation for the not-occurrence of these correlations is, that the clicking accuracy is not only dependent on a person's personality, but also on the used device.

Not hypothesized and also not very surprising is the significant negative correlation between the careless responding score (i.e., the number of failed attention checks) and its relationship to the "objective" conscientiousness. It is reasonable that careless responding persons who failed one or more attention checks also miss information in the informed consent, which results in a bad performance in the retrieval task. Further, this finding supports the assumption that the IC retrieval score more reflects a person's state conscientiousness.

Unexpectedly, a significant negative correlation between self-reported conscientiousness and the log-transformed response time per page was found, which will be discussed in the next section.

No sign of an inverted U relationship between response time and conscientiousness (H1)

An inverted U-shaped relationship was expected between self-reported conscientiousness and the log-transformed average response time per page. After excluding careless responders, this effect even increased, supporting the claims of negative effects of careless responders (Nichols & Edlund, 2020; Arias et al., 2020; Goldhammer et al., 2020). Further analyses showed homoscedasticity in both samples and additional plotting of the relationship supports the assumption of a linear relationship. Consequently, persons who have a longer average response time per survey page also report lower conscientiousness. This stands in contrast to previous findings of Ward and Pont (2015) and Bowling

et al. (2021), who connected faster response times with more careless responding (i.e., a less conscientious behavior). Further, this result is also contrary to the previous found zero-correlation (McKay et al, 2018; Maniaci & Rogge, 2012). Nevertheless, such zero-correlation was also replicated in the raw data set, which contained outliers and raw (i.e., not log-transformed) response times. Bowling et al. (2021) further found, that "local" item response times are superior in detecting careless responders (i.e., less conscientiously behaving participants) than the "global" total survey time. Since both Maniaci and Rogge (2012) and McKay et al. (2018) used the total survey times, a more liberal outlier analysis (e.g., windsoring) and did not log-transform their response times, their found zero-correlations might be due to all these design decisions. Generally, it stands to reason that outliers (i.e., extremely fast and extremely slow responders) as well as the (not) conducted log-transformation are the origin for the heterogeneous findings in literature. To sum it up altogether: fast response times can be associated with low conscientiousness both theoretically and empirically (Ward & Pond, 2015; Bowling et al., 2021) indicating a positive correlation. In contrast to that, we found small empirical evidence for a negative linear relationship between response time and self-report conscientiousness when applying a conservative outlier analysis and log-transformation on the response time variables. It stands to reason that there might be indeed an inverted U-shaped relationship between conscientiousness and response times, when a bigger range of response times is used. But we also found small evidence for higher conscientiousness in slow responders. It can be argued that those few slow responders were slow because of their careful and effortful responding due to their trait conscientiousness. Alternatively, the used global self-report conscientiousness score is heterogeneous, as multiple behaviors, thoughts and feelings were asked and subsumed to an overall score. Maybe those slow responders are usually conscientious in most of the facets within the IPIP-240 but take breaks and do other careless things (i.e., be not very conscientious) in low stake online surveys, which made them noticeable for the outlier analysis or the careless responding detection method. This is consistent with the quite average distribution of self-report conscientiousness scores, where none of the participants reached the end of the scale. Consequently, *hypothesis H1.1 is not supported, as a linear relationship instead of a nonlinear relationship between response time and self-reported conscientiousness was found.*

When testing for the inverted U-shape effect between "objective" conscientiousness (IC retrieval score) and the average log-transformed response time no significant linear or curvilinear effect occurred in neither the linear, quadratic or two-lines analysis. *Consequently H1.2 is also not supported*. There might be several explanations for the lack of any relationship between those two variables. Like discussed earlier the measure for "objective" conscientiousness is likely to be flawed or measures state conscientiousness.

Another reason which also applies on self-reported conscientiousness could be the exclusion of outliers. These exclusion criteria might also have been the reason for the linear relationship between self-reported conscientiousness and the average log-transformed response time. The outlier analysis was conservative, leading to the exclusion of around 20% of the cases. The idea behind this was to test the hypotheses only in a more normal (and less extreme) population which ultimately led to more stable and robust statistical results. On the other hand, persons with a very low conscientiousness were expected to respond carelessly in the survey. We expected them to be either extremely fast or extremely slow. And exactly these cases were excluded. There are two sides on this coin: on the one side are the interesting outliers, which might support the inverted U-shape effect. On the other side the same participants bias statistical analyses leading to unstable statistical results. To make sure that outliers are in fact a very interesting group of participants, who have low conscientiousness, a bigger sample size is needed. Assuming that (some) outliers are in fact just random "noise", and some are not, a bigger sample with consequently more outliers and therefore higher statistical power can show, if their tendency to be more or less conscientious is statistically relevant. In our exploratory two-lines analysis (see figure 8) the low(er) self-reported conscientiousness of the very fast responders is mainly influenced by 2 or 3 participants. If these cases were excluded the seemingly curvilinear relationship on the left side of the breakpoint would be weakened or even vanish. Ultimately, depending on the outlier analysis, trimming or data transformation one could show a linear (with our robust outlier analysis), an inverted U-shaped (excluding the slow responders on the right side of the breakpoint), a U-shaped (excluding only few very fast and very slow responders) or even a 3rd degree polynomial (not exclusion at all) relationship. This emphasizes the importance of a robust and well documented data preparation and outlier analysis, so that such approaches can be discussed in the scientific community in order to address the replication crisis. In sum, a stable negative relationship between response time and self-reported conscientiousness was found within a sample of average responding participants. No final conclusion can be made regarding an inverted U-shaped relationship, as the full range (i.e., including outliers) of response times is needed for that. To determine if fast and slow outliers are a special subpopulation with lower conscientiousness, a bigger sample size is necessary.

