Improving User Engagement to Reduce Dropout Rates in Long Web Surveys

Exploring the Effectiveness of Achievement Primes Amongst Intrinsically and Extrinsically Motivated Respondents

Technische Universiteit Delft

T. van Tussenbroek

TU Delft

Improving User Engagement to Reduce Dropout Rates in Long Web Surveys

Exploring the Effectiveness of Achievement Primes Amongst Intrinsically and Extrinsically Motivated Respondents

by

T. van Tussenbroek

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday July 27, 2023 at 2:00 PM.

Student number: 4534794
Project duration: December 20, 2022 – July 27, 2023
Thesis committee: Dr. U. Gadiraju, TU Delft, supervisor
Dr. F. Broz, TU Delft

TU Delft

Preface

The cover photo of this thesis[1] is an artistic representation of the theory applied in this work. It features a flight of stairs, symbolising the journey of progressing through a survey. Like climbing stairs, surveys can become more challenging as you proceed. The exit on the stair landing signifies the temptation to drop out, especially when the survey or stairway is lengthy.

At the point of considering an exit, a motivational quote appears, much like the passive achievement primes employed in this study. The quote “All ideas grow out of other ideas” by Anish Kapoor reflects how this research is built on knowledge gaps identified in previous work. If you persist in ascending the stairs, the quote disappears from view, hinting that the stimulus may lose its effect until another quote appears at the next landing. Interestingly, the individual in the photo opted not to continue up the stairs, a playful nod to this study’s findings. Moreover, the picture draws to me personally as the black and white palette of the photo pays tribute to the preferred photographic style of Petra and Peter, my aunt and uncle.

I deeply thank my thesis advisor, Ujwal Gadiraju, and my co-supervisor, Garrett Allen. Your input have shaped this work, making me proud of the result. I hope you share this pride. A big thanks goes to the Business Intelligence team at KPMG, who welcomed me as their thesis intern and gave me a glimpse of what it is like to work in a professional environment.

This thesis marks the close of my academic journey and the start of my professional one. My family and friends have been a constant source of joy and support over the years. My mother, Corine, sister, Marleen, and father, Aart, in particular, have witnessed my transformation from a first-year student to an engineer, and our bond has only strengthened since we lived under the same roof. My friends deserve a shout-out for their humour and companionship, especially Femke and Cristel, who were always there for coffee breaks on campus during the writing of this thesis.

Looking ahead, I’m eager to explore the opportunities that await me, but more importantly, to discover new ways to enrich my life with happiness.

As Audrey Hepburn once said, “The most important thing is to enjoy your life — to be happy — it’s all that matters.”

T. van Tussenbroek
Delft, July 2023

[1] Photo by CJ Dayrit on Unsplash

Contents

List of Figures

List of Tables

Abstract

Web surveys have increasingly been used to collect data from respondents over the years. They offer several advantages compared to other methods of obtaining data. Researchers benefit from a broad demographic representation to make generalized conclusions, and satisfaction surveys allow employees to explain shortcomings or improvements anonymously. Both examples demand comprehensive information, thereby requiring a lengthy survey. However, dropout increases with the length of a survey, which is a big problem on web surveys as it decreases the statistical significance of the results. Proposed solutions, such as reducing the number of questions or rewarding respondents with an incentive, may not always be feasible due to the preciseness of information required or limited financial capabilities.

Achievement primes have been shown to reduce dropout on short surveys targeting extrinsically motivated respondents without additional costs or the need to reduce survey length. As repeated exposure to primes reinforces the stimuli, long surveys may also benefit from achievement primes. In this study, respondents are exposed to a questionnaire of more than 15 minutes on health whilst working behind a computer containing either no prime, passive achievement primes, or active achievement primes. Besides extrinsically motivated respondents, recruited via the crowdworking platform Prolific, intrinsically motivated respondents are also targeted in this study, recruited via snowball sampling.

Through a 2×3 factorial design, we discovered no statistical difference in dropout, perceived workload, and user engagement across the three questionnaire variants when evaluating intrinsically ($N = 88$) and extrinsically motivated respondents ($N = 140$) individually. By comparing intrinsically with extrinsically motivated respondents, we discovered extrinsically motivated respondents were more engaged and dropped out less.

1

Introduction

Web surveys are used to collect data from a pool of respondents. Compared to other data collection forms, such as face-to-face interviews or physical (paper) surveys, web surveys offer several advantages. These include the ability to attract a large number of participants, independent of geographical location, as well as the convenience of easy data management and analysis because of digitally stored answers (Menon & Muraleedharan, 2020). The use of web surveys in research has increased over the years as a means to gather data (Wu et al., 2022).

The quality of the survey answers depends both on the survey design (Sanchez, 1992) and the respondents' motivation to participate (Vriesema & Gehlbach, 2021). Survey design encompasses various elements, including structure, language, user interface, and other design decisions. On the other hand, respondents' motivation can stem from intrinsic or extrinsic factors (Veit et al., 2011). An example of intrinsic motivation is the desire to contribute to research, whereas an example of extrinsic motivation is the will to fill out a survey in exchange for compensation.

Dropout is a predominant problem in this area (Galesic, 2006; Nestler et al., 2015). Dropouts are respondents who prematurely leave a survey without submitting their answers (Standard Definitions, 2016). By not submitting answers, the resulting data is incomplete, which leads to a decrease in statistical power (O'neil et al., 2003). Dropout is especially a problem on long web surveys as dropout increases with the length of a survey (Galesic, 2006; Hoerger, 2010; Liu & Wronski, 2018b; Tijdens, 2014). Reducing the survey length may not always be a solution to reducing dropout. For example, job satisfaction surveys or surveys for research often include multiple discrete questionnaires, each designed to measure specific effects such as user engagement (O'Brien et al., 2018) or perceived workload (S. Hart, 1986). While these questionnaires may be short, the combination can extend the overall survey length. Discarding these questionnaires is not an option, as it would result in essential effects remaining undetected.

Providing incentives has been found to effectively reduce survey dropout (Frick et al., 2001; O'neil et al., 2003; Ziegenfuss et al., 2013). An incentive is a form of compensation for responding to a web survey. Whereas incentives that do not involve monetary costs, such as informational rewards, lead to increased dropout (Kalleitner et al., 2022), incentives in the form of money (Frick et al., 2001; O'neil et al., 2003) or material gifts (Ziegenfuss et al., 2013) have demonstrated a reduction in dropout. Despite reducing dropout, providing respondents with monetary incentives may not always be feasible, as organisations have limited financial resources.

Achievement priming has proven to reduce dropouts on web surveys without requiring a reduction in survey length or cost increase (Gadiraju & Dietze, 2017). Primes are stimuli in the environment which cause the user to change their thoughts or perform an action without them being aware of it (J. Bargh & Chartrand, 2000). A distinction can be made between active primes, which require interaction, and passive primes, which may serve as decoration. An example of a passive achievement prime is a picture of a marathon runner crossing the finish line placed amongst the questions with no means of interaction. An illustration of an active achievement prime would be the same picture, accompanied by a question for the respondent to answer, "Which emotion does this picture evoke for you?".

Thus far, achievement priming has been applied to *short surveys* targeting *extrinsically motivated* respondents (Gadiraju & Dietze, 2017). However, it has the potential to benefit long surveys as well,

as repeated exposure to stimuli can reinstate the priming effect (Hermans et al., 2001). Furthermore, similar to extrinsically motivated respondents, intrinsically motivated respondents may also be influenced by achievement primes. Nevertheless, the effect of external stimuli on intrinsically motivated respondents could be lower than on extrinsically motivated respondents, as only the latter group participate because of external factors.

This research aims to close the knowledge gap of the effect of the different forms of achievement primes in combination with the two types of respondent groups on dropout, perceived workload, and user engagement. The following Research Questions (**RQ**) are answered by means of the corresponding Hypotheses (**H**):

RQ1: How does achievement priming affect the dropout rate on long surveys?

- **H1A:** There is less dropout when respondents are exposed to active achievement priming compared to no priming.
- **H1B:** There is less dropout when respondents are exposed to passive achievement priming.
- **H1C:** Active achievement primes are more effective than passive achievement primes at reducing dropout.

RQ2: How does achievement priming influence respondents with different motivations to participate?

- **H2:** People are less affected by primes when they participate based on intrinsic motivation than when they participate because of extrinsic motivation.

To answer these questions, we design a 3×2 between-subjects factorial user study, where we consider three questionnaire variations containing either no primes, passive achievement primes, or active achievement primes alongside intrinsically and extrinsically motivated respondent groups. The participants shall be recruited through snowball sampling and crowdworking platforms, as these methods are suitable for targeting intrinsically and extrinsically motivated participants. Respondents will be asked to fill in a long web survey of roughly 20 minutes which asks about their health while working behind a computer. Participants who were shown the passive or active versions will be primed by inspirational quotes.

Thesis Contribution and Research Outline

The contributions made by this thesis are threefold:

- An extensive evaluation of best practices for survey design based on previous research.
- A 3×2 factorial user study exploring the effect of the different forms of achievement primes in combination with the two types of respondent groups on dropout, perceived workload, and user engagement.
- Publicly accessible (anonymised) data and code produced for this thesis. [1]

We find no significant difference improvement in dropout between the questionnaire variants for either extrinsically or intrinsically motivated respondent group. We do find a statistically significant difference in dropout when comparing intrinsically to extrinsically motivated respondents, with the latter having a more favourable outcome. Moreover, extrinsically motivated respondents report higher engagement compared to intrinsically motivated respondents.

The remainder of this thesis is structured as follows. Best practices for survey design and related literature is summarised and discussed in Chapter 2. After this, the experimental setup of the main research as well as the curation of the achievement primes is discussed in Chapter 3. Chapter 4 reports the results of the final study. These results and the study's limitations are then discussed in Chapter 5. The conclusions of this study are provided in Chapter 6.

[1] The data and code can be found at: https://osf.io/pxfba/?view_only=77cda642e7a84a9cba63b1714c326fdb

2

Background and Related Literature

This chapter contains background information on web survey design. First, a distinction between respondent groups and recommendations for designing web surveys is presented. These recommendations are based on functional and non-functional requirements derived from previous research. Second, the problem of dropout on long surveys and existing solutions to this problem are discussed.

2.1. Best Practices for Survey Design

This section combines individual findings into a set of best practices, which should help to acquire improved data from web surveys. Since the quality of survey data depends on the design of the survey (Sanchez, 1992), researchers have investigated how to improve upon the design.

Another factor affecting the quality of survey data is the respondent's motivation to participate (Vriesema & Gehlbach, 2021), and it is therefore important to differentiate the groups of participants based on their initial motivation. As an example, someone who is asked to fill out a survey by their employer may want to rush through the questions as they may face deadlines that have their priority. Oppositely, an enthusiast filling out a survey might use their spare time to carefully construct answers.

Veit et al. (2011) identified five reasons why someone may participate in a survey. The groups of participants are distinguished based on:

- Intrinsic Motivation
  - Enjoyment Based Motivation (EB) *(e.g. boredom, fun)*
  - Community Based Motivation (CB) *(e.g. addition to a community or to research)*
- Extrinsic Motivation
  - Immediate Payoff (IP) *(primary or secondary source of income)*
  - Delayed Payoff (DP) *(skill improvement, reputation)*
  - Social Motivation (SM) *(forced, network)*

These recommendations are categorised in the following survey design features: trust and contact, incentive and engagement, questions, and progress. A tabular representation of the requirements, the participant group, duration, number of respondents, and more details can be found in appendix A.

The literature has been obtained by means of searching on the academic research database Scopus, internet search engines Google and Google Scholar, as well as the databases of the conferences Human Factors in Computing Systems (CHI), Conference On Computer-Supported Cooperative Work (CSCW), Intelligent User Interfaces (IUI), User Interface Software and Technology (UIST), and Conference on Human Computation and Crowdsourcing (HCOMP). The following search terms were used to find the literature: `("web survey" OR "online survey" OR "e-survey" OR crowdwork OR crowdsourcing OR microtasks OR "questionnaire") AND (breakoff OR dropout OR non-response OR attrition OR "task abandonment" OR retainment OR burden) AND (engagement OR motivation OR curiosity OR`

`interest OR prime OR performance OR "goal setting" OR incentive)` . Research cited by literature found on these databases have also been evaluated. Literature was not used if it reported no or insignificant results, are longitudinal surveys, establish definitions, or are otherwise not fit to derive functional and non-functional requirements from (such as defining a model, explaining best practices without scientific support, or are a research draft).

2.1.1. Trust and Contact

By accurately and honestly reporting information in surveys, a sense of trust can be established with the respondents. First, before publicizing the survey, there should be approval of the research by an official ethics committee (Lavidas et al., 2022). This approval should be communicated when contacting possible respondents and when done via e-mail, it is important to disclose how the surveyor acquired the respondent's e-mail address (Lavidas et al., 2022). If the respondent chooses to participate, the respondent's e-mail address should not be asked for in the survey (Lavidas et al., 2022) as the survey should ensure that the anonymity of the respondent is maintained at all times (Wen & Fang, 2012). Asking for an e-mail address creates the possibility to link a response to an individual, compromising their anonymity. Furthermore, the survey provider (or platform on which the survey is hosted) should have a positive reputation in the eyes of the respondent (Fang et al., 2012), as should any potential sponsorship affiliated with the survey (Allen & Roberts, 2016; Fang et al., 2012; Lavidas et al., 2022; Pan et al., 2014; Wen & Fang, 2012).

A surveyor might want to directly contact respondents via other media than the survey itself for various reasons. They might want to attract new potential respondents, have dropouts return to the survey, or want to have previous respondents return for future surveys.

The surveyor may send a pre-notification to respondents, which would state that they will be asked to participate in a survey at a later time. These pre-notifications may increase response rate (Bosnjak et al., 2008; Cook et al., 2000; Keusch, 2012), and should preferably be sent out via SMS instead of via e-mail (Bosnjak et al., 2008).

The service via which messages are sent to respondents should be carefully chosen for the actual invitation or reminders. There are mixed conclusions to the question whether invitations should be sent via e-mail (Bosnjak et al., 2008), or that they should be sent out via physical mail or via phone surveys instead in order to increase response rate (Saleh & Bista, 2017), as the effect differs per motivation group or demographic. For surveys which are meant to be filled out on phones however, it is clear that invitations are best sent via SMS (Couper & Mavletova, 2014). Moreover, when surveys are filled out on phones, they should be administered via an application rather than via a website in order to decrease dropout (Roberts et al., 2022).

There are also mixed conclusions on whether the text in the invitation should appeal to the egotistic need for approval (Pedersen & Nielsen, 2016) or that the survey should include altruistic messaging (Hall et al., 2019; Lavidas et al., 2022) or create a feeling of moral obligation (Wen & Fang, 2012). What is unanimously agreed upon, however, is that all text should be personalized (Cook et al., 2000; Kaczmirek, 2008; Sánchez-Fernández et al., 2012). For example, the user should be saluted personally (Heerwegh & Loosveldt, 2006), preferably by first name only instead of with their full name (Joinson & Reips, 2007; Sauermann & Roach, 2013). The invitation itself should contain an authoritative, trustworthy, and consistent solicitation message (Pan et al., 2014) and should contain information on the importance of the outcome of the survey (Kaczmirek, 2008). The respondents should also be promised to be informed about said outcome (Lavidas et al., 2022).

Regardless of the recipient's gender, the sender should also be someone who is higher in power than the recipient (Joinson & Reips, 2007). A male recipient should be contacted by a female surveyor in order to increase response rate (Keusch, 2012). There is no research which has shown that the same effect holds the other way around, where females would be addressed by males. To our knowledge, there is no research on the effect of the recipient's gender on the response rate of non-binary genders. When the recipient and surveyor do not personally know each other, the invitation should also include a picture of the surveyor as to create the feeling that there is a personal connection (Guéguen & Jacob, 2002). Moreover, the surveyor's name should be of a similar cultural background as the recipient's background (Guéguen & Jacob, 2002; Pan et al., 2014).