Multiple regression examining the unique effect of paradata indicators (H3)

When estimating the unique effect of a single paradata indicator while controlling for the others, only the scaled average log-transformed response time per page showed a significant relationship to the self-reported conscientiousness. The earlier found significant negative correlation between the

variability in a persons' response time and self-reported conscientiousness vanishes. This indicates, that the earlier found effect was a result of shared variance between the average and variability in response time, which was also found in their high positive correlation with each other. *Consequently, H3.1 is not supported as only the average response time is a significant predictor for self-reported conscientiousness while its variability as well the average and variability of click inaccuracy are not.*

Regarding the "objective" conscientiousness as the criterion, none of the average and variability of response time and clicking inaccuracy was a significant predictor. Here again it can be assumed that retrieval questions of the informed consent are more likely a state measure for conscientiousness. And this state measure (i.e., IC retrieval score) depicts another behavior within the domain of conscientiousness than behavioral consistency, response time or click accuracy. *Because in neither sample any paradata indicator reached significance in explaining the IC retrieval score, H3.2 is not supported.*

Limitations

One limitation is the measure for "objective" conscientiousness. Even when expecting a low correlation between self-report and the behavioral observations as a result of the method specificity (Koch et al., 2014), the absolute absence of any relationship indicates either a flawed measure. Or a measure which captures state conscientiousness instead of trait conscientiousness, when regarding its correlation to the careless responding indicator (FAC score). Future research should only use validated measures as behavioral indicators for conscientiousness like OPTs when examining the validity of paradata indicators as indicators for conscientiousness.

Another limitation was the decision to conduct a conservative outlier analysis in order to make statistical results less biased and better replicable. Outliers might be a special group with low conscientiousness, who might be to blame for the inconsistent findings in literature regarding the relationship between response time and conscientiousness. Outlier response times might either be the result of personality (i.e., conscientiousness) or a result of random error like a distraction through a ringing doorbell, although the environment (e.g., the own flat) normally is quiet and suitable for answering surveys. To address such random error and simultaneously examine if outliers are truly a relevant subpopulation, a big sample size is necessary to get enough statistical power to average out random error and detect true effects. Our conservative outlier analysis identified approximately 20% of our total sample (N = 194) as outliers. The error variance within an outlier sample is expected to be higher than in the "average" sample, which is also the case in this thesis (compare table 2 and table A1 in the appendix). Additionally, the 40 outliers have a very similar self-reported conscientiousness in both outlier free samples. As our outlier sample size is only small and its variance is higher than in

the "average" outlier free samples, this small effect size (i.e., mean difference) might be underestimated. But solely based on this small mean difference between both groups, one could calculate the estimated sample size to detect mean differences between outliers and "average" samples. When a small Cohen's d = 0.2 is expected, a two-sided $\alpha = 5\%$ and a statistical power of 80% with an outlier to "average" sample ratio of 0.2 (because of the 20% identified outliers) is assumed, one would need a sample size of N = 1416 participants. The needed sample size gets even bigger when a smaller fraction of outliers is expected or when a higher statistical power is wanted. But in contrast, a smaller faction of outliers might be even "less normal" in comparison to the outlier free sample, which would assumably result in bigger mean differences, which reduce the needed sample size. In conclusion, to examine outliers as a special group in a low-stake situation like online surveys a very big sample is necessary, which is hard to achieve. But if outliers truly are a subgroup with a low conscientiousness, the easy identification of an outlier consequently would indicate their personality. This would be an advantage in personality assessment.

Neither the average nor the variability in clicking accuracy was a significant predictor for conscientiousness. This could be due to the used device of the participants, as smartphone or tablet users might have a higher clicking inaccuracy in comparison to computer mouse users solely by the usage of a touch screen and therefore irrespective of their personality. Future research should explicitly control for the used device when using paradata indicators to assess ones' personality.

Additionally, it should be mentioned, that paradata indicators can be calculated differently. The 6 response times and click-inaccuracies over 6 survey pages were averaged with robust measures like the median and MAD. If we had used parametric measures like the arithmetic mean and the standard deviation, outlier values would have had significantly more influence on the average values. This would have led to a bigger variance, bigger bias through outliers and ultimately to less stable statistical results. On the other hand, potentially more evidence for the true relationship between conscientiousness and response times could have been found. Future research should therefore additionally focus on other paradata calculation methods. For instance, Bowling et al. (2021) reported a successful validation of a 2 seconds per item response time cut-off value as an indicator for careless responding, which can be examined as an indicator for conscientiousness in future research.

Finally, this thesis shed light onto the relationship between conscientiousness and its manifestation in paradata indicators. Structurally speaking, those regression models predicting conscientiousness are wrong, as the personality trait is expected to be causally responsible for the manifestation in the paradata indicators. But in order to test their ability being manifest variables of a personality trait this procedure is justified. Future research should follow up from this point and do structural equation modelling about conscientiousness being the responsible trait for different paradata indicators.

Conclusion

This thesis examined if paradata indicators like response time, mouse/touch click accuracy and response consistency show shared variance to conscientiousness measures. To address bias and increased alpha and beta errors a conservative exclusion of outliers and further analyses with and without careless responders were conducted. A person's clicking accuracy, as well as response consistency measures like the variability in response time and the variability in clicking inaccuracy showed no significant correlation to conscientiousness. The assumed reason for the inability of clicking inaccuracy to explain conscientiousness is its dependency of the used device and therefore its input modality (i.e., touch screen). But before dismissing clicking accuracy as irrelevant, future research should test its relationship to conscientiousness measure while controlling for the input device.

The person's response time showed a significant negative linear relationship with self-reported conscientiousness. Consequently, no (inverted) U-shape relationship was found in this thesis. Nevertheless, the used outlier analysis and data preparation approach was identified as one explanation how other researchers found either linear positive (Berry et al., 2019), linear negative (Bowling et al., 2016; Nichols & Edlund, 2020) or zero-relationships (Maniaci Rogge, 2012; McKay et al., 2018) between response time and conscientiousness. Future research should focus on outliers to test, if and under which circumstances a U-shaped relationship occurs.

Multiple regression analyses showed that the originally significant negative relationship between a person's variability in response time and self-reported conscientiousness is due to shared variance with a person's average response time per survey page. Consequently, the variability in response time seems to be no predictor of conscientiousness.

While self-reports are only one possible measure for personality assessments, other assessments approaches should be used as well. This thesis assumed, that a person's ability in recalling information about the informed consent is an indicator for trait conscientiousness, as conscientious persons are expected to read the informed consent more thoroughly (Theiss et al., 2014, Bowling et al., 2021). The complete absence of a correlation between self-report and informed consent retrieval items questions the assessed construct. Future research should use best practice assessment tools to test, if paradata indicators are economic indicators for the personality.

Overall, this thesis covered only a tiny fraction of possible paradata indicators. From those examined, response time seems to be a useful and economic indicator for conscientiousness, if data preparation and an outlier analysis are performed. Although the other paradata indicators used in this thesis failed to show shared variance to conscientiousness, the depicted approaches in data preparation, outlier analysis and additional identified confounders provide new starting points for future research regarding data quality in online research. Although much future research has to be

done in this field, it seems possible to create a set of paradata indicators to assess personality. And the benefits in research and aptitude testing arising from this will be worth the effort.

Literature

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*(6), 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Arthur Jr., W., Hagen, E., & George Jr., F. (2020). *The Lazy or Dishonest Respondent: Detection and Prevention*. 33.

Asendorpf, J. B. (2011). *Persönlichkeitspsychologie für Bachelor* (2. überarbeitete und aktualisierte Auflage). Springer.

Auguie, B., & Antonov, A. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics* (2.3) [Computer software]. https://CRAN.R-project.org/package=gridExtra

Barrick, M. R., & Mount, M. K. (1991). THE BIG FIVE PERSONALITY DIMENSIONS AND JOB PERFORMANCE: A META-ANALYSIS. *Personnel Psychology, 44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Berry, K., Rana, R., Lockwood, A., Fletcher, L., & Pratt, D. (2019). Factors associated with inattentive responding in online survey research. *Personality and Individual Differences, 149*, 157–159. https://doi.org/10.1016/j.paid.2019.05.043

Bleidorn, W., & Hopwood, C. J. (2019). Using Machine Learning to Advance Personality Assessment and Theory. *Personality and Social Psychology Review, 23*(2), 190–203. doi: https://doi.org/10.1177/1088868318772

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., Brower, C. K., & Bragg, C. B. (2021). The Quick and the Careless: The Construct Validity of Page Time as a Measure of Insufficient Effort Responding to Surveys. *Organizational Research Methods*. https://doi.org/10.1177/10944281211056520

Buss, D. M., & Craik, K. H. (n.d.). *The Act Frequency Approach to Personality*. 22.