The ease of starting a survey also has impact on the response rate, for example in the case of a survey which requires the participant to log in. When a respondent receives a personal invitation via an electronic platform (e.g. e-mail or SMS), the link to the survey may contain a unique query

parameter which associates the respondent's answers to that person. This removes the need for the respondent to login and increases the likelihood of the respondent wanting to participate. Automated entry of credentials does, however, decrease the honesty of reporting compared when the respondent is required to manually enter them (Crawford & Couper Mark J Lamias, 2001).

Reminders have proven to be efficient in reducing the number of non-responses. It is best to have few days in between the invitation and the reminder, as a follow-up reminder which is sent out two days after the invitation results in a higher response rate than one which is sent out after five days (Crawford & Couper Mark J Lamias, 2001). Researchers have not settled on the number of reminders, however, as the recommendations fluctuate between one (Christensen et al., 2015) and as many reminders as possible (Sauermann & Roach, 2013). The optimal frequency of reminders also depends on whether the time in which the survey can be conducted is limited or not. When time is limited, the frequency should be high, whereas a lower frequency may be suitable in the absence of time constraints (Sánchez-Fernández et al., 2012). In any case, when opting for multiple reminders, each reminder should use different wording and the user should always be presented with an option to opt-out of receiving reminders (Sauermann & Roach, 2013).

After completing a questionnaire, it might be worthwhile to contact the same respondents for new studies. Contacting panelists who took part in previous studies results in a higher completion rate compared to when new participants are recruited (Goritz, 2014). One way of doing so would be to include an option at the end of the previous survey to be kept up to date for future research. When providing the user with such an option, it is best to have the default option be to opt in (Jin, 2011). When recontacting previous participants, however, there should be a long time lag between the previous study and the new study, and the length of the new survey should be kept shorter than the previous study (Goritz, 2014).

2.1.2. Incentive and Engagement

One of the most predominant reasons for people to participate in web surveys is that participants are offered an incentive (Hall et al., 2019; Howell, 2020; Ziegenfuss et al., 2013). When given the choice between a longer survey accompanied by a monetary or gift reward or no reward and a shorter survey, respondents most often chose the former (Ziegenfuss et al., 2013). Such a reward can be granted to all respondents in a fixed fashion or to a select group of respondents in the form of a lottery. Respondents are more likely to participate when they take part in a lottery than when the money is donated to a good cause (Pedersen & Nielsen, 2016). The incentive should be offered after completion instead of up-front (Manfreda & Vehovar, 2002), further indicated by pre-paid incentives resulting in a lower response rate than the promise of a lottery (Sauermann & Roach, 2013). When participants may partake in a lottery, this should be mentioned at the beginning of the survey (Frick et al., 2001).

When the survey is conducted on a crowdworking platform, where the crowdworker and the surveyor have an employee-employer relation, the hourly pay should be fair with respect to the target demographic (L. Han et al., 2019). A survey with monetary incentive should never allow for multiple submissions (Howell, 2020).

Respondents should continuously be engaged in order to improve their performance or to prevent dropout (Manfreda & Vehovar, 2002). In general, this can be done by having the survey be fun (Howell, 2020), trigger curiosity (Law et al., 2016), or be subjectively interesting (Galesic, 2006).

By means of historical data and user behaviour, it can be predicted when a user may want to drop out. At these points, it may be useful to dynamically introduce the user to subjectively enjoyable questions. This might relieve the user from the burden they experience and thereby prevent them from dropping out (Kobren et al., 2015). The questions introduced when the attention span is reduced can also be dummy events (Elmalech et al., 2016), which are brief, enjoyable activities that are unrelated to the survey itself.

These micro-diversions can come in the form of a game (Dai et al., 2015; Rzeszotarski et al., 2013), a short story (Dai et al., 2015), a comic (Rzeszotarski et al., 2013), or might even let the respondents gamble to obtain a higher monetary reward (Rzeszotarski et al., 2013).

Gamification may also be implemented in the survey in order to increase engagement. By including a leaderboard, levels, badges, feedback alerts, bonus points, or treasure points, the respondent may feel the need to compete against other respondents (Feyisetan et al., 2015).

The question's style may also be of influence to the user's feeling of engagement, such as when the respondent feels as though they are actively communicating with a person in a conversational user

interface (Qiu et al., 2020). In such a conversational style, the answers are posed in the form of text messages, and the user may respond back in this style as well. Contrary to normal conversations, the respondent should not be forced to react in the form of an open question. Other types of answers, such as multiple-choice, single-choice, or sliders may also be used in this conversational style. When using the conversational style, the conversational agent (asking the questions) and the respondent should both have an avatar (Haque et al., 2022). Moreover, the questions should be asked in an informal manner (Kim et al., 2019).

2.1.3. Questions

Questions themselves should cause little subjective burden to the user (Galesic, 2006; L. Han et al., 2019). This applies to the question types (open questions, sliders, multiple choice, etc.) as well as to the text used to ask the question.

Open-ended questions should typically be avoided (Crawford & Couper Mark J Lamias, 2001; Galesic, 2006; Lavidas et al., 2022; Liu & Wronski, 2018b; Manfreda & Vehovar, 2002; Peytchev, 2009). Even though they allow the user to answer in their own words, open-ended questions require constant attention and energy. Nonetheless, they may be necessary depending on the survey's purpose. When open-ended questions are incorporated into the survey, the area where the answer can be typed should be limited, as this lowers the respondent's expected burden (Zuell et al., 2015). An experiment in which the respondents were asked to draw their answer instead of writing in a text area also caused an increase in burden and therefore did not serve as a proper substitute (Bosch et al., 2022).

Other types of questions that should be avoided are attitudinal (measuring opinions, emotions, perceptions or judgments) (Peytchev, 2009), numeric (Peytchev, 2009), or matrix questions (Kaczmirek, 2008). Oppositely, types of questions which cause little burden are single- or multiple-choice questions, as long as the number of options in the list is not too long (Tijdens, 2014). The options in single- or multiple-choice questions should consist of text, as logo's representing the text causes an increase in burden (Manfreda et al., 2002).

Survey questions should also be well-defined and consistent (Kaczmirek, 2008; Keusch, 2012). For questions where confusion may occur due to the difficult nature of the question, notes and instructions should be included (Kaczmirek, 2008). To steer the user in the right direction when there is a large body of text, portions of the text may be highlighted. As long as the selected highlighted text is of good quality, this will reduce the required effort and thereby reduce the respondent's decision time (Ramírez et al., 2019).

The order in which the questions are asked should also be considered, as should the structure (such as pagination). Demographic information, for example, should be acquired at the beginning rather than at the end of the questionnaire (Howell, 2020). Respondents regard demographic information as non-trivial. As the respondent's perceived level of burden is at its height at the end of a survey, when asked at the end, participants may be inclined to leave the survey when they are asked to fill out demographic information. Furthermore, the questions at the start should be relatively simple (Peytchev, 2009) and interest-related (Shropshire et al., 2009) to ease the respondent into more demanding tasks and introduce them to the subject. The first questions should preferably be multiple-choice rather than open-ended questions (Liu & Wronski, 2018b).

The remainder of the survey should alternate burdensome questions with more enjoyable questions to prevent the accumulation of burden at one specific point (Howell, 2020). Specifically on mobile devices, all questions should be placed on one page rather than spread over multiple pages. This enables the user to scroll indefinitely until the end of the survey, decreasing the dropout rate (Couper & Mavletova, 2014). On desktop, there appears to be no difference between the dropout rate of scrolling or paging surveys (Manfreda et al., 2002; Peytchev et al., 2006).

2.1.4. Progress

Respondents prefer a survey not to be long (Galesic, 2006; Liu & Wronski, 2018b; Revilla & Ochoa, 2017) as the burden respondents perceive increases the longer they have to spend on a task. The progress indicator helps the respondent to decide whether they want to continue or drop out by comparing the perceived burden thus far with the expected burden in the future (Howell, 2020).

The speed at which the progress bar moves can be calculated in various ways, as the speed at which a respondent answers questions may vary from person to person and from question to question.

If the progress bar does not report accurate or promising feedback, the respondent is likelier to drop out if the progress indicator is shown than when it is hidden (Kaczmirek, 2008).

However, even if the speed at which the progress bar advances accurately reflects the speed at which questions are answered, filter questions may pose a problem. Depending on the answer to filter questions, a selection of questions may or may not be shown to the respondent, which greatly impacts the respondent's progress. There has been a great effort to dynamically calculate the progress when filter questions are included (Kaczmirek, 2008), for which observing historical data may be beneficial (Prinz et al., 2020).

The dropout rate can be reduced by artificially accelerating the speed at which the progress bar moves at the start of the survey and slowing down the speed towards the end (F. Conrad et al., 2005; F. G. Conrad et al., 2010; Kaczmirek, 2008; Maliković & Tončić, 2021; Peytchev, 2009; Villar et al., 2013). This makes the respondent feel optimistic about the duration of the survey. Artificially accelerating the speed at which the progress is reported does come with serious ethical considerations, however. To have the progress appear to be moving faster whilst not artificially modifying the speed, a ribbon effect moving opposite the direction of the progress may be introduced. This creates the illusion of acceleration (Harisson et al., 2010) as opposed to artificially manipulating the actual progress.

There are some contradicting findings on when, where, or even whether the progress bar should be placed on a questionnaire. There are claims that progress indicators should never be placed on surveys which take 20 minutes or longer to complete (Matzat et al., 2009), and they should also not be shown on surveys which contain many open-ended questions (Crawford & Couper Mark J Lamias, 2001) or should never be shown at all (Liu & Wronski, 2018b). Others claim that on shorter surveys, the progress indicator should always be shown (Yan et al., 2011a), that it should only be shown early in the survey (Matzat et al., 2009), or that it should be shown intermittently (F. Conrad et al., 2005; Vicente & Reis, 2010). However, when the progress indicator is present on all pages, it should be placed at the top of the web page rather than at the bottom (Liu & Wronski, 2018b).

By following these best practices in terms of trust and contact, incentive and engagement, questions, and progress, the quality of the results of any general survey should be improved.

2.2. Dropout on Long Surveys

This section discusses the issue of dropout, specifically on long surveys. First, we define what constitutes a long survey. Then, we explain why dropout is an issue specifically for long surveys and provide current solutions to this problem. Dropouts are recognised by the point at which the respondent leaves the web survey prematurely and therefore does not submit the data (Standard Definitions, 2016).

2.2.1. Definition of Long Surveys

One way of expressing the measurement of survey length is the time required to complete the survey. However, the suitability of time as a measurement has been a subject of debate. Rolstad et al. (2011) argue that potential respondents decide whether or not to start filling out a survey based on their perception of the expected burden. They claim time alone cannot serve as an indication of the questionnaire's level of burden and may therefore not be suitable. Several studies have countered this argument by demonstrating that response rate is negatively correlated with survey length (Deutskens et al., 2004; Galesic & Bosnjak, 2009; Marcus et al., 2007; Trouteaud, 2004; Walston et al., n.d.), which shows that time is indeed an indicator of expected burden. Based on these findings, we use time as the unit of measurement for the length of a survey.

Revilla and Ochoa (2017) asked respondents their opinion on the ideal and maximum length for a survey, which they concluded to be 10 and 20 minutes, respectively. However, their research was regarding *written* surveys and not *online* surveys. Revilla and Höhne (2020) followed up this research by asking respondents what the ideal and maximum length should be when focused on *web* surveys. They concluded that web surveys' ideal and maximum length should be 10-15 and 20-28 minutes, respectively. The research outcome is based on subjective measurements, and therefore the actual ideal length of a survey might differ from the subjective ideal length. Respondents might specify that the ideal length of a survey is 10 minutes, but when they are actually filling out a survey themselves, they might think this length is too long.

Informed by the aforementioned research, we define long web surveys to be those that exceed 15 minutes, as by that point, a web survey would exceed the ideal survey length.

2.2.2. Reducing Survey Length

The length of the survey has a significant impact on the dropout rate. Hoerger (2010) discovered a linear relationship between the number of questions and the dropout rate, with 2% of respondents dropping out per 100 questions. Liu and Wronski (2018b) examined 25,080 surveys on SurveyMonkey Audience[1] and discovered that an increase in the number of pages or an increase in the number of questions results in a higher dropout rate. Tijdens (2014) researched the behaviour of (self-)employed and unemployed filling out a search tree and discovered the dropout rate is affected by the size of the items in the survey. Galesic (2006) investigated the effects of interest and burden on dropout rates in web surveys of different lengths (10, 20, and 30 minutes). The data revealed a peak in dropout at the beginning of all three surveys, after which a consistent dropout rate was observed until the end.

As survey length is positively correlated with the dropout rate, it may be compelling to limit the survey length. Indeed, several online blogs advocate for shorter surveys as a strategy to increase response rate or to reduce dropout ("10 Tips to Shorten Your Survey Length", n.d.; "5 Ways Your Survey Tool Can Help to Shorten Your Questionnaire", 2016; Antarika, 2021; Chinn, 2023; Chudoba, n.d.; DeFranzo, n.d.; Deshpande, n.d.; Henning, 2023; "Keep your Survey Length to 20 Minutes or Less", n.d.; Nielsen, 2004; Quinn, 2010; "Six Tips to Shorten Your Survey", 2013; "Survey Length Best Practices: Are Shorter Surveys Better?", n.d.; Vanette, n.d.; Wigmore, 2022). Nonetheless, these posts often lack empirical evidence to support their claims. Moreover, there is a unified perspective on the optimal survey duration, with suggestions varying from as brief as 4 minutes to as long as 13 minutes.

Several post-questionnaires, designed to capture respondents' experiences, have been condensed to decrease the completion time. For instance, the User Engagement Scale (UES), originally taking 15 minutes to complete, was shortened to a 5-minute version called the User Engagement Scale - Short Form (UES-SF) while still capturing the same concepts (O'Brien et al., 2018). Similarly, the original NASA's TLX (S. Hart, 1986) questionnaire requires respondents to compare six dimensions pairwise. Removing this comparison stage does not significantly impact the outcome (Byers et al., 1989; Nygren, 1991). Yet, the completion time is dramatically reduced from up to 60 minutes ("NASA Task Load Index (NASA TLX)", n.d.) to 3 minutes (Bell et al., 2022). The System Usability Scale (SUS) is another example where the creators promote the short completion time, with the associated article being titled "SUS: A quick and dirty usability scale" (Brooke, 1995).

However, these post-questionnaires can only capture a singular effect and cannot be reduced further. Evaluating multiple effects in studies, such as in research or job satisfaction surveys, is often necessary. Having multiple post-questionnaires can still result in a lengthier survey. Therefore, it is not always possible to decrease the survey length without compromising the data required.

2.2.3. Incentives

Incentives also play a role in the dropout rate on long surveys, but they are limited in their effect. Incentives may take the form of money, information, or products, and can be provided to a select group of respondents in the form of a lottery or be granted to all respondents.

Kalleitner et al. (2022) informed participants they would be rewarded with an informational video after participating. However, the increase in bandwidth use, the disinterest in the topic at hand, and the burden of additional information decreased the response rate. Frick et al. (2001) and O'neil et al. (2003) demonstrate that using a lottery as compensation results in half the number of dropouts compared to receiving no compensation. Ziegenfuss et al. (2013) provided respondents of a web survey with three options, being a survey of 5-7 minutes with no incentive or a survey of 10 minutes accompanied by either a guaranteed incentive in the form of a $5 gift card or a chance to win an iPad. The majority opted for the longer survey, preferring a chance to win an iPad over a guaranteed gift card. The findings of Ziegenfuss et al. (2013) contradict the research of Dykema et al. (2011), which stated a more significant reduction in dropout when all respondents are certain of being granted compensation as opposed to having a chance at winning the lottery, even if the amount they are sure to receive is less than the reward of the lottery (Dykema et al., 2011).

Both products and money may therefore be a suitable way to reduce dropout. However,

[1] https://www.surveymonkey.com/

organizations may only be able to allocate a finite amount of money. The impact that incentives may have on the dropout rate is limited, which is why researchers have tried to find alternative options to reduce dropout on long surveys.