Callegaro, M. (2013). Paradata in Web Surveys. In F. Kreuter (Ed.), *Improving Surveys with Paradata* (pp. 259–279). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118596869.ch11

Credé, M. (2010). Random Responding as a Threat to the Validity of Effect Size Estimates in Correlational Research. *Educational and Psychological Measurement, 70*(4), 596–612. https://doi.org/10.1177/0013164410366686

Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. (1989). *Journal of Clinical Psychology, 45*(2), 12.

Crutzen, R., & Peters, G.-J. Y. (2017). Targeting Next Generations to Change the Common Practice of Underpowered Research. *Frontiers in Psychology*, *8*, 1184. https://doi.org/10.3389/fpsyg.2017.01184

Day, N. E., Hudson, D., Dobies, P. R., & Waris, R. (2011). Student or situation? Personality and classroom context as predictors of attitudes about business school cheating. *Social Psychology of Education*, *14*(2), 261–282. https://doi.org/10.1007/s11218-010-9145-8

de Bruin, G. P., & Rudnick, H. (2007). Examining the Cheats: The Role of Conscientiousness and Excitement Seeking in Academic Dishonesty. *South African Journal of Psychology*, *37*(1), 153–164. https://doi.org/10.1177/008124630703700111

De Raad, B., & Peabody, D. (2005). Cross-culturally recurrent personality factors: Analyses of three factors. *European Journal of Personality*, *19*(6), 451–474. https://doi.org/10.1002/per.550

de Vries, R. E., & van Gelder, J.-L. (2013). Tales of two self-control scales: Relations with Five-Factor and HEXACO traits. *Personality and Individual Differences*, *54*(6), 756–760. https://doi.org/10.1016/j.paid.2012.12.023

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The Differential Impacts of Two Forms of Insufficient Effort Responding: IMPACT OF DIFFERENT TYPES OF IER. *Applied Psychology*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Diedenhofen, B., & Musch, J. (2017). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, *49*(4), 1444–1459. https://doi.org/10.3758/s13428-016-0800-7

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., Iwasaki, W., Wenchel, S., … Schwen, B. (2021). *data.table: Extension of "data.frame"* (1.14.0) [Computer software]. https://CRAN.R-project.org/package=data.table

Evans', R. G. (n.d.). *Response consistency among high F scale scorers on the MMPI*. 3.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., … R-Core. (2021). *car: Companion to Applied Regression* (3.0-11) [Computer software]. https://CRAN.R-project.org/package=car

Gagolewski, M., Tartanus, B., & and others. (2021). stringi: Character String Processing Facilities (Version 1.7.4). Retrieved from https://CRAN.R-project.org/package=stringi

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Heller, D., Komar, J., & Lee, W. B. (2007). The Dynamics of Personality States, Goals, and Well-Being. *Personality and Social Psychology Bulletin*, *33*(6), 898–910. https://doi.org/10.1177/0146167207301010

Hendy, N. T., & Montargot, N. (2019). Understanding Academic dishonesty among business school students in France using the theory of planned behavior. *The International Journal of Management Education*, *17*(1), 85–93. https://doi.org/10.1016/j.ijme.2018.12.003

Herzberg, P. Y., & Roth, M. (2014). *Persönlichkeitspsychologie*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-93467-9

Höhne, J. K., & Schlosser, S. (2018). Investigating the Adequacy of Response Time Outlier Definitions in Computer-Based Web Surveys Using Paradata SurveyFocus. *Social Science Computer Review*, *36*(3), 369–378. https://doi.org/10.1177/0894439317710450

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, G., & Mitchell, D. (2020). lmtest: Testing Linear Regression Models (Version 0.9-38). Retrieved from https://CRAN.R-project.org/package=lmtest

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511. https://doi.org/10.1016/j.jrp.2010.06.005

Koch, T., Ortner, T. M., Eid, M., Caspers, J., & Schmitt, M. (2014). Evaluating the Construct Validity of Objective Personality Tests Using a Multitrait-Multimethod-Multioccasion-(MTMM-MO)-Approach. *European Journal of Psychological Assessment*, *30*(3), 208–230. https://doi.org/10.1027/1015-5759/a000212

Kreuter, F. (2013). Improving Surveys with Paradata: Introduction. In F. Kreuter (Ed.), *Improving Surveys with Paradata* (pp. 1–9). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118596869.ch1

Kreuter, F., & Olson, K. (2013). Paradata for Nonresponse Error Investigation. In F. Kreuter (Ed.), *Improving Surveys with Paradata* (pp. 11–42). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118596869.ch2

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

Leiner, D. J. (2019). Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys. *Survey Research Methods*, 229-248 Pages. https://doi.org/10.18148/SRM/2019.V13I3.7403

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24*(7), 1003–1020. https://doi.org/10.1002/acp.1602

Long, J. A. (2021). *jtools: Analysis and Presentation of Social Scientific Data* (2.1.4) [Computer software]. https://CRAN.R-project.org/package=jtools

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of Conscientiousness. *Learning and Individual Differences*, *19*(4), 451–458. https://doi.org/10.1016/j.lindif.2009.03.007

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

Mayerl, J. (2013). *Response Latency Measurement in Surveys. Detecting Strong Attitudes and Response Effects*. https://doi.org/10.13094/SMIF-2013-00005

McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior*, *84*, 295–303. https://doi.org/10.1016/j.chb.2018.03.007

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Muggeo, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, *22*(19), 3055–3071. doi: 10.1002/sim.1545

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

Perugini, M., & Gallucci, M. (1997). *A hierarchical faceted model of the Big Five*. *11*, 23.

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Zivot, E., Cornilly, D., Hung, E., Lestel, M., Balkissoon, K., Wuertz, D., Christidis, A. A., Martin, R. D., Zhou, Z. "Zenith," & Shea, J. M. (2020). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis* (2.0.4) [Computer software]. https://CRAN.R-project.org/package=PerformanceAnalytics

R Core Team. (2021). *R: The R Project for Statistical Computing*. https://www.r-project.org/

Ranger, J. (2013). *Modeling responses and response times in personality tests with rating scales*. 22.

Ranger, J., & Ortner, T. M. (2011). Assessing Personality Traits Through Response Latencies Using Item Response Theory. *Educational and Psychological Measurement*, *71*(2), 389–406. https://doi.org/10.1177/0013164410382895

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research* (2.1.6) [Computer software]. https://CRAN.R-project.org/package=psych

Rinker, T., Kurkiewicz, D., Hughitt, K., Wang, A., Aden-Buie, G., Wang, A., & Burk, L. (2019). *pacman: Package Management Tool* (0.5.1) [Computer software]. https://CRAN.R-project.org/package=pacman

Roberts, B. W., Bogg, T., Walton, K. E., Chernyshenko, O. S., & Stark, S. E. (2004). A lexical investigation of the lower-order structure of conscientiousness. *Journal of Research in Personality*, *38*(2), 164–178. https://doi.org/10.1016/S0092-6566(03)00065-5

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). THE STRUCTURE OF CONSCIENTIOUSNESS: AN EMPIRICAL INVESTIGATION BASED ON SEVEN MAJOR PERSONALITY QUESTIONNAIRES. *Personnel Psychology*, *58*(1), 103–139. https://doi.org/10.1111/j.1744-6570.2005.00301.x

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & Meints, J. (2009). Conscientiousness. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 369–381). The Guilford Press.

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & Hill, P. L. (2014). What is conscientiousness and how can it be assessed? *Developmental Psychology*, *50*(5), 1315–1330. https://doi.org/10.1037/a0031109

Sabin, T. E., & Stafford, S. G. (n.d.). *Assessing the Need for Transformation of Response Variables*. 38.

Saucier, G., & Ostendorf, F. (n.d.). *Hierarchical Subcomponents of the Big Five Personality Factors: A Cross-Language Replication*. 15.

Schmidt-Atzert, L., & Amelang, M. (2012). *Psychologische Diagnostik* (5., vollständig überarbeitete und erweiterte Auflage). Springer.

Schnell, R. (1994). *Graphisch gestützte Datenanalyse:* DE GRUYTER. https://doi.org/10.1515/9783486787320

Schreiber, M., & Iller, M.-L. (2016). *Handbuch Fragebogen zur Erfassung der Persönlichkeit (IPIP-240)*. 60.

Simonsohn, U. (2017, September 18). [62] Two-lines: The First Valid Test of U-Shaped Relationships. Retrieved September 8, 2021, from Data Colada website: http://datacolada.org/62

Simonsohn, U. (2018). Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships With Quadratic Regressions. *Advances in Methods and Practices in Psychological Science*, *1*(4), 538–555. https://doi.org/10.1177/2515245918805755

Simonton, D. K. (n.d.). *Biographical Determinants of Achieved Eminence: A Multivariate Approach to the Cox Data*. 9.

Soetewey, A. (2020). *Outliers detection in R*. Stats and R. https://statsandr.com/blog/outliers-detection-in-r/

Stanek, K. C., & Ones, D. S. (2018). Taxonomies and Compendia of Cognitive Ability and Personality Constructs and Measures Relevant to Industrial, Work and Organizational Psychology. In D. Ones, N. Anderson, C. Viswesvaran, & H. Sinangil, *The SAGE Handbook of Industrial, Work and Organizational Psychology: Personnel Psychology and Employee Performance* (pp. 366–407). SAGE Publications Ltd. https://doi.org/10.4135/9781473914940.n14

Stanley, D. (2021). *apaTables: Create American Psychological Association (APA) Style Tables* (2.0.8) [Computer software]. https://CRAN.R-project.org/package=apaTables

Theiss, J. D., Hobbs, W. B., Giordano, P. J., & Brunson, O. M. (2014). Undergraduate Consent Form Reading in Relation to Conscientiousness, Procrastination, and the Point-of-Time Effect. *Journal of Empirical Research on Human Research Ethics*, *9*(3), 11–17. https://doi.org/10.1177/1556264614540593