2.2.4. Priming

Priming has been proven to be an effective alternative to incentives to alter the behaviour of respondents. Primes are stimuli in the environment which cause the user to change their thoughts or perform an action without them being aware of it (J. Bargh & Chartrand, 2000) and may have different effects on the respondents' behaviour depending on their type.

Harrison et al. (2013) demonstrated that presenting a positive brief narrative can enhance precision in visual assessments. First, respondents were assigned to read either a positive or negative story. After this, they were shown a chart and had to estimate the relative size of one data entry to another. The accuracy in common graphical perception tasks of negatively primed respondents did not decrease, whereas the performance of positively primed respondents increased.

Positive images can improve creative performance as demonstrated by Lewis et al. (2011) and Morris et al. (2012). Respondents were shown an image next to the task, which requested them to describe unique usages of everyday objects. Respondents who were shown a positive or negative picture produced more original answers than the control group.

Morris et al. (2013), Morris et al. (2012) also demonstrated how music could increase performance on insight-based problem-solving tasks. The participants listened to either a cheerful jazz piece or a section of a negative classical composition. Respondents in the positive priming condition outperformed respondents in the negative condition.

Another form of priming is achievement priming. Achievement primes are stimuli which activate achievement motivation, a theory that describes how the desire to accomplish a goal depends on someone's "chronic achievement motivation" (the amount of pleasure someone experiences doing so) (W. Hart & Albarracín, 2009). When primed, people with high chronic achievement motivation seek to excel, whereas people with low chronic achievement motivation seek to enjoy themselves. These two groups are mutually exclusive: those who seek to excel do it at the expense of having fun, and vice versa.

J. A. Bargh et al. (2001) demonstrated how achievement-related words could increase performance in information retrieval tasks. Respondents exposed to these words in a word search puzzle or Scrabble game found or created more words than the control group. Oikawa (2004) exposed the respondents to either positive activation words or neutral words before starting an unrelated mathematical task and observed a performance improvement when exposed to achievement priming.

However, the words used for the achievement primes and neutral words were of a different word class. The activation primes were verbs (e.g. win, master, or excel), and the neutral words were nouns (e.g. ranch, carpet, or river). Since verbs are a call to action and nouns are not, the difference in performance might have been caused by this difference in word class instead of being caused by the use of activation words. Engeser (2009) therefore performed similar experiments as J. A. Bargh et al. (2001) and Oikawa (2004), but made the neutral words and activation words be of the same word class. The results showed that the use of activation primes was indeed the cause of the increased performance. In later research, W. Hart and Albarracín (2009) further confirmed that using motivational words as positive achievement primes increases performance compared to respondents exposed to neutral words.

All aforementioned research, however, apply achievement motivation to problems that are not representative of real-world survey tasks. Gadiraju and Dietze (2017) were the first to research whether achievement primes can be used in generalized surveys by incorporating motivational quotes as *active primes* or *passive primes*. Active primes are embedded within the task and are therefore required to be interacted with. On the other hand, passive primes serve as supplementary information and therefore do not require interaction. Gadiraju and Dietze (2017) demonstrated that crowdworkers inherently have high chronic achievement motivation, and as a result, when exposed to the primes, they did not drop out or dropped out at a later stage. Particularly, active achievement primes are more effective at reducing dropout than passive achievement primes.

The advantage of adding achievement primes over increasing incentives is that no additional costs are required. However, two important restrictions must be considered when adding achievement primes to surveys. Firstly, achievement primes should be continuously reintroduced to respondents, as the risk

of losing the priming effect increases with the length of a study (Hermans et al., 2001). Secondly, adding achievement primes should minimally add to respondents' cognitive load. A higher cognitive load can prolong the survey completion time. As established before, this results in more dropouts, which is contrary to the purpose of this study.

In summary, adding achievement primes to surveys is a promising method to reduce dropout rates in long web surveys, provided that the primes are continuously reintroduced, and the cognitive load is impacted as little as possible.

3

Experimental Setup

This study aims to understand how to use engagement to reduce dropout on long web surveys. In this section we design an experiment where we expose respondents of a web survey to achievement primes. Through this design, we are able to investigate a number of hypotheses and answer a set of research questions.

3.1. Goal

The goal of this study is to answer the following Research Questions (**RQ**) by means of the corresponding Hypotheses (**H**):

RQ1 **How does achievement priming affect the dropout rate on long surveys?**
Primes are stimuli in the environment which cause the user to change their thoughts or perform an action without them being aware of it (J. Bargh & Chartrand, 2000). The effect of primes can be reinstated through repeated exposure (Hermans et al., 2001). As achievement primes have been shown to prevent or delay dropout on short surveys (Gadiraju & Dietze, 2017), repeated exposure should cause the same effect on long surveys. The effect of achievement primes depends on how they are presented and interacted with. Active achievement primes require respondents to actively engage with the prime, whereas passive achievement primes may be ignored. The priming effect is therefore expected to be stronger when the user is tasked to engage with them actively. This research question will be answered by means of a between-subject experiment, which will confirm or refute the following hypotheses:

- **H1A** There is less dropout when respondents are exposed to active achievement priming compared to no priming.
- **H1B** There is less dropout when respondents are exposed to passive achievement priming.
- **H1C** Active achievement primes are more effective than passive achievement primes at reducing dropout.

RQ2 **How does achievement priming influence respondents with different motivations to participate?**
The initial motivation to participate has a strong impact on the behaviour of respondents. Respondents may either be extrinsically motivated, e.g. filling out incentivised surveys as a source of income, or internally motivated, e.g. filling out to contribute to research (Veit et al., 2011). Crowdworkers fill out surveys in exchange for income and are therefore classified as extrinsically motivated. They exhibit high chronic achievement motivation and have been shown to be susceptible to achievement primes (Gadiraju & Dietze, 2017). Extrinsically motivated respondents start a survey based on external stimuli and may be more susceptible to the exposure of primes than internally motivated participants. However, to the best of our knowledge, there is no research on this for intrinsically motivated respondents. This research question will be answered by means of a between-subject experiment, which will confirm or refute the following hypothesis:

H2 People are less affected by primes when they participate based on intrinsic motivation than when they participate because of extrinsic motivation.

3.2. Variables

In this section, we provide details on the independent, dependent, and descriptive variables.

3.2.1. Independent Variables

Independent variables are variables that are manipulated or changed by the researchers and may influence the dependent variables. There are two independent variables, being respondent group and questionnaire variant.

There are two groups of respondents *(categorical)*, those who participate based on intrinsic motivation (**IM**) and those who participate based on extrinsic motivation (**EM**). For *RQ1*, both respondent groups are evaluated independently from one another.

The three variations of the questionnaire *(categorical)* will contain either no primes (**NP**), passive achievement primes (**AP-P**), or active achievement primes (**AP-A**). For *RQ2*, `AP-P` and `AP-A` are evaluated independently from one another.

3.2.2. Dependent Variables

Dependent variables are the variables that are measured in the study, and changes in these variables are believed to be caused by changes in the independent variables. Dropout, perceived workload, and user engagement are used as dependent variables in this study.

Dropout

(Categorical/continuous; between-subjects)

Dropouts are recognised by the point at which the respondent leaves the web survey prematurely and therefore does not submit the data (Standard Definitions, 2016). In the context of this survey, dropout constitutes of three factors.

Based on the definition of dropouts *(categorical)*, the first factor is whether or not a respondent finished the entire survey, including post-questionnaires.

The second factor is the duration (in seconds) filling out the questionnaire *(continuous)*. The duration is partially linked to dropout, as a respondent who dropped out early in the survey likely spent less time on the survey compared to those who finished it. By considering the duration, we gain insights into the relationship between survey completion and time invested by respondents. Additionally, the time spent on the survey has financial implications, particularly when using crowdworkers as respondents. Crowdworkers are typically compensated based on the amount of time they spend completing the survey. Therefore, by measuring the duration of the survey, we can assess the financial costs associated with running the survey.

Thirdly, the page surpassed by the respondent *(continuous)* is also evaluated as part of dropout, allowing us to assess the impact of achievement primes on dropout at specific stages of the survey. The `AP-P` and `AP-A` variants include one achievement prime per page and thereby we can capture the effect that a single prime has on dropout on the corresponding page.

Perceived Workload

(Continuous; between-subjects)

Perceived workload is used to measure whether the primes cause additional cognitive load. An increase in cognitive load may result in increased dropout, which is contrary to the goal of this study. The perceived workload is measured using NASA's Task Load Index (TLX) (S. Hart, 1986).

Participants are asked to rate their experience across six dimensions: mental demand, physical demand, temporal demand, effort, performance, and frustration level. Each dimension is rated on a scale ranging from (1) *Very low* to (20) *Very high*, except for performance, which ranges from (1) *Perfect* to (20) *Failure*. TLX typically includes a pairwise-comparison task of each domain to capture the subjective weighting before asking the users to rate their experience on six scales. For the purpose of this study, the pairwise-comparison has been removed as it is considered time-consuming (Stanton

et al., 2005). Instead, all domains are weighted equally. Participants receive a briefing on the purpose of NASA TLX before filling it out.

To calculate the workload score, the ratings for all dimensions are summed. The resulting score ranges from 6 to 100.

User Engagement
(Continuous; between-subjects)
User engagement is included as a dependent variable as this study aims to reduce dropout rates through increased engagement. By measuring user engagement, we can assess the effectiveness of achievement primes. Even if there is no significant difference in dropout rates, the reported level of engagement may still be influenced by the presence of achievement primes. To capture user engagement, we used the User Engagement Scale - Short Form (UES-SF) (O'Brien et al., 2018). For the same reason why the weighting of NASA TLX was removed for our study, the Short Form version of the UES was chosen to ensure the post-questionnaire does not consume excessive time. The UES-SF consists of four subscales: Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward (RW). Each subscale comprises three questions that are rated on a 5-point Likert scale. The following questions were posed to the respondents:

Focused Attention (FA):

FA-S.1: I lost myself in this experience.

FA-S.2: The time I spent filling out this web survey just slipped away.

FA-S.3: I was absorbed in this experience.

Perceived Usability (PU):

PU-S.1: I felt frustrated while filling out this web survey.

PU-S.2: I found this web survey confusing to fill out.

PU-S.3: Filling out this web survey was taxing.

Aesthetic Appeal (AE):

AE-S.1: This web survey was attractive.

AE-S.2: This web survey was aesthetically appealing.

AE-S.3: This web survey appealed to my senses.

Reward (RW):

RW-S.1: Filling out this web survey was worthwhile.

RW-S.2: My experience was rewarding.

RW-S.3: I felt interested in this experience.

When calculating the overall engagement measure, the scores for the Perceived Usability (PU) subscale are reversed. Engagement scores for each subscale are calculated by taking the mean score, and the overall engagement score can be obtained by taking the mean score across all subscales.

3.2.3. Descriptive Variables

Descriptive variables encompass demographic variables, such as gender and age, along with exploratory variables, like mood.

Demographic Variables
Demographic variables are not controlled for but may influence the dependent variables. They help to better understand and differentiate between population groups. The descriptive variables captured for this study are gender and age.

Gender *(categorical)* is included as people may act differently when responding to web surveys depending on gender. For example, males are more likely to respond to questionnaires if they are sent by female surveyors (Keusch, 2012). It has not been researched whether this also holds the other way

around or what the effect would be for non-binary genders.
Options: Male, Female, Other, Prefer not to say.

Age *(continuous)* was found to be (negatively) linked to what respondents regard a survey's maximum length to be (Revilla & Ochoa, 2017). Since the length of a survey is positively correlated with dropout (Hoerger, 2010; Liu & Wronski, 2018b), age may be of influence on the dropout rate of the survey.
Options: number input.

Exploratory Variables
Exploratory variables are also captured, as prior work suggests these factors may be influential. Similar to descriptive variables, they are not controlled yet may influence the dependent variables. They can change from time to time, however. This study's exploratory variable is mood.

Mood *(categorical)* has proven to be of influence on the user's experience while performing search tasks (Xu et al., 2019), and may therefore influence the user's perceived burden. If someone is already in a cheerful or excited mood, their threshold of burden may be prolonged when compared to people who start a survey whilst in an irritated or tense mood. Thereby, they may drop out later. Figure D.1 shows a picture of the moods (Desmet, Vastenburg, & Romero, 2016) provided to respondents to assist them with their choice. The picture can be found in Appendix D.
Options: Bored, Calm, Cheerful, Excited, Irritated, Neutral, Relaxed, Sad, Tense (Desmet, Vastenburg, Romero, & Pieter, 2016).

3.3. Achievement Prime Curation

Similar to research by Gadiraju and Dietze (2017), inspiring quotes were chosen as achievement primes. To select the achievement primes, a total of 100 quotes were collected and rated by crowdworkers on their inspirational value. The 25 quotes that received the highest average ratings were ultimately chosen to be used as achievement primes.

3.3.1. Collecting Quotes as Achievement Primes

To investigate the general applicability of inspiring quotes as achievement primes in long surveys we selected those which were context- and domain-independent. This ensures that the results of this study can be generalised and applied to questionnaires across various domains. The procedure of collecting the quotes consisted of (i) obtaining quotes, (ii) sorting based on likes, (iii) limiting length, (iv) performing sentiment analysis, (v) limiting quotes per author, and (vi) reviewing manually.

(i) *Obtaining quotes:* We obtained the selected quotes from GoodReads[1], a social media platform where users can submit quotes and assign them tags. All quotes assigned with the following tags associated with inspiration were collected: `inspiration`, `inspirational`, `inspirational-quotes`, `inspire`, and `inspiring`.

(ii) *Sorting based on likes:* GoodReads' users may indicate their approval of quotes by giving them likes, which creates a natural ranking of quotes. As such, the selected quotes were sorted based on the number of likes they received.

(iii) *Limiting length:* After collecting the quotes from GoodReads, we removed quotes with a length exceeding 25 words from consideration. Achievement primes must not exceed a certain length, which would increase the cognitive load and potentially increase the respondent's burden. The length was chosen based on the question in "A Questionnaire for Understanding Worker Health" consisting of the most words, as exceeding the length of the longest question in the survey could increase the cognitive load.

(iv) *Performing sentiment analysis:* Subsequently, we performed sentiment analysis on the quotes, and quotes with a negative sentiment were removed from consideration. Achievement primes should cause a respondent to feel the need to succeed or attain excellence (W. Hart & Albarracín, 2009), and should not contain negative sentiment. The quotes' sentiment was

[1] https://www.goodreads.com

analysed by the `DistilBERT base uncased finetuned SST-2`[2] model using the Python packages `Transformers`[3] and `Pytorch`[4].

(v) *Limiting quotes per author:* Following the sentiment analysis, we limited the number of quotes to one per author to ensure a variance. Specific authors received more likes on their quotes compared to others. By removing duplicate authors, the possible influence of the authors' quotes on the study's validity is reduced.

(vi) *Reviewing manually:* Finally, we manually reviewed the quotes to remove quotes that did not fit the definition of achievement primes. Examples of reasons why quotes were removed were that they falsely passed the automated sentiment analysis ("Hell is empty and all the devils are here."), imposed religious beliefs ("In the name of God, stop a moment, cease your work, look around you."), or focused on concluding rather than continuing an activity ("So comes snow after fire, and even dragons have their endings.").

3.3.2. Evaluation of the Inspiration Value

To select only the most inspirational quotes as achievement primes, the top 100 quotes were evaluated by crowdworkers on the Prolific[5] platform for which they were rewarded a fair wage. The workers were tasked to judge how inspiring the quotes were on a 5-point Likert scale with the values: "very uninspiring" (1), "somewhat uninspiring" (2), "neither uninspiring nor inspiring" (3), "somewhat inspiring" (4), and "very inspiring" (5) to ensure explicitness in the differences in answers. The format of all questions was identical, as illustrated in Figure 3.1. The quotes' authors were deliberately not shown to prevent possible bias towards specific authors. The best practices for survey design, as defined in Section 2.1, have been considered.