Trapmann, S., Hell, B., Hirn, J.-O. W., & Schuler, H. (2007). Meta-Analysis of the Relationship Between the Big Five and Academic Success at University. *Zeitschrift Für Psychologie / Journal of Psychology*, *215*(2), 132–151. https://doi.org/10.1027/0044-3409.215.2.132

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., EISPACK authors, Heisterkamp, S., … R Core Team. (2021). nlme: Linear and Nonlinear Mixed Effects Models (Version 3.1-153). Retrieved from https://CRAN.R-project.org/package=nlme

Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554–568. https://doi.org/10.1016/j.chb.2015.01.070

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., Zemla, J., Freidank, M., Cai, J., & Protivinsky, T. (2021). *corrplot: Visualization of a Correlation Matrix* (0.90) [Computer software]. https://CRAN.R-project.org/package=corrplot

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & RStudio. (2021). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (3.3.5) [Computer software]. https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2021). *dplyr: A Grammar of Data Manipulation* (1.0.7) [Computer software]. https://CRAN.R-project.org/package=dplyr

Wickham, H., & RStudio. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations* (1.4.0) [Computer software]. https://CRAN.R-project.org/package=stringr

Wickham, H., & RStudio. (2021). *tidyr: Tidy Messy Data* (1.1.3) [Computer software]. https://CRAN.R-project.org/package=tidyr

Wilmot, M. P., & Ones, D. S. (2019). A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences*, *116*(46), 23004–23010. https://doi.org/10.1073/pnas.1908430116

Wood, S. (2021). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation* (1.8-36) [Computer software]. https://CRAN.R-project.org/package=mgcv

Zeileis, A., Lumley, T., Graham, N., & Koell, S. (2021). *sandwich: Robust Covariance Matrix Estimators* (3.0-1) [Computer software]. https://CRAN.R-project.org/package=sandwich

Appendix
A Careless Responders and Outliers

Table A1 Descriptive statistics about the excluded outliers. Cells with "/" within the column "Number of identified outliers by variable" indicate, that the respective variable was not included in the outlier analysis and therefore no participant was excluded based his/her value in the variable.

		Mean (Median)			SD (MAD)		
Variables	Number of identified outliers by variable	All (N = 40)	Only outlier (N = 21)	Outlier & CR (N = 19)	All (N = 40)	Only outlier (N = 21)	Outlier & CR (N = 19)
self-report conscientiousness	9	3.04 (3.04)	3.04 (3.12)	3.04 (2.92)	0.58 (0.56)	0.49 (0.49)	0.67 (0.37)
"objective" conscientiousness	/	0.61 (0.60)	0.66 (0.60)	0.56 (0.60)	0.21 (0.30)	0.16 (0.30)	0.25 (0.30)
average response time in sec	/	80.79 (73.87)	89.80 (95.25)	70.82 (63.83)	42.52 (55.49)	46.81 (60.06)	35.81 (36.24)
average response time (log-transformed)	17	4.23 (4.30)	4.32 (4.56)	4.13 (4.16)	0.61 (0.70)	0.65 (0.54)	0.56 (0.57)
variability in response time in sec	/	24.91 (8.87)	27.32 (21.52)	22.25 (6.77)	30.34 (9.44)	28.68 (26.74)	32.66 (6.03)
variability in response time (log-transformed)	22	2.46 (2.18)	2.61 (3.07)	2.28 (1.91)	1.29 (1.57)	1.31 (1.95)	1.28 (1.32)
average click-inaccuracy	0	5.39 (5.25)	5.93 (7.50)	4.49 (4.00)	3.47 (4.45)	3.64 (2.97)	3.26 (3.71)
variability in click-inaccuracy	5	1.89 (1.48)	1.66 (1.48)	2.15 (1.48)	1.70 (1.10)	1.57 (1.10)	1.85 (2.20)

Table A2 Distribution of careless responders in the raw average response time per survey page in the full sample (N = 194) containing all outliers and careless responders.

	Average Time per Page (not transformed) - Quartiles			
	Q0 – Q1	Q1 – Q2	Q2 – Q3	Q3 – Q4
failed any attention check	24	26	16	20
failed attention check 1	16	19	12	16
failed attention check 2	11	14	8	12
Failed both attention checks	3	7	4	8

Table A3 Distribution of careless responders in the raw average response time per survey page in the sample without outliers (N = 154) containing careless responders.