To ensure high-quality responses, we selected only workers who were fluent in English and had an approval rate of at least 90 out of 100. As per our testing, the completion time of our questionnaire was roughly 15 minutes. The hourly wage was set to £10 to provide fair compensation, resulting in a payout of £2,50 after completing the questionnaire. To ensure a sufficient number of responses for analysis, we gathered ten judgements per quote from distinct workers. To prevent an unequal distribution of ratings due to possible dropout, we randomized the order of the quotes for each respondent.

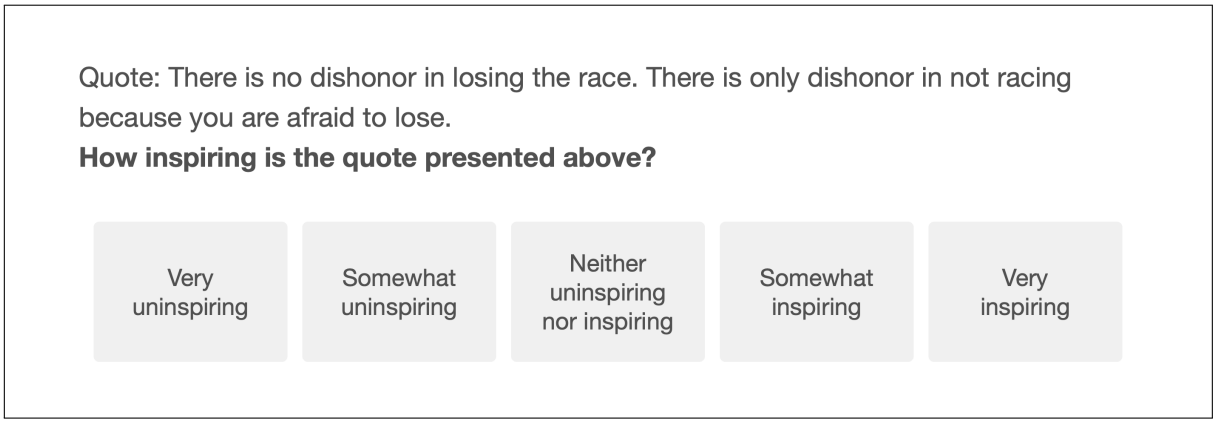


Figure 3.1: Example of a survey question used for assessing the inspirational value of a quote.

The opening statement of the questionnaire contained information regarding Informed Consent, the text of which can be found in Appendix B.1. The Informed Consent was based on TU Delft's Informed Consent Templates and Guide[6] and included all critical points as advised by the TU Delft's Human Research Ethics Committee. The opening statement was included in the job description on Prolific and on the first page of the Qualtrics survey to ensure that the respondent did not unintentionally skip

[2] https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english
[3] https://pypi.org/project/transformers/4.26.1/
[4] https://pypi.org/project/torch/1.13.1
[5] https://www.prolific.co
[6] https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics/informed-consent-templates-and-guide

the information. We automatically collected the respondents' Prolific Personal IDs to ensure that the respondents would be paid out and deleted this information once the payment successfully ensured the respondents' privacy.

Two attention check questions were included to filter out data from inattentive respondents. We used directed queries as attention checks, which are highly effective at detecting malicious workers (Abbey & Meloy, 2017) and are the only type allowed on Prolific. Directed queries ask the user to answer the question in a precise manner and can often stand out from the rest of the questions. To blend in with the regular questions, the attention check questions were identical to the regular questions except for the replacement of the quote with a request to choose a specific value on the Likert scale. The decision to include two attention checks was based on previous research, which found that using one attention check in an incentivised survey that takes 3.5 minutes to complete is effective (Liu & Wronski, 2018a), and using two attention checks in a *non*-incentivised survey that takes 15 minutes to complete is the best amount (Olamijuwon, 2021). Although our survey is incentivised, similar to the former research, in our testing, the completion time of our questionnaire matched the latter research, and we therefore adhered to their recommendation.

The data of respondents for whom there were high suspicions of extreme speeding were also filtered out. Speeding occurs when respondents prioritize finishing a survey over providing answers of high quality, which negatively affects the quality of the results (Zhang et al., 2014). Similar to research by Zhang et al. (2014), we set the reading speed threshold to be 300 milliseconds per word. To account for repetitive text in the question (specifically *"Quote:"* and *"How inspiring is the quote presented above?"*), we excluded them from the calculation of the minimum completion speed. We also excluded the opening statement, as it is displayed on the Prolific and Qualtrics platforms. Consequently, the minimum completion time depended only on the quotes presented in the survey. As we could not record the answer time per question, we compared the time it took the respondents to complete the entire survey to the calculated minimum time.

After collecting the results, the 25 quotes which scored the highest on inspirational value were selected as achievement primes.

3.4. Study Design

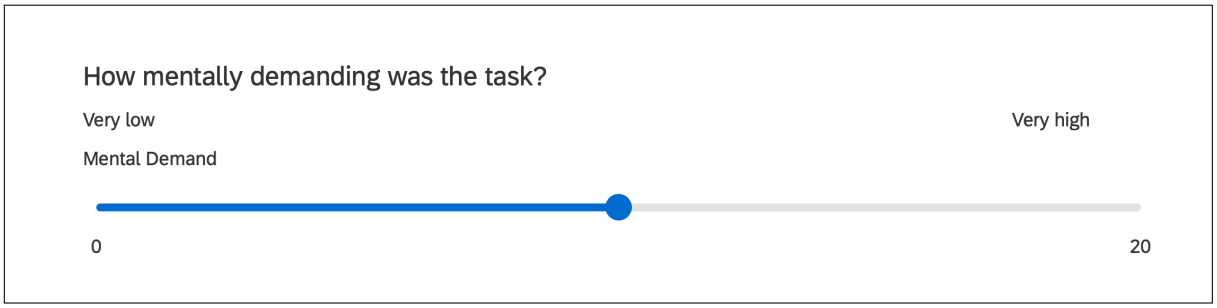
This section provides details on the design of the main study, where we discuss the design of the questionnaire and the independent variables.

3.4.1. Questionnaire Design

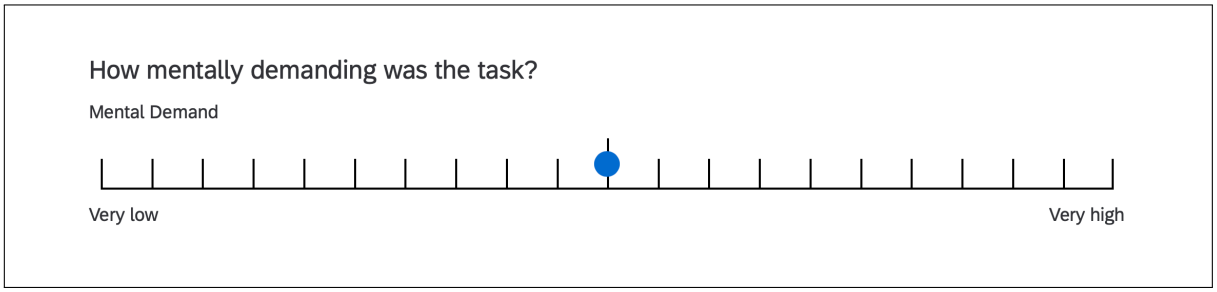
The questionnaire that was chosen for this study is "A Questionnaire for Understanding Worker Health" and was appended with NASA's Task Load Index (TLX) (S. Hart, 1986) and User Engagement Scale Short Form (UES-SF) (O'Brien et al., 2018) surveys to measure perceived workload and user engagement respectively. The combined survey constitutes a long survey as defined as taking 15 minutes or longer in Section 2.2.1, as the main questionnaire's average survey completion time was 14 minutes (850 sec) in a pilot study, 3 minutes for TLX (Bell et al., 2022), and 5 minutes for UES-SF (O'Brien et al., 2018). Moreover, the questionnaire does not require prior knowledge as the questionnaire's questions are designed to ask about the subjective experience of the respondents' work environment. Because of these two factors, the questionnaire is well-suited for this study.

Modified Text

To ensure that both respondent groups could answer the questionnaire, some questions from the original survey were modified. The original survey was designed with crowdworkers in mind. As non-crowdworkers should also complete the same survey, questions that mentioned crowdwork were generalised. The main focus of the question remained the same after making modifications. For example, question 1.5 originally read as *"How many hours do you work on MTurk/Prolific each day on average?"*, and was revised to *"How many hours do you work on a computer each day on average?"*. Both questionnaire versions are in Appendix D. The crowdworkers were presented with the original questionnaire, as they might have other jobs besides crowdwork which may influence their answers otherwise. The non-crowdworkers received the modified version. The answer type and options for each question can also be found in Appendix D. All answer types of the original questionnaire were



(a) Example of the default design of a slider in Qualtrics.



(b) Example of the custom design of the slider, which more closely matches the original design of NASA TLX's sliders.

Figure 3.2: Examples of the default and custom design of the slider for the NASA TLX post-questionnaire.

provided to us. However, not all options were available at the time of creating the survey. For questions with ambiguous options, we chose options that best fit the intended meaning of the question. Moreover, we slightly altered Q10 to better align with the intentions of the source material. Instead of giving examples ("control, display, compatibility, layout, posture etc.") the revised question is now explicit ("layout, seating, displays, and compatibility"). Additionally, we added "How are your hands/wrists and arms?" as a question to Matrix starting with ID 24.

Modified Questions and Design

The questionnaire was designed in Qualtrics[7], and the default design elements of the "Simple" layout were used. The selection options were displayed vertically to make them easy to compare, whereas all Likert scales were displayed horizontally to insinuate the continuous property of this answer type. The third question asks the user to explain their mood, which is accompanied by the Pick-A-Mood scale (Desmet, Vastenburg, & Romero, 2016), to assist respondents in establishing their current feeling. The picture can be seen in Appendix D. Matrix questions were used for Questions 24 up until and including 29, and Questions 41 up until and including 49, despite the recommendations of Section 2.1. This suggestion was made due to the increased perceived burden caused by the repetitive nature of matrices. The questions themselves, however, already repeat a portion of the preceding question. This repetition can be removed by displaying the questions as one matrix answer type instead of separate Likert scales, thereby reducing the perceived burden. The range input for the TLX sub-questionnaire was custom designed to better reflect the original style of the official survey compared to the default range slider provided by Qualtrics (S. Hart, 1986). The differences between the default and custom-built slider can be seen in Figure 3.2. The respondent was shown the value they selected on the right side of the scale (ranging from 1 to 20).

Pagination

The main questionnaire was organized into separate pages for each of its four parts, which all focus on distinct topics. However, an exception was made for Question 15, originally Part II's first question. When using *skip logic* in Qualtrics, subsequent questions are automatically put on a new page. Since Question 15 had skip logic for Questions 17, 18, and 19, leaving it as the first question on a page would have created a separate page for only one question, disrupting the survey flow. We moved Question 15 to the end of Part I to avoid this issue. After completing the main questionnaire, a separate page

[7] https://www.qualtrics.com

explains that the user will be asked to complete the TLX and UES-SF post-questionnaires. Both take up one page, bringing the total to seven pages. The instructions from S. Hart (1986) and O'Brien et al. (2018) were followed while constructing the supplementary surveys.

Attention Checks
Two attention checks were added to the survey to filter out inattentive respondents, which is the same amount as the design of the pre-study described in Section 3.3 which has a comparable survey length. The attention checks were in the form of directed queries. These attention checks are placed in the main survey's matrix questions 41 to 49 and in the matrix question of the supplementary survey UES-SF, which are on the third and last page, respectively. The attention checks were deliberately placed in Matrices because such questions are considered burdensome (Kaczmirek, 2008). This is the cause of the inattention we were trying to detect. Moreover, these Matrices are distributed evenly across the survey.

Opening Statement
This main study was preceded by information regarding Informed Consent, the text of which can be found in Appendix B.2 and was based on TU Delft's Informed Consent Templates and Guide[8]. Contrary to the Informed Consent of the pre-study in Section 3.3, this opening statement does not contain the title or purpose of this research as it contains too many details on the research and might cause bias. Like the aforementioned Informed Consent, the opening statement was also shown on Prolific, and the respondents' Prolific Personal IDs were collected and deleted after payment was successful.

3.4.2. Questionnaire Variants

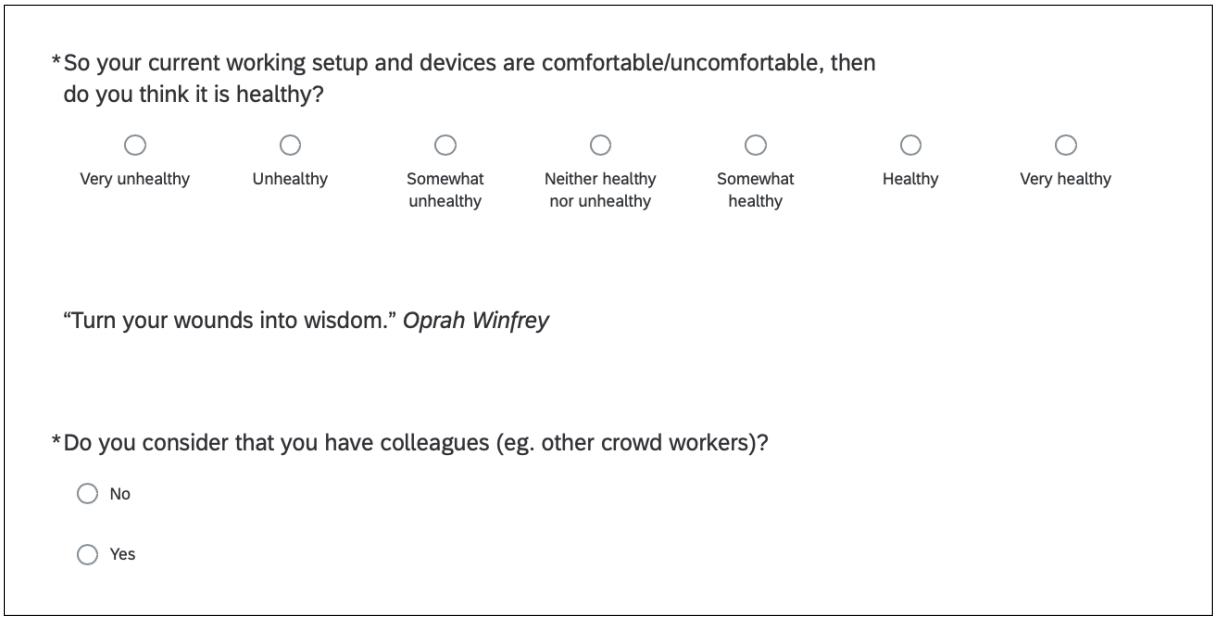
The survey consists of "A Questionnaire for Understanding Worker Health", which is appended by an altered version of the TLX and the UES-SF. The surveys differ in the main questionnaire, in which either (i) No Achievement Primes (NAP), (ii) Passive Achievement Primes (AP-P), or (iii) Active Achievement Primes (AP-A) are introduced.

The `NP` variant is a one-on-one copy of the main questionnaire. This survey version does not contain achievement primes and is hypothesised to have the worst dropout of all variations across respondent groups.