	Average Time per Page (not transformed) - Quartiles			
	Q0 – Q1	Q1 – Q2	Q2 – Q3	Q3 – Q4
failed any attention check	21	16	14	16
failed attention check 1	12	13	12	13
failed attention check 2	11	8	6	10
Failed both attention checks	2	5	4	7

B Additional Information

I Two-lines Analysis

In this thesis a so called curvilinear inverted U-shaped effect is hypothesized in the relationship between the time spent on a specific Item and the person's assumed indicator of conscientiousness. Before we get to description of the Two-Lines Analysis by Uri Simonsohn (2018) itself, we need first define some terms. A U-shaped or inverted U-shaped effect consists of a continuous and symmetric line with an extreme point, i.d. maximum or minimum, and a sign flip between the slope before and after this extreme point. From a mathematical perspective the slope on this extreme point is the derivative $f'(x) = 0$ of the function $f(x)$. This depiction accounts for quadratic, but also all other polynomials with an even exponent (e.g., x^4, x^6, x^8). In the non-mathematical literature, especially in economics and social sciences a U-shaped effect most often is called, when a sign flip in slopes happens. So, like Simonsohn (2018) we use the term "U-shape", if a sign flip is found (see, e.g., Cohen, Cohen, West, & Aiken, 2003, p. 576, in Simonsohn, 2018; Simonton, 1976).

Usually, U-shaped effects have been tested with a quadratic regression. A quadratic function is a continuous, symmetric line with an extreme value and a sign flip, like the depiction of U-effect in the section above. As noted above, our hypotheses just need the sign flip to be interpretable. Based on the linear regression in the form $y = b_0+b_1*x_1$, where x is the independent variable and y the criterium, the quadratic regression introduces a second predictor, which is in fact the interaction of the predictor with itself. The form is $y = b_0+b_1*x_1+b_2*(x_1)^2$. The quadratic regression assumes a quadratic relationship between predictor and criterium when the slope b_2 of the quadratic term is significant. Like in the linear regression, the t-test of each slope coefficient tests its deviation from 0, which also accounts for the quadratic coefficient. If the quadratic b_2 is significantly different to 0, the line is also assumed to be U-shaped. The problem of quadratic regression arises with the data, which indicate a non-linear relationship, which does not necessarily have to be U-shaped (e.g., logarithmic). While fitting the model line, it is quite often the case, that a "curved" quadratic line has less residuals, than a linear line. This makes the quadratic regression overly sensitive for curved relationships. Consequently, the quadratic regression often wrongly diagnoses a U-shaped relationship, even if the true relationship is somehow different (e.g., logarithmic). This over-sensitivity results in a high false-positive rate, which Simonsohn (2017) also showed in a simulation analysis. On the other hand, higher degree polynomials with even exponents (e.g., x^8) are often not detected, as those differ strongly from the quadratic term x^2. To sum it up: quadratic regression has a bad test accuracy with higher degree polynomial or logarithmic relationships and therefore has a high error rate. Other problems from the quadratic regression are, that the calculated breakpoint might be above or below all measured values of the predictor. Although the quadratic coefficient might be significant, the model can not be interpreted as only one side of the sign flip is seen. Another problem could

potentially arise when the predictor is not centered. The quadratic model could also assume negative values as an input, which in our case are not possible, as we use a Likert-Scale for the self-report conscientiousness measure.

Based on those limitations Simonsohn (2018) proposed a different approach of diagnosing a curvilinear U-shaped relationship: the two-lines test, which is basically an interrupted regression. The idea behind this approach relies on our definition of U-shape. As it solely relies on a sign flip up to a specific breakpoint, we just need a significant linear regression up to this breakpoint and a second significant linear regression with an opposite sign starting at this breakpoint (see figure BI.1). For the sake of completeness, the interrupted regression has the formulation:

$$(1)\ y = a + bx_{low} + cx_{high} + d * high + ZB_Z \begin{cases} x_{low} = x - x_C, x_{high} = 0, high = 0, if\ x < x_C \\ x_{low} = 0, x_{high} = x - x_C, high = 1, if\ x \geq x_C \end{cases}$$

Where Z is the (optional) covariate matrix and B_Z its matrix of coefficients. The important specifications of formula (1) are: If $x < x_C$, $x_{low} = x - x_C$, $x_{high} = 0$ and high = 0. When $x \geq x_C$, $x_{low} = 0$, high = 1 and $x_{high} = x - x_C$.