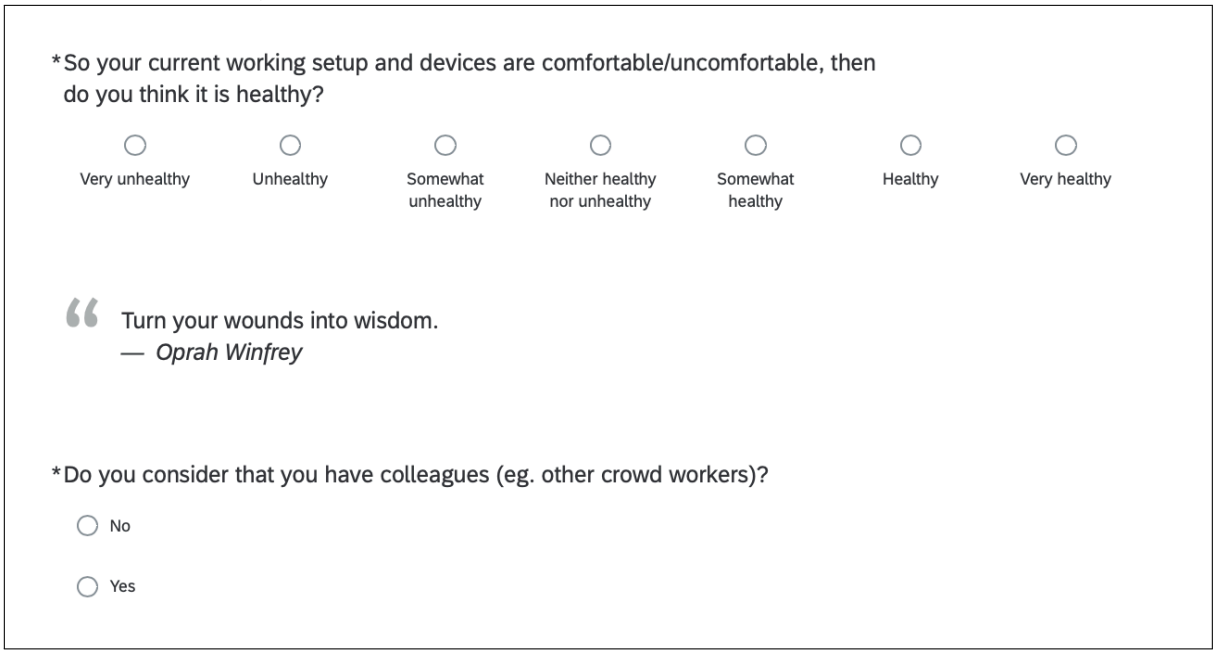
The `AP-P` variant contains quotes interspersed at specific locations between the regular quotes in the main questionnaires. Per the definition of `AP-P` , the quotes are non-interactive. The design of the quotes is slightly changed from regular questions, so they are not confused to be part of a succeeding question. The default design of the quotes blends in with the rest of the questionnaires. The change, which puts a large open-quote mark before the quote and the large gaps between the questions, gives the impression that the quote is a standalone element. The change in appearance can be seen in Figure 3.3. Unlike the pre-study, these questions include the author's name, which may create a positive connection with the prime and increase its effects.

The `AP-A` variant contains questions interspersed at the same locations as the Passive Achievement Primes. Respondents are presented with a quote and are presented with four options. They are tasked to choose which word would be the best replacement of one of the words in the quote. One of the options is a synonym or a likely replacement, whereas the other options are obviously the wrong answer. We opted for a *multiple selection* question format instead of a *free text* question to prevent ambiguity and to increase the probability they interact with the quote correctly. Moreover, by providing pre-defined answer choices, we aim to prevent users from leaving the survey to find information elsewhere, which would otherwise be the case in an *information finding* task. We prefer to ask for a replacement of a noun as it is usually the subject matter of the quote. If a noun is not present in the sentence, we ask for a replacement of an adjective instead as it is typically the next most important word in the sentence. If both a noun and an adjective are absent, we ask for a replacement of a verb. If a sentence contains multiple nouns, adjectives, or verbs, we ask for a replacement of the last one in the sentence. By doing so, we provide the user with more context to base their answer on as they will have been able to read more information which may help them choose the correct word. These alternatives are intentionally incorrect, prompting the user to select

[8] https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics/informed-consent-templates-and-guide



(a) Example of the default design of the passive achievement primes.



(b) Example of the custom design of the passive primes used to make the prime easier to differentiate from preceding and succeeding questions.

Figure 3.3: Examples of the default and custom design of the passive achievement primes.

the correct word and actively engage with the quote. The synonyms were collected from Thesaurus[9], and alternatives were collected from a Random Word Generator[10]. The options in the multiple choice question were randomised for each participant in order to prevent option ordering effect (Terentev & Maloshonok, 2019). As we want to test the prime variation's effect, not the prime's design, the `AP-A` quotes are styled in the same way as the `AP-P` . By using the same design of the primes in the `AP-P` and `AP-A` versions, the difference in behaviour is caused by the differences of the prime itself and not the differences in the design. The quotes, words, and alternatives can be found in Appendix C. Figure 3.4 shows an example of the question.

[9]https://www.thesaurus.com
[10]https://randomwordgenerator.com

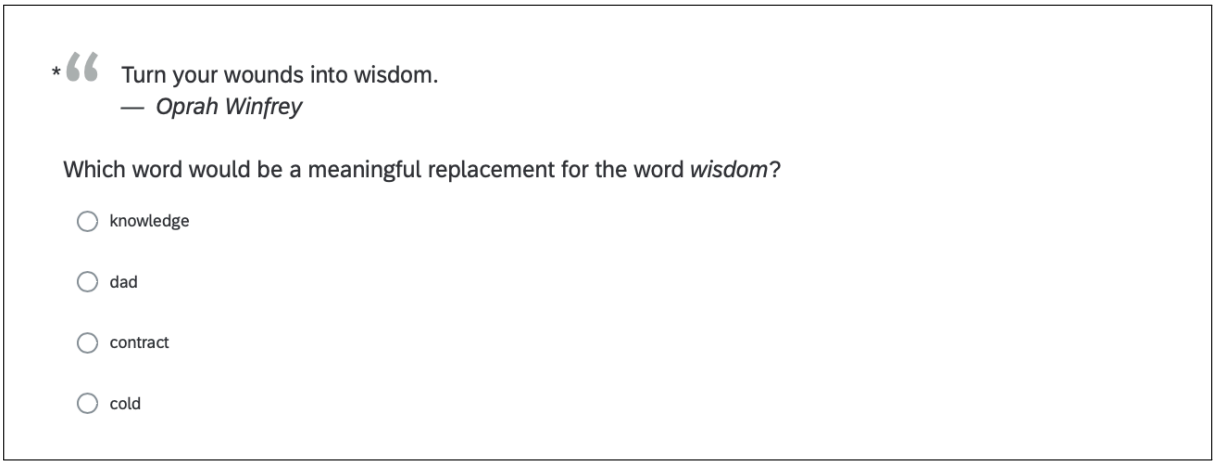


Figure 3.4: Example of the custom design of the active achievement prime questions, made to closely match the passive achievement primes.

Frequency and Placement of Primes

One prime was placed pseudo-randomly on each of the four pages of the main questionnaire. The stimulating effect of primes should not be noticeable to the respondent. Therefore they were consistently placed on every page to blend in with the rest of the survey. Moreover, questions and their order should appear persistent (Kaczmirek, 2008; Keusch, 2012), and due to the different appearance of the primes compared to other questions, they should be consistently placed across the pages. The primes were deliberately not inserted into the supplementary TLX and UES-SF surveys to prevent introducing the primes to the respondent while asking the respondents' subjective experience of the effects of these primes. To avoid an ordering effect, where the respondent's actions are caused by the order of the questions rather than the primes, the primes were placed pseudo-randomly. Using a randomisation algorithm, the locations of the primes were determined to be after questions 11, 18, 35, and 56. The locations within matrices were excluded from the algorithm, as the primes were of an incompatible answer type. The primes were divided into groups, each being assigned a different location. Qualtrics does not allow a set of questions to be randomised across multiple locations, limiting randomisation to one location. The primes are distributed amongst the four locations to prevent the prime from being shown to the same respondent multiple times. This problem could have occurred if one prime was randomly selected from the entire collection and would limit the effect of the primes. As splitting the groups according to their rank order would result in inconsistent average inspirational values between groups, the following mathematical equation was used to assign quotes to prime locations:

$$g_i = q_{id} \mod 4 + i$$

with g_i being group $i \in \{1, 2, 3, 4\}$, and q_{id} the question with ID id. Appendix D includes the primes' location between the questions and options.

3.4.3. Groups of Participants

There are two groups of participants, being Intrinsically Motivated (IM) or Extrinsically Motivated (EM).

`IM` respondents will be recruited from the public social media channels Reddit and Facebook, and via private social media channels such as WhatsApp or Instagram. The following is a list of the URLs where the web survey was posted:

- https://www.reddit.com/r/SampleSize/
- https://www.reddit.com/r/takemysurvey/
- https://www.reddit.com/r/Favors/
- https://www.facebook.com/groups/surveysharing/

This group was also asked to participate only if they (i) are 18 years or older due to legal restrictions, (ii) speak English fluently due to the language in the questionnaire, and (iii) perform some kind of work behind a computer due to the questionnaire topic.

EM respondents were recruited from the online crowdworking platform Prolific. Respondents were rewarded with £10,00/h for completing the task in order to allow fair pay. The participants were pre-screened such that they (i) are 18 years or older due to legal restrictions, (ii) speak English fluently due to the language in the questionnaire, and (iii) have a minimum approval rate of 90 (out of 100) to ensure high quality responses. The criteria that respondent must perform some kind of work behind a computer, as established for the Intrinsically Motivated group, already holds for this group due to the nature of crowdwork.

3.5. Procedure

Participants went through four phases during this research: (i) introduction, (ii) main questionnaire, (iii) post-task questionnaire, and (iv) completion.

(i) *Introduction:* Participants are explained the purpose of the survey, as well as informed on who the surveyors are and whom to contact in case there are any questions. The duration of the questionnaire is provided, and the participant is ensured they may withdraw at any time. Moreover, the participant is ensured that their data is be anonymised.

(ii) *Main questionnaire:* Participants are asked demographic information, after which they are asked questions regarding their health while doing crowdwork or while working behind a computer. The first part asks about the worker's background, the second about their working ergonomics and physical health, the third about psychosocial conditions and mental health, and the last about the worker's needs. The answers to these questions are recorded but are not reported in the results.

(iii) *Post-task questionnaires:* After completing the main questionnaire, the participants are shown NASA's Task Load Index (TLX) and User Engagement Scale Short Form (UES-SF), in that order.

(iv) *Completion:* Crowdworkers from Prolific are sent back to the platform, and their submission is automatically approved. In case users drop out, their submissions are manually approved. Volunteers are thanked for their participation.

3.6. Analysis

In order to answer the research questions, we needed to analyse the data to examine the influence of the independent variables on the dependent variables. To perform these analyses, we applied various statistical tests, selected according to the flowchart created by Draws (2021).

Data obtained from the respondents was subject to be filtered out in two cases. The first case would be if the respondent wrongfully answered an attention check, indicating they were not attentive during the survey. Secondly, implicit refusals were filtered out as they do not fit the definition of dropout established in Section 3.2.2. Implicit refusals are respondents who drop out without completing a single survey item (Standard Definitions, 2016). It is impossible to know whether the respondents who dropped out on the first page of the main questionnaire are true dropouts, having answered at least one question, or are implicit refusals. Moreover, in the questionnaire variants containing a prime, we cannot know whether a respondent has observed at least one prime if they have not surpassed the first page. As the prime is the element we want to observe the effect of, we can only use data entries where we are certain the user has been exposed to the prime. Therefore, we did not consider the data of any respondent who dropped out on the first page.

3.6.1. Statistical Tests

For both the IM and EM respondent groups, we analysed the impact of the questionnaire variant to address RQ1. Furthermore, for both passive and active prime variants, we analysed the influence of the respondent group to address RQ2.

Besides analysing the data by means of significance, we used estimation plots to visualize the distribution of data across the various different experimental conditions by means of the visualisation tool DABEST by Ho et al. (2019). Specifically, we established the NP condition as the baseline for addressing RQ1, while the IM condition served as the baseline for addressing RQ2. The statistical tests were performed using the software JASP by JASP Team (2023).

Effect of Questionaire Variant
The first research question focuses on the effect of the questionnaire variant on the dependent variables. This analysis is performed separately for two respondent groups: `IM` and `EM` respondents. For analysing whether the independent variable has an impact on whether the respondent finished the entire questionnaire, we employed the Chi-Squared Test. The continuous dependent variables of duration, page surpassed, perceived workload, and user engagement, were analysed using the One-Way Analysis of Variance (ANOVA) test. ANOVA establishes *whether* there is a difference between groups, after which we required the post hoc test Tukey's Honest Significant Difference (Tukey HSD) to determine *between which groups* there was a difference. It is important to note that the One-Way ANOVA test assumes the data follows a normal distribution, which is verified by means of a normal quantile-quantile (Q-Q) plot. If the data points fall approximately along a straight line, this indicates the data follows a normal distribution. Deviation from the straight line may indicate a light or heavy tail, a skew or non-normality. In case we observe no normality, the non-parametric Kruskal-Wallis Test was used as an alternative, with the post hoc test being the Dunn test.

Effect of Respondent Group
The second research question examines the dependent variables across two respondent groups, namely `IM` and `EM` respondents, within the `AP-P` and `AP-A` variants of the questionnaire. The impact of the respondent group on finishing the questionnaire was analysed using Fisher's Exact Test. All continuous dependent variables were analysed using the Independent Samples t-Test. Similar to the One-Way ANOVA test, the Independent Samples t-Test assumes a normal distribution, which was tested by means of Shapiro-Wilk. For the Shapiro-Wilk tests, if the p-value is less than the significance threshold, the data is non-normal. If there was no normality, the non-parametric Mann-Whitney U-Test was used as an alternative.

3.6.2. Significance
We aimed to test against an overall significance level of $\alpha = 0.05$ to control the probability of Type I errors (falsely accepting a hypothesis). However, when conducting multiple statistical tests on the same dataset, the family-wise error rate (FWER) increases, leading to an elevated chance of Type I errors occurring. For example, if 20 significant tests are performed on the same dataset with a significance level of 0.05, the probability of at least one Type I error rises to 0.64 (Napierala, 1995).

To address this issue, we applied the Bonferroni-Holm correction to adjust the obtained p-values (Abdi, 2010). The Bonferroni-Holm correction is stricter than the Bonferroni correction alone, reducing the occurrence of FWER. To implement the Bonferroni-Holm correction post-hoc, we calculate the p-values for each test, order them from lowest to highest, and assign them a rank (i) within a family of tests. Using the formula:

$$p'_{i|C} = (C - i + 1) \times p$$

we obtain the adjusted p-value, $p'_{i|C}$, for each test, where C represents the number of p-values in the family. If the adjusted $p'_{i|C}$ is greater than 1, it is set equal to 1. The significance threshold is adjusted to the overall significance threshold divided by the number of families. In our design, there were four families, leading to an $\alpha = \frac{0.05}{4} = 0.0125$.

For all p'_i-values, if the adjusted p-value is lower than the significance threshold α, it indicates a significant difference in the associated dependent variable between groups. Else, there is no statistical difference in the dependent variable between groups, and subsequent dependent variables with larger p' values are also not considered statistically different.

3.7. Participants
We performed a power analysis to determine the required sample size, which we found to be 288 respondents. This sample size ensures that any observed difference in dropout rates between conditions can be considered statistically significant. We used the software `G*Power`[11] to conduct the power analysis.

The statistical test used for calculating the sample size was One-Way ANOVA. The options used in `G*Power` correspond to test family `F tests` and statistical test `ANOVA: Fixed effects, omnibus,`

[11] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

`one-way`. Since the sample size was calculated beforehand, the type of power analysis was set to `A priori`.

The input parameters used for calculating the sample size were effect size, significance level, power, numerator degrees of freedom and number of groups. We set the effect size to $f = 0.25$, corresponding to a medium effect size according to Cohen (1992). The statistical significance level α is corrected by the amount of hypotheses that are tested, resulting in an α value of $\frac{0.05}{4} = 0.0125$, being the probability that we wrongfully claim that there is no significant difference. The Power $(1 - \beta)$ was set to 0.8, being the probability that we correctly claim that there is a significant difference. Due to our $A \times B$ factorial design, we had 6 groups.

4

Results

In this chapter, we present the results from executing the experimental setup outlined in Chapter 3, by which we mean to address our proposed research questions.

4.1. Achievement Prime Curation

To curate the achievement primes, ten crowdworkers were tasked to rate the inspirational value of 100 pre-processed quotes on a 5pt-Likert scale. The 25 most inspirational quotes were then used for the main survey.

To exclude the data from speeders, respondents must reach a minimum threshold of 400 seconds. This threshold was calculated based on a total of 1335 words to read (excluding the informed consent or repetitive text), with an assumed reading speed of 300 milliseconds per word. Amongst the participants, three individuals failed to meet the threshold and submitted their answers in 287, 309, and 383 seconds. This indicates they did not read all quotes thoroughly, so we filtered out their answers. On the other hand, the remaining seven respondents who surpassed the speeding threshold also passed both attention checks. Therefore, the average inspirational value of quotes is based on the data provided by these participants.

The quotes used as achievement primes range from an inspirational value of 4.00 to 4.71 out of 5.00 ($M=4.22, SD=0.18$). The results of Crowdworker Evaluation of the Inspiration Value of the Quotes can be found in Appendix C.

4.2. Descriptive Statistics

The total number of participants, before and after exclusions for implicit refusals or failure on at least one attention check, are summarized in Table 4.1. From the 129 intrinsically motivated (**IM**) participants, 41 data entries were removed due to 5 respondents failing at least one attention check and 36 instances of implicit refusal. From the 145 extrinsically motivated (**EM**) participants, 5 data entries were removed from the analysis, caused by three respondents failing at least one attention check and two instances of implicit refusal. Consequently, 88 `IM` participants and 140 `EM` participants were included in the final analysis.

The gender distribution amongst all participants comprised 134 females (59%), and 93 males (41%), with one participant opting not to disclose their gender (1%). Detailed gender distribution for each respondent group is presented in Table 4.2.

The participants ranged from 19 to 66 years, with a mean age of 30 and a standard deviation of 11 years. The average age, standard deviation, and minimum and maximum ages were consistent between the respondent groups. Detailed demographic data, including exact numbers and distribution plots, are provided in Table 4.3 and Figure 4.1, respectively.

Table 4.1: Amount of respondents before and after excluding implicit refusals or respondents who failed at least one attention check.

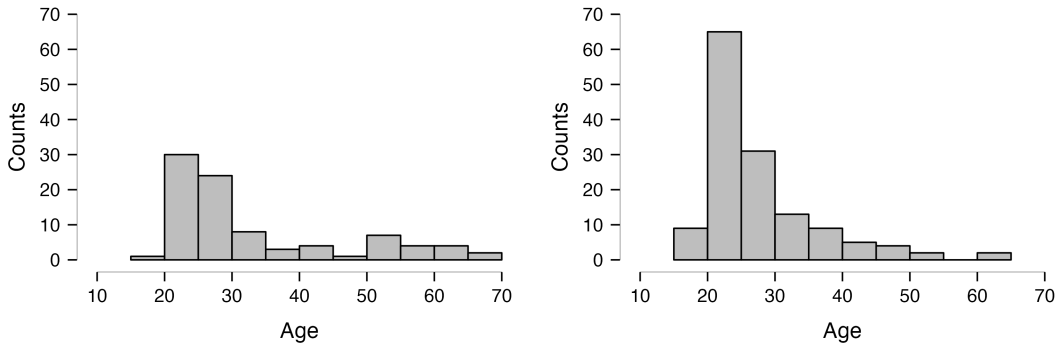
	Respondent group			
	IM		EM	
Prime variants	Before	After	Before	After
NP	45	30	48	47
AP-P	41	29	50	48
AP-A	43	29	47	45

Table 4.2: Frequences of gender per respondent group.

Respondent group	Gender	Frequency	Percent
IM	Female	47	53
	Male	40	45
	Prefer not to say	1	1
EM	Female	87	62
	Male	53	38
	Prefer not to say	0	0

Table 4.3: Descriptive statistics of age per respondent group.

	Respondent group	
	EM	IM
Mean	28	24
Std. Deviation	9	13
Minimum	19	19
Maximum	65	66



(a) Age distribution of intrinsically motivated respondents.

(b) Age distribution of extrinsically motivated respondents.

Figure 4.1: Bar plots illustrating the age distribution of intrinsically and extrinsically motivated respondents.

4.3. Questionnaire Variant

By means of analysing the data of IM and EM respondents separately, we evaluated the impact of the questionnaire variant on *dropout*, *perceived workload*, and *user engagement*. This analysis was performed to answer the first research question: *RQ1) How does achievement priming affect the dropout rate on long surveys?* The three hypotheses associated with these research questions are: *H1A) There is less dropout when respondents are exposed to active achievement priming compared to no priming*; *H1B) There is less dropout when respondents are exposed to passive achievement priming*; and *H1C) Active achievement primes are more effective than passive achievement primes at reducing dropout*.

For continuous variables, we used the analysis of variance (ANOVA) in cases of a normal distribution and employed the Kruskal-Wallis test for non-parametric analysis. Categorical variables were evaluated with the Chi-Squared Test. We separately controlled for the covariates of gender, age, and mood. For ANOVA, we used its covariate counterpart, analysis of covariance (ANCOVA). Since ANCOVA is a robust measure, we also applied it to control for the covariates of non-normal continuous variables. We employed Logistic Regression to control covariates in the case of categorical variables. The quantile-quantile plots (Q-Q plots) used for checking normality can be found in Appendix E.1. The results for the questionnaire variant are summarised in Table 4.4.

Table 4.4: Statistical analysis results of dependent variables for Research Question 1.

The p' is the corrected p-value by means of the post-hoc Bonferroni-Holm method as described in Section 3.6.2.
No measurement had a $p' < 0.0125$, and therefore no significant differences were observed.

Respondent group	Dependent variable	Measurement	Outcome test	p	p'_{Holm}
IM	Dropout	Finished	$\mathrm{X}^2(2) = 0.75$	0.686	1.000
		Duration	$\mathrm{X}^2(2) = 8.52$	0.014	0.252
		Surpassed page	$\mathrm{X}^2(2) = 1.01$	0.603	1.000
	Percieved workload	TLX Sum	$F(2, 51) = 0.41$	0.663	1.000
	User engagement	UES Mean	$F(2, 49) = 0.70$	0.500	1.000
		UES FA	$F(2, 49) = 1.43$	0.249	1.000
		UES PU	$F(2, 49) = 0.52$	0.598	1.000
		UES AE	$F(2, 49) = 0.19$	0.829	1.000
		UES RW	$F(2, 49) = 0.18$	0.825	1.000
EM	Dropout	Finished	$\mathrm{X}^2(2) = 1.97$	0.374	1.000
		Duration	$F(2, 137) = 2.65$	0.744	1.000
		Surpassed page	$\mathrm{X}^2(2) = 1.97$	0.374	1.000
	Percieved workload	TLX Sum	$F(2, 134) = 0.17$	0.847	0.847
	User engagement	UES Mean	$F(2, 134) = 3.25$	0.042	0.672
		UES FA	$F(2, 134) = 0.52$	0.597	1.000
		UES PU	$\mathrm{X}^2(2) = 3.19$	0.203	1.000
		UES AE	$F(2, 134) = 1.91$	0.152	1.000
		UES RW	$F(2, 134) = 3.72$	0.027	0.459

4.3.1. Intrinsic Motivation

Out of the 88 IM participants, 24 participants dropped out of the survey before filling out the TLX post-questionnaire, and 2 more dropped out before filling out the UES-SF post-questionnaire. The number of users considered corresponding to each dependent variable can be found in Table 4.5.

Table 4.5: Amount of intrinsically motivated respondents per dependent variable: dropout, perceived workload, and user engagement.

	Dependent variables		
Prime variants	Dropout	Perceived workload	User engagement
NP	30	18	17
AP-P	29	19	19
AP-A	29	17	16

Dropout

Dropout consists of whether the respondent finished, the time spent on the survey, and the last page of the main questionnaire the respondent completed before leaving (either continuing with the post-questionnaires or dropping out).

The categorical variable finished was analysed using the Chi-Squared Test, which did not reveal any significant effect between the questionnaire variants and whether a respondent completed the survey, $\mathrm{X}^2(2) = 0.75$, $p' = 1.000$. The distribution plots found in Figure 4.2 show a relatively larger, though statistically insignificant, number of respondents finishing the AP-P variant compared to the NP and AP-A variants.

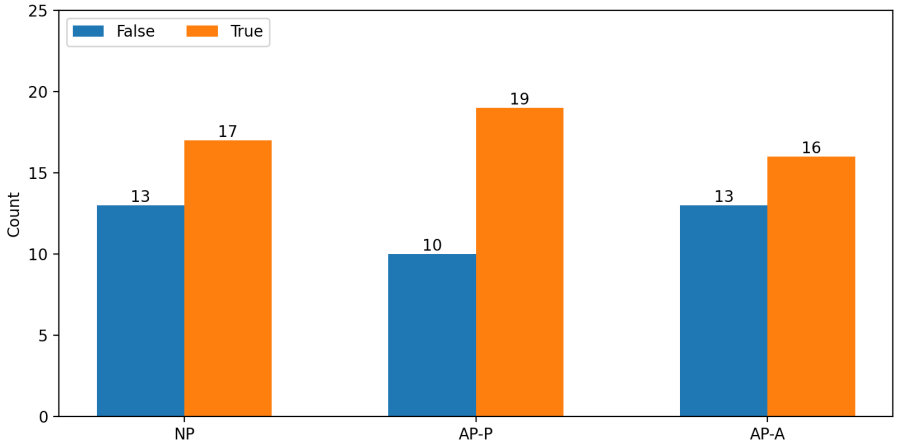


Figure 4.2: Bar plot illustrating the number of intrinsically motivated respondents who finished across the three questionnaire variants.

False means the respondent dropped out during the questionnaire, whereas *True* indicates the respondent finished.

As duration (in seconds) was not normally distributed, Kruskal-Wallis was applied to this variable. The analysis shows a non-significant effect of the NP ($M = 695.90$, $SD = 803.11$), AP-P ($M = 1024.00$, $SD = 753.00$), and AP-A ($M = 1055.52$, $SD = 789.27$) variants on the duration, $\mathrm{X}^2(2) = 8.52$, $p' = 1.000$. The results were also non-significant when controlled for the covariates. The estimation plot as shown in Figure 4.3, shows a slightly higher duration in both AP-P and AP-P variants compared to the NP variant.

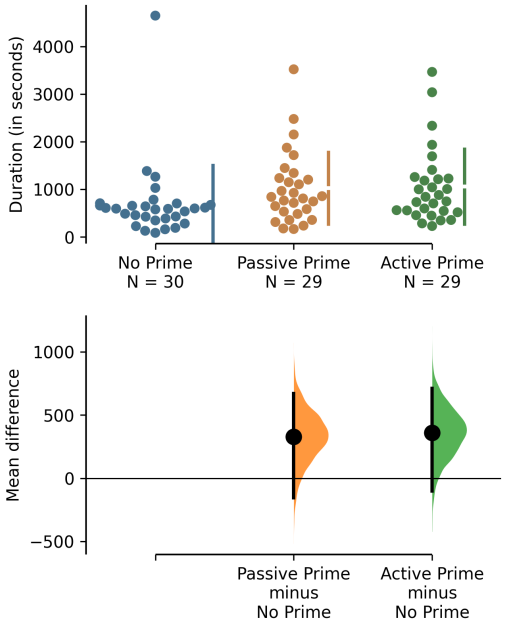


Figure 4.3: Estimation plot illustrating the time spent (in seconds) amongst intrinsically motivated respondents on the survey across the three questionnaire variants.

The variable page surpassed was also not normally distributed. The Kruskal-Wallis analysis demonstrated a non-significant effect of the `NP` ($M = 3.20$, $SD = 1.19$), `AP-P` ($M = 3.45$, $SD = 1.02$), and `AP-A` ($M = 3.45$, $SD = 0.15$) variants on the number of pages surpassed by the respondent, $X^2(2) = 1.01$, $p' = 1.000$. The results remained non-significant when controlled for covariates. However, the distribution plots in Figure 4.4 suggest that respondents in the `NP` group dropped out earlier than those in the `AP-P` and `AP-A` groups, although this difference was not statistically significant.

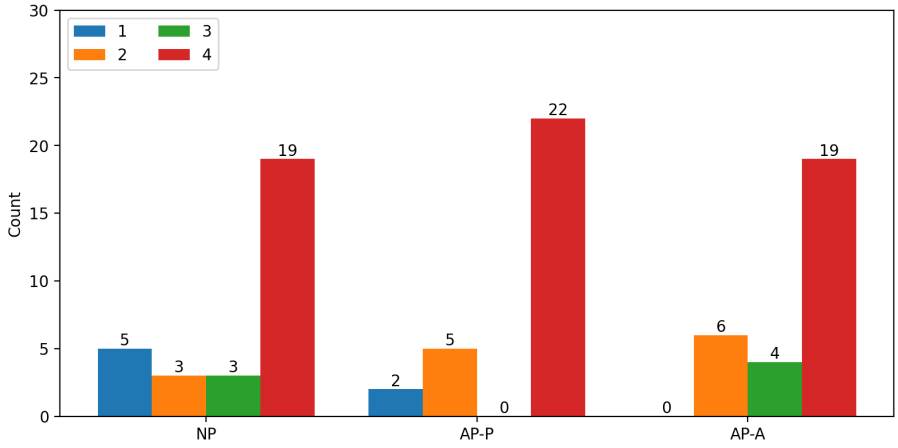


Figure 4.4: Bar plot depicting the last page surpassed by intrinsically motivated respondents before leaving the survey across the three questionnaire variants.

The bar in the legend corresponding to the number 1 portrays the number of respondents having surpassed the first page before dropping out, 2 on the second page, 3 on the third page, and 4 portrays the number of respondents who did not drop out of the main questionnaire.

Perceived Workload

The perceived workload was measured using NASA's TLX, which is a summed score ranging from 6 to 120. The data of TLX followed a normal distribution. The ANOVA indicated no significant effect of the NP ($M = 32.06$, $SD = 23.60$), AP-P ($M = 39.11$, $SD = 19.49$), and AP-A ($M = 36.12$, $SD = 27.48$) variants on the perceived workload, $F(2, 51) = 0.41$, $p' = 1.000$. The estimation plot depicted in Figure 4.5 demonstrates a slightly higher, but statistically insignificant, TLX score for the AP-P variant compared to the NP variant. The standard deviations were similar, although the AP-A variant showed a slightly larger one. Note that a lower TLX score is desirable.

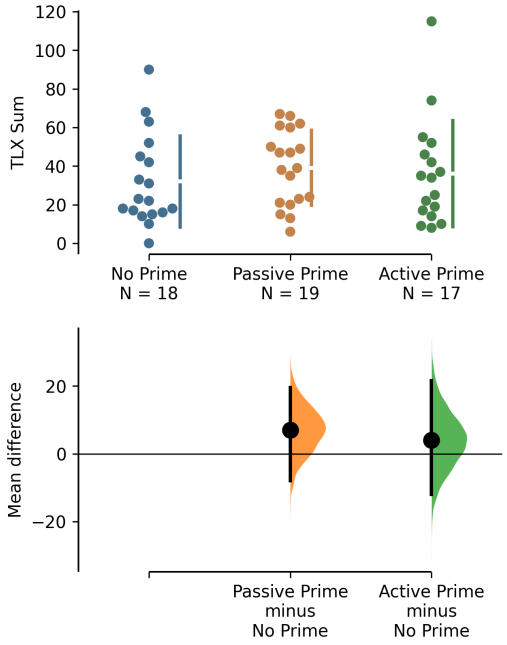


Figure 4.5: Estimation plot illustrating the perceived workload amongst intrinsically motivated respondents across the three questionnaire variants.

User Engagement

User engagement was evaluated using the User Engagement Scale-Short Form (UES-SF). We examined the overall UES Mean and its subscales - Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward (RW) - to determine if the prime variant influenced user engagement. Since the data for all variables conformed to a normal distribution, we employed ANOVA for the analysis.

ANOVA revealed no significant effect of the NP ($M = 3.47$, $SD = 0.51$), AP-P ($M = 3.62$, $SD = 0.53$), and AP-A ($M = 3.73$, $SD = 0.45$) prime variants on UES Mean, $F(2, 134) = 3.25$, $p' = 1.000$. Figure 4.6 depicts the estimation plot for UES Mean. While the NP and AP-P variants scored similarly and had similar standard deviations, AP-A had a slightly lower score with a larger spread, though not significantly.