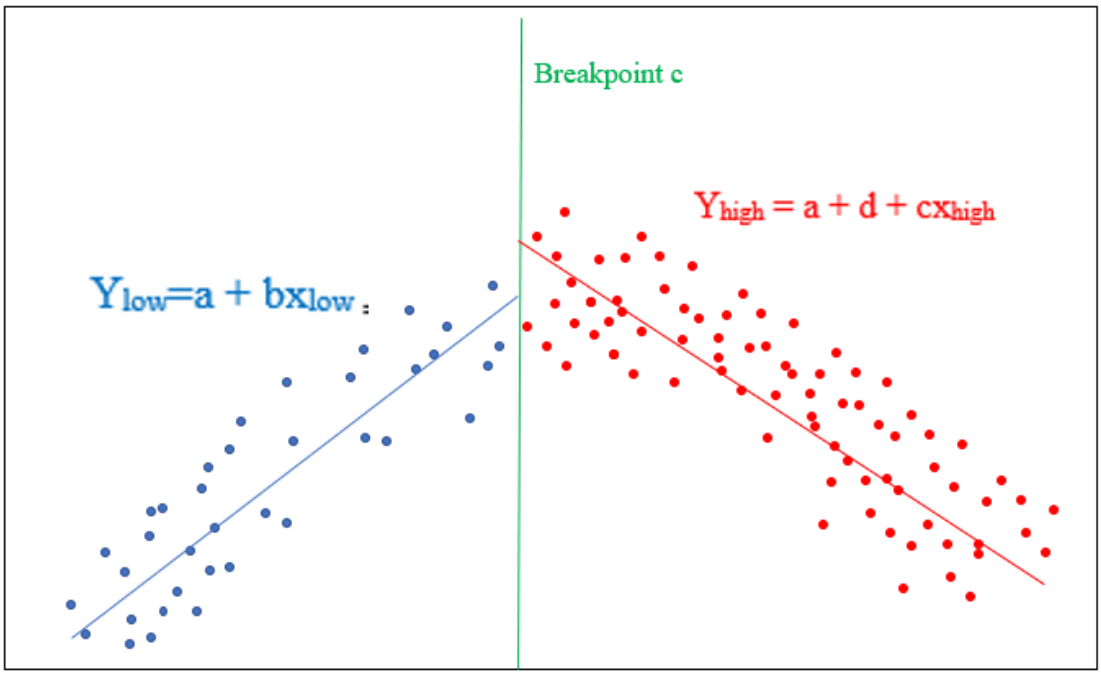

Figure BI.1 An exemplary representation of the interrupted regression in formula (1).

To break this formula down as simple as possible, we ignore the optional covariate matrix and its coefficients ZB_Z. For the lower rectilinear regression line (left of the breakpoint) the formula is $Y_{low}=a + bx_{low}$, which is identical to the "normal" simple regression (or correlation). For visualization see figure BI.1. For the right-hand side, the newly arranged formula is $Y_{high} = a + d + cx_{high}$. The binary variable high, which indicates the "high" group contributes to the intercept of the rectilinear regression for the high values. It should be mentioned that the d coefficient can be set to 0, which

would result in a so-called segmented regression, where the regression line after the breakpoint does not have its "own intercept" (Muggeo, 2003).

Originally the idea was, to determine the breakpoint by calculating a quadratic regression and use its extreme point as a breakpoint. Simonsohn (2018) found a better approach with higher statistical power to determine the U-shaped effect, which he called the "Robin Hood" algorithm. Basically, the statistical power for finding a significant regression line rises, if either the slope is steep or when there are many observations (i.e., small standard error). To raise the overall power, the Robin Hood procedure calculates both regression lines from a starting breakpoint and then "takes" observations from the stronger regression and "gives" it to the weaker regression. Consequently, the (new) breakpoint is set, that the originally weaker regression gets a boost in statistical power at the cost of the stronger regression. For further information see Simonsohn (2018) or his blog Data Colada [62] (Simonsohn, 2017).

II Outlier Analysis

Regarding response latencies, research suggests different approaches to detect outliers. None of which are perfect used alone (Höhne & Schlosser, 2018). Some approaches excluded all cases, which are two SD above or below the arithmetic mean (Mayerl, 2013). Schnell (1994) proposed 1.5 times the IQR as threshold. Hoaglin, Mosteller & Tukey, 2000 in Höhne & Schlosser, 2018) used 3 times the upper and lower quartile range above or below the median to exclude outliers. Lenzner, Kaczmirek, and Lenzner (2009) excluded the lowest and highest 1% of the cases. Soetewey (2020) proposed the Hampel filter, which excludes cases 3 times the MAD above and below the median.

Alas, there is no one-fits-all solution for outlier detection. As Mayerl's (2013) approach is vulnerable to outliers itself by relying on the mean and SD and Lenzner et al.'s (2009) approach seems too liberal, neither of those is used. The Hampel filter (Soetewey, 2020) as well as the approach by Schnell (1994) are relatively simple approaches, where a symmetric cut-off threshold around the median is used. In contrast, the approach proposed by Hoaglin et al. (2000 in Höhn, 2018) use non-symmetric thresholds around the median. They use the quartile range, which is either the 3rd quartile minus median for the upper threshold, or the median minus 1st quartile for the lower threshold. The advantage of this approach is, that symmetrically distributed data will have the same quartile ranges for both sides, while skewed data will have adjusted thresholds. As response latencies are right-skewed without transformation (Fazzio, 1990 in Höhne & Schlosser, 2018) this more flexible approach seems reasonable, as even the log-transformed median time per page is not normally distributed yet. When assuming a normal distribution, this approach would be equal to using an absolute z score of z = 2.022 (97.84%) as a cut-off threshold, which is quite similar to Mayerl's (2013)

criterium. Consequently, approximately 4.32% cases get excluded as outliers. As we expect careless responders to have a tendency of showing outlier behavior, we would like to keep the careless responders in our data and therefore having a relatively liberal cut-off threshold. We will use two data sets for our analyses: one with careless responders and one without them.