For FA, the analysis resulted in $F(2, 49) = 1.43$, $p' = 1.000$, for the variants NP ($M = 2.37$, $SD = 0.55$), AP-P ($M = 2.58$, $SD = 0.63$), and AP-A ($M = 2.25$, $SD = 0.56$). For PU, the results were $F(2, 49) = 0.52$, $p' = 1.000$ for NP ($M = 3.57$, $SD = 0.65$), AP-P ($M = 3.60$, $SD = 0.62$), and AP-A ($M = 3.35$, $SD = 0.96$) variants. For AE, the results were $F(2, 49) = 0.19$, $p' = 1.000$ for NP ($M = 2.90$, $SD = 0.60$), AP-P ($M = 2.96$, $SD = 0.71$), and AP-A ($M = 2.81$, $SD = 0.88$). Lastly, for RW, the results were $F(2, 49) = 0.18$, $p' = 1.000$ for NP ($M = 3.18$, $SD = 0.59$), AP-P ($M = 3.11$, $SD = 0.79$), and AP-A ($M = 3.02$, $SD = 0.83$). The estimation plots for the subscales can be found in Figure **??**. The AP-A variant scored the lowest in all subscales, most notably in the PU subscale. AP-P also scored higher in the FA category than the NP and AP-A variants. However, given the insignificance of the results, no generalised conclusions can be drawn.

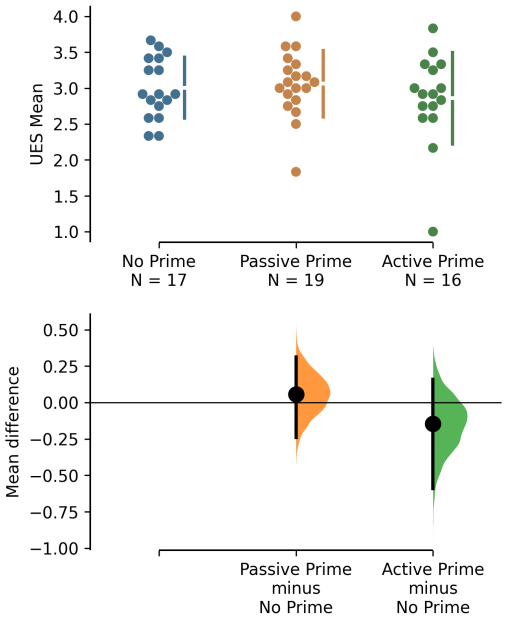
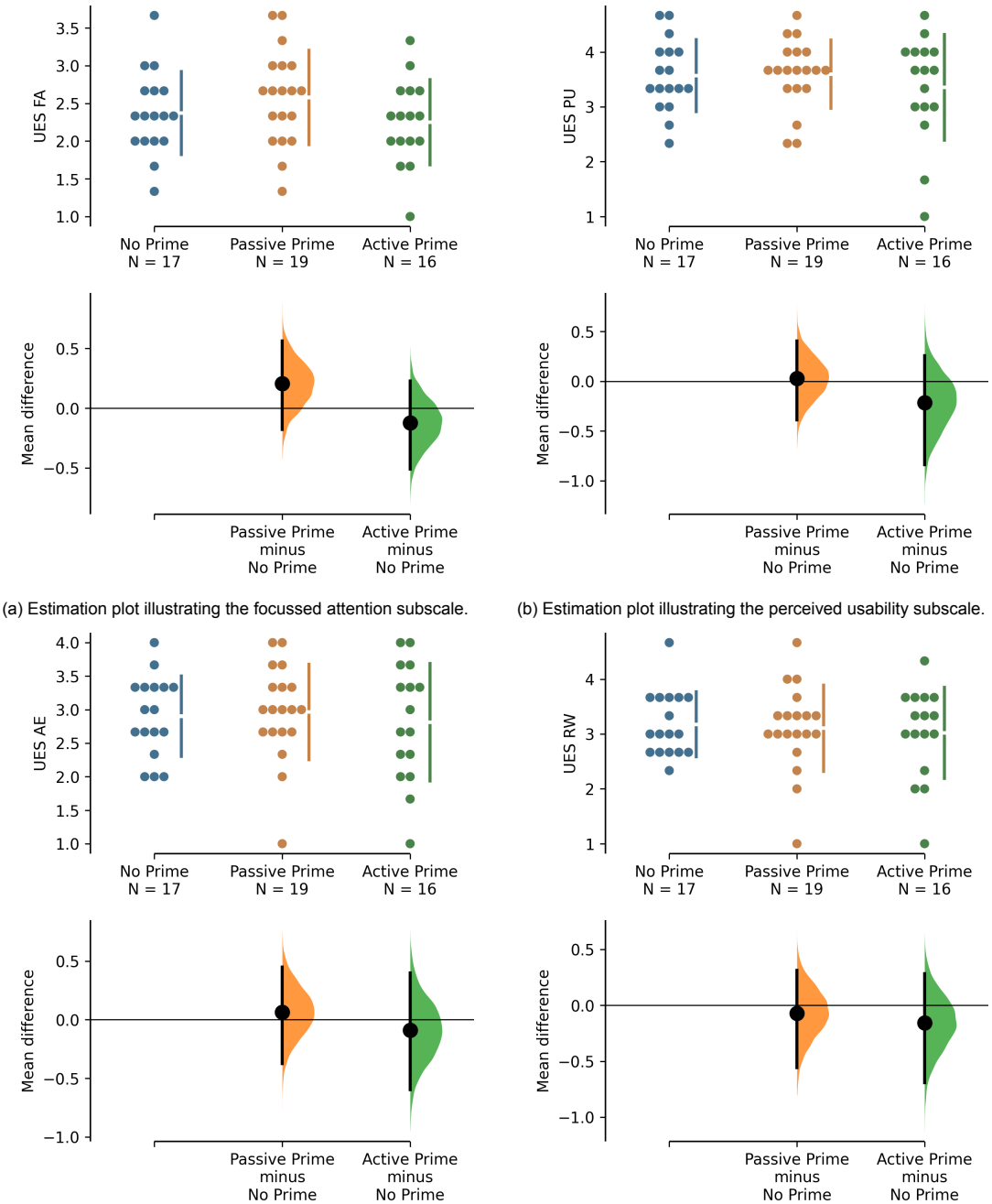


Figure 4.6: Estimation plot illustrating the mean user engagement amongst intrinsically motivated respondents across the three questionnaire variants.



(a) Estimation plot illustrating the focussed attention subscale.

(b) Estimation plot illustrating the perceived usability subscale.

(c) Estimation plot illustrating the aesthetic appeal subscale.

(d) Estimation plot illustrating the reward subscale.

Figure 4.7: Estimation plots illustrating the user engagement subscales amongst intrinsically motivated respondents across the three questionnaire variants.

4.3.2. Extrinsic Motivation

Of the 140 extrinsically motivated participants, 3 dropped out of the survey before filling out the TLX and UES-SF post-questionnaires. The number of users considered for each dependent variable can be found in Table 4.6.

Table 4.6: Amount of extrinsically motivated respondents evaluated per dependent variable: dropout, perceived workload, and user engagement.

	Dependent variables		
Prime variant	Dropout	Perceived workload	User engagement
NP	47	47	47
AP-P	48	46	46
AP-A	45	44	44

Dropout

The categorical variable finished was analysed using the Chi-Squared Test. The test revealed no significant effect between the questionnaire variants and whether a respondent completed the survey, $X^2(2) = 1.97$, $p' = 1.000$. The distribution plots for the data can be found in Figure 4.8. As only three respondents dropped out in total, the distribution plots show very little difference amongst the three questionnaire variants.

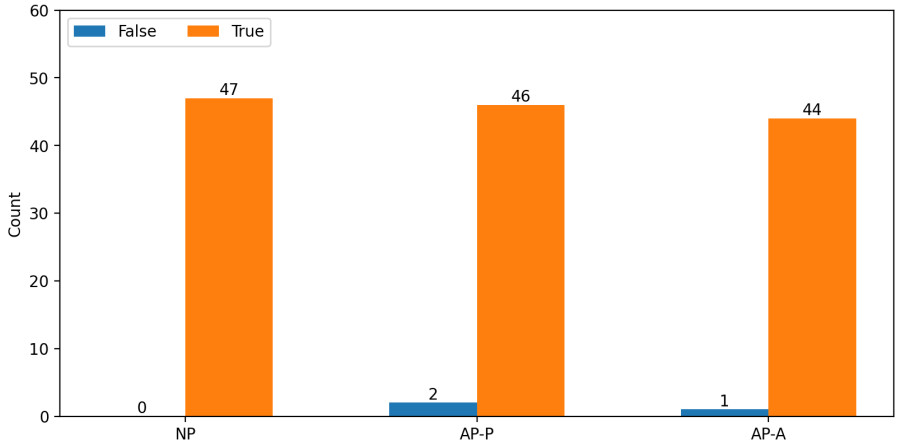


Figure 4.8: Bar plot illustrating the number of extrinsically motivated respondents who finished across the three questionnaire variants.

False means the respondent dropped out during the questionnaire, whereas *True* indicates the respondent finished.

Duration (in seconds) followed a normal distribution for the EM participants. ANOVA revealed a non-significant effect of the NP ($M = 908$, $SD = 416$), AP-P ($M = 1059$, $SD = 509$), and AP-A ($M = 1137$, $SD = 530$) variants on the duration, $F(2, 137) = 2.65$, $p' = 1.000$. The estimation plot for the duration can be found in Figure 4.9. With similar standard deviations, respondents exposed to the NP variant spent the least amount of time on the questionnaire, followed by AP-P , while AP-A variant users took the longest.

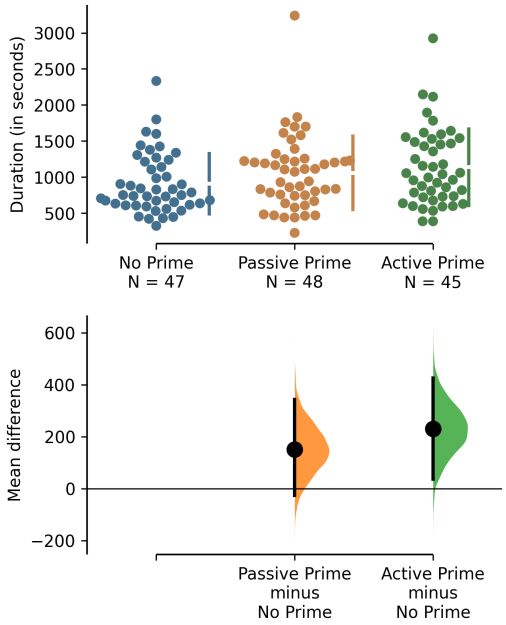


Figure 4.9: Estimation plot illustrating the time spent (in seconds) amongst extrinsically motivated respondents on the survey across the three questionnaire variants.

The variable page surpassed was found not to be normally distributed. The Kruskal-Wallis test indicated a non-significant effect between the NP ($M = 4.00$, $SD = 0.00$), AP−P ($M = 3.94$, $SD = 0.32$), and AP−A ($M = 3.98$, $SD = 1.00$) variants, $\mathrm{X}^2(2) = 1.97$, $p' = 1.000$. The distribution plots in Figure 4.10 confirmed this. Given that only 3 out of 140 participants dropped out, this result is to be expected.

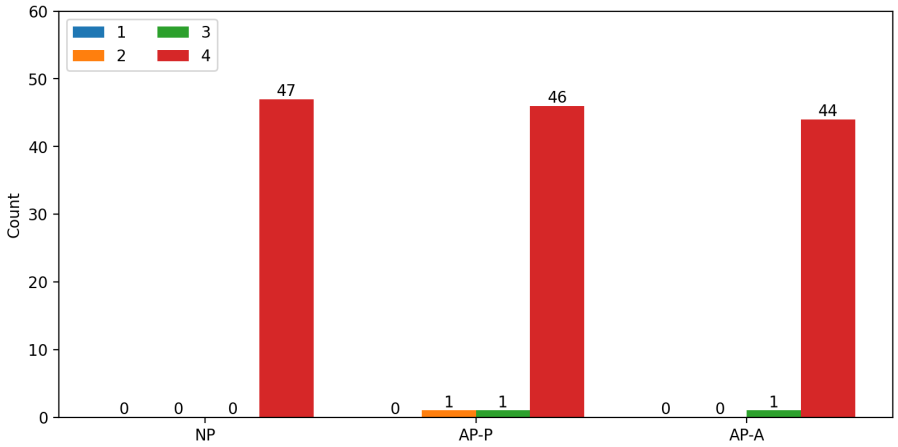


Figure 4.10: Bar plot depicting the last page surpassed by extrinsically motivated respondents before leaving the survey across the three questionnaire variants.

The bar in the legend corresponding to the number 1 portrays the number of respondents having surpassed the first page before dropping out, 2 on the second page, 3 on the third page, and 4 portrays the number of respondents who did not drop out of the main questionnaire.

Perceived Workload

Perceived workload, measured by means of NASA's TLX, followed a normal distribution. ANOVA test revealed a non-significant effect of the NP ($M = 30.14$, $SD = 18.43$), AP−P ($M = 27.85$, $SD = 18.29$), and AP−A ($M = 28.80$, $SD = 20.73$) variants on the perceived workload, with $F(2, 134) = 0.17$, $p' = 0.847$. The estimation plot, displayed in Figure 4.11, showed little difference in perceived workload amongst the prime variants.

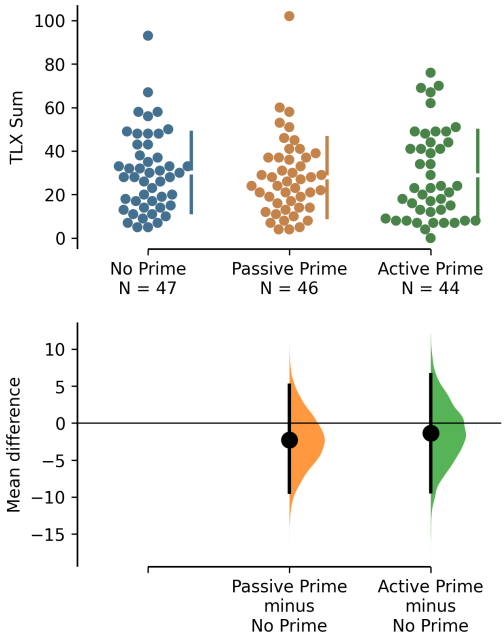


Figure 4.11: Estimation plot illustrating the perceived workload amongst extrinsically motivated respondents across the three questionnaire variants.

User Engagement

User engagement was measured using the UES Mean and its subscales: FA, PU, AE, and RW. The UES Mean exhibited a normal distribution, and the ANOVA test showed no significant effect of the `NP` ($M = 3.47$, $SD = 0.51$), `AP-P` ($M = 3.62$, $SD = 0.53$), and `AP-A` ($M = 3.73$, $SD = 0.45$) prime variants on the user engagement, with $F(2, 134) = 3.25$, $p' = 0.672$. Though the effect was insignificant, the estimation plot shown in Figure 4.12 reveals that the `NP` scored lowest, followed by the `AP-P` variant, with the `AP-A` variant achieving the highest score.

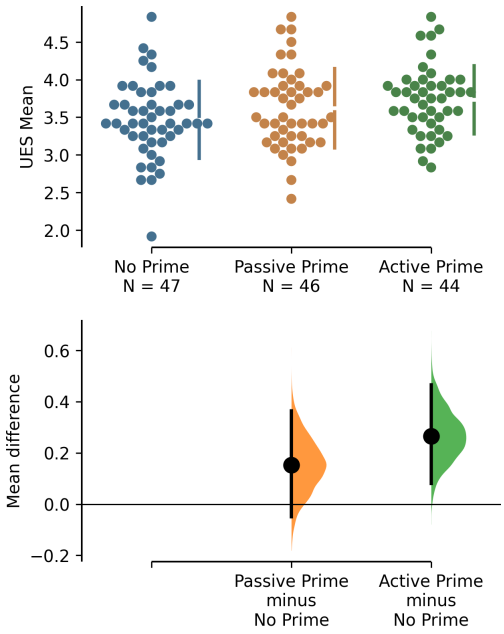
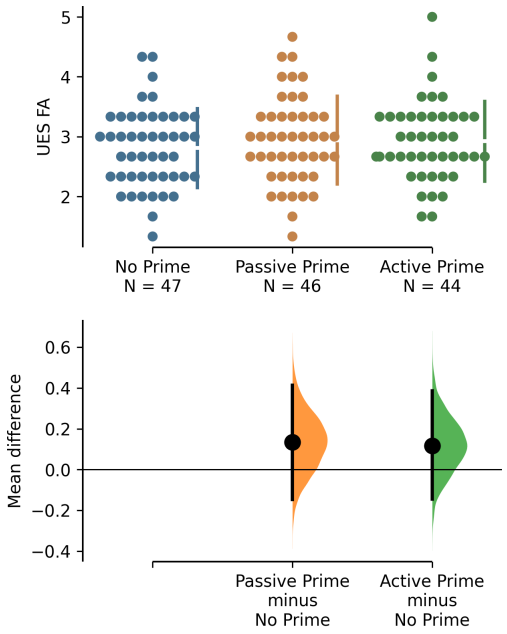
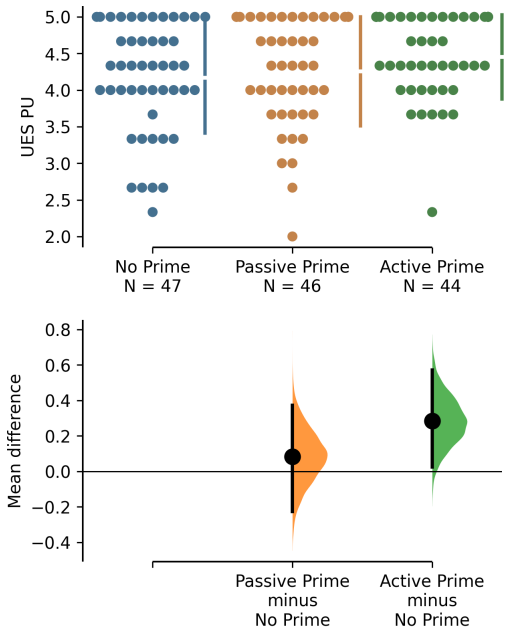


Figure 4.12: Estimation plot illustrating the mean user engagement amongst extrinsically motivated respondents across the three questionnaire variants.

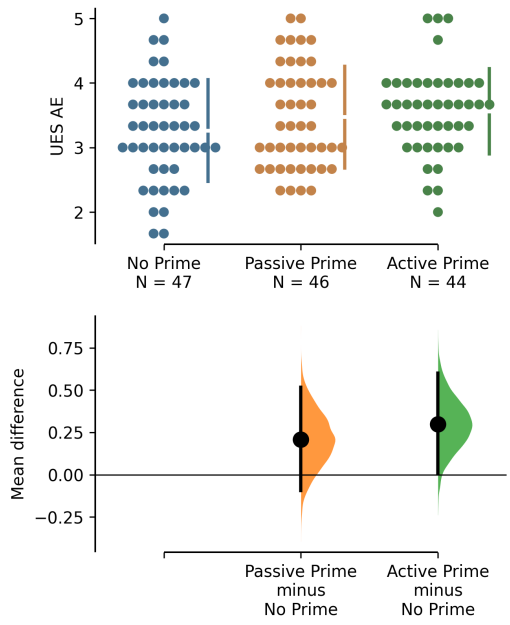
When investigating the subscales, all subscales except for PU showed a normal distribution, yet the effect was non-significant. For FA, ANOVA resulted in an $F(2, 134) = 0.52$, $p' = 1.000$ for the variants `NP` ($M = 2.81$, $SD = 0.66$), `AP-P` ($M = 2.94$, $SD = 0.73$), and `AP-A` ($M = 2.92$, $SD = 0.67$). For PU, the Kruskal-Wallis test gave $X^2(2) = 3.19$, $p' = 1.000$ for the `NP` ($M = 4.17$, $SD = 0.76$), `AP-P` ($M = 4.25$, $SD = 0.74$), and `AP-A` ($M = 4.45$, $SD = 0.58$) variants. For AE, ANOVA resulted in $F(2, 134) = 1.91$, $p' = 1.000$ for `NP` ($M = 3.26$, $SD = 0.79$), `AP-P` ($M = 3.47$, $SD = 0.79$), and `AP-A` ($M = 3.56$, $SD = 0.70$) prime variants. Lastly, for RW, $F(2, 134) = 3.72$ and $p' = 0.0459$ were calculated for `NP` ($M = 3.62$, $SD = 0.64$), `AP-P` ($M = 3.81$, $SD = 0.65$), and `AP-A` ($M = 3.98$, $SD = 0.61$). The estimation plots of the subscales can be found in Figure 4.13. The `NP` variant scored the lowest in all subscales. For FA, both AP variants scored similarly. However, for all other subscales, `AP-A` scored higher than both the `NP` and `AP-P` variants.



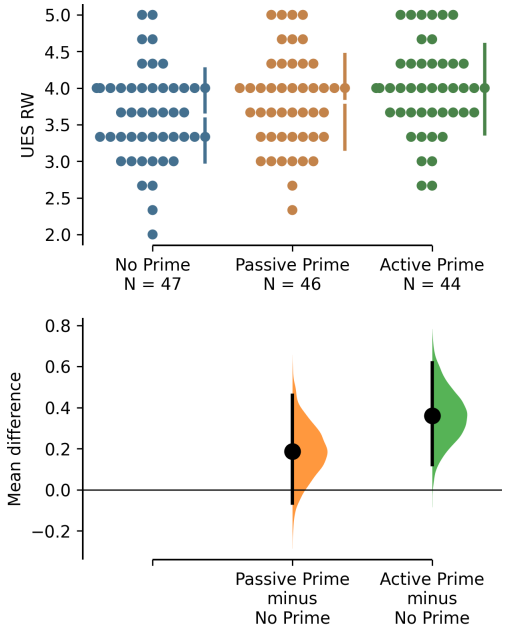
(a) Estimation plot illustrating the focussed attention subscale.



(b) Estimation plot illustrating the perceived usability subscale.



(c) Estimation plot illustrating the aesthetic appeal subscale.



(d) Estimation plot illustrating the reward subscale.

Figure 4.13: Estimation plots illustrating the user engagement subscales amongst extrinsically motivated respondents across the three questionnaire variants.

4.3.3. Summary

Our data analysis reveals no significant effect of the achievement prime on any of the dependent variables for both intrinsically and extrinsically motivated respondents. These finding prohibit the verification of Hypotheses *H1A*, *H1B*, and *H1C* related to *RQ1*. Nevertheless, patterns were observed in the dependent variables for both respondent groups. These descriptions should be interpreted cautiously as the differences are statistically non-significant.

Amongst the intrinsically motivated respondents, the AP-P variant of the questionnaire had the highest completion rates, with the NP and AP-A variants having comparable completion rates. The AP-P variant, despite having the lowest dropout rate, had dropouts occur earlier compared to the NP and AP-A variants. Both AP variants required more time than the NP variant.

On the other hand, nearly all EM respondents completed the survey, regardless of the variant they were presented with, resulting in minor differences in dropout. Regarding duration, NP respondents spent the least time on the survey, followed by the AP-P respondents, with the AP-A variant being the most time-consuming.

Therefore, EM respondents showed similar behaviour across all questionnaire variants. Including achievement primes might have increased the cognitive load, leading to a longer survey completion time. The AP-P variant yielded fewer dropouts for intrinsically motivated respondents than the AP-A and NP variants, despite requiring additional time.

There were contrasting trends between the respondent groups in the context of perceived workload. The NP variant yielded the lowest workload for intrinsically motivated respondents, while it was highest for the extrinsically motivated group. The AP-P variant showed the lowest workload for EM respondents and the highest for IM . The AP-A variant consistently performed in the middle for both groups. Even though the mean scores vary, the differences are minor, with the best and worst-performing variants differing by just 7 points out of a possible 120, whereas the standard deviation of around 30 points was relatively large.

Regarding user engagement, the AP-P variant recorded the highest scores for IM respondents, while the AP-A variant registered the lowest for the mean and all subscales except for RW, where NP and AP-P had similar scores. In contrast, for extrinsically motivated respondents, the AP-A variant scored highest in all subscales. The AP-P variant scored second highest, except for the FA subscale, where AP-P and AP-A achieved similar scores.

In this study, the AP-P variant appears to be the most effective for IM respondents, whereas the AP-A variant performs best for EM respondents.

4.4. Respondent Group

We analysed the data of the AP-P and AP-A variations of the questionnaire separately to understand whether there were any differences across the respondent groups. Doing so helped us answer the second research question: *RQ2) How do achievement primes influence respondents with different motivations to participate?* The associated hypothesis is: *H2) People are less affected by primes when they participate based on intrinsic motivation than when they participate because of extrinsic motivation*.

Categorical variables were examined using Fisher's Exact Test, while the continuous variables were evaluated using the Independent Samples t-Test in case of a normal distribution and the Mann-Whitney U test otherwise. Normality was determined by means of the Shapiro-Wilk test for normality, the result of which can be found in Appendix E.2.

The covariates were controlled for separately, using Logistic Regression for categorical variables and ANCOVA for continuous variables. We only report instances where controlling for categorical variables led to a non-significant result becoming significant. The results of these analyses are summarised in Table 4.7. For a result to be considered significant, the adjusted p'-value must be less than 0.0125.

Table 4.7: Statistical analysis results of dependent variables for Research Question 2.

The p' is the corrected p-value by means of the post-hoc Bonferroni-Holm method as described in Section 3.6.2.

The measurements in bold have a $p' < 0.0125$, and therefore show a significant difference.

* When controlling for the covariate *Age*, the measurement *Surpassed page* has a $p < 0.001$ and $p' = 0.001$ which would make the result significant. Controlling *Surpassed page* for the covariate *Age* does not change the significance of other measurements.

Prime variant	Dependent variable	Measurement	Outcome test	p	p'
AP-P	Dropout	**Finished**		**< 0.001**	**0.008**
		Duration	$U = 600.50$	0.318	0.318
		Surpassed page*	$\boldsymbol{U = 552.50}$	**0.007**	**0.049**
	Percieved workload	TLX Sum	$t(63) = 2.21$	0.024	0.146
	User engagement	**UES Mean**	$\boldsymbol{t(63) = -4.02}$	**< 0.001**	**0.003**
		UES FA	$t(63) = -1.89$	0.063	0.252
		UES PU	$\boldsymbol{U = 516}$	**< 0.001**	**0.007**
		UES AE	$U = 516$	0.054	0.271
		UES RW	$\boldsymbol{t(63) = -3.76}$	**< 0.001**	**0.005**
AP-A	Dropout	**Finished**		**< 0.001**	**< 0.001**
		Duration	$U = 516.00$	0.132	0.396
		Surpassed page	$\boldsymbol{U = 439.00}$	**< 0.001**	**0.001**
	Percieved workload	TLX Sum	$U = 439.00$	0.299	0.598
	User engagement	**UES Mean**	$\boldsymbol{t(58) = -5.91}$	**< 0.001**	**< 0.001**
		UES FA	$\boldsymbol{t(58) = -3.60}$	**< 0.001**	**0.008**
		UES PU	$\boldsymbol{U = 94}$	**< 0.001**	**< 0.001**
		UES AE	$\boldsymbol{t(58) = -3.54}$	**< 0.001**	**0.007**
		UES RW	$\boldsymbol{t(58) = -4.90}$	**< 0.001**	**< 0.001**

4.4.1. Passive Prime

In the group exposed to the AP-P variant, out of 77 participants, 12 did not complete the TLX and UES-SF post-questionnaires. The breakdown of respondents considered for each dependent variable is documented in Table 4.8.

Table 4.8: Amount of respondents shown the passive achievement prime questionnaire variation per dependent variable: dropout, perceived workload, and user engagement.

	Dependent variables		
Respondent group	Dropout	Perceived workload	User engagement
IM	29	19	19
EM	48	46	46

Dropout

Like with the first research question, we captured dropout by means of the variables finished, duration, and the last surpassed page before leaving the survey.

Fisher's Exact test was used to determine if respondent motivation affected completion rates. The results demonstrated a significant difference between IM and EM groups with a $p' = 0.008$. The distribution plot can be seen in Figure 4.14. This suggests that EM respondents were significantly more likely to complete a questionnaire when passive primes were used compared to IM respondents.

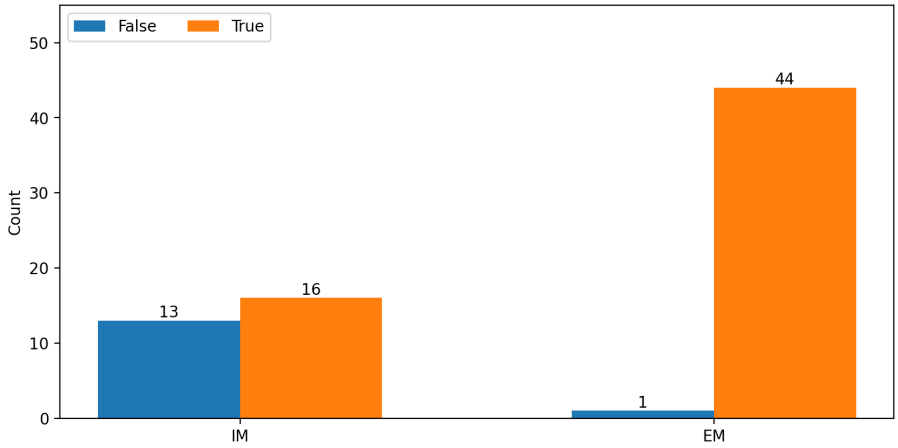


Figure 4.14: Bar plot illustrating the number of respondents who finished the passive achievement prime variant of the questionnaire.

False means the respondent dropped out during the questionnaire, whereas *True* indicates the respondent finished.

Duration (in seconds) was not normally distributed, and therefore, we used the Mann-Whitney U test for analysis. No significant impact of respondent motivation on duration was observed for IM ($M = 1024$, $SD = 753$) and EM ($M = 1059$, $SD = 509$) groups, with $U = 600.50$, $p' = 0.318$. As seen in Figure 4.15, the mean survey duration for both IM and EM respondents was just above 17 minutes. There is an important distinction in standard deviation, which was larger for IM respondents, potentially due to the smaller sample size.

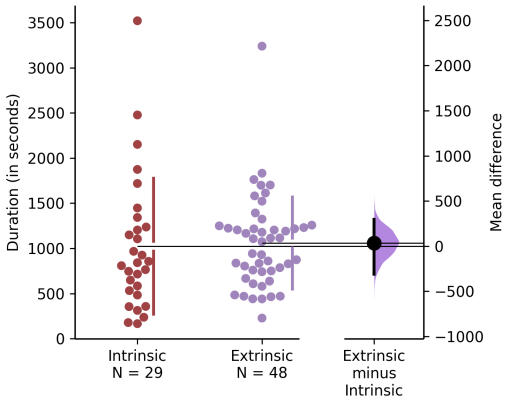


Figure 4.15: Estimation plot illustrating the time spent (in seconds) amongst respondents who were shown the passive achievement prime variant of the questionnaire.

The surpassed page data was also non-normally distributed. The Mann-Whitney U test revealed no significant difference between IM ($M = 1024$, $SD = 753$) and EM ($M = 1059$, $SD = 509$) groups, with $U = 552.50$, $p' = 0.049$. However, when controlling for the covariate *Age* by means of ANCOVA, a significant difference appeared between the groups, with $F(1, 72) = 16.02$, $p' = 0.001$. Controlling for other covariates, *Gender* and *Mood*, did not result in a significant difference. Importantly, controlling for *Age* did not affect the significance of other measurements. The distribution plot without controlling for *Age* can be found in Figure 4.16, while the flex plot with age control is in Figure 4.17. Dropouts were aged between 20 and 30 in both groups, yet, more IM respondents in this age range exit at earlier pages compared to their EM counterparts.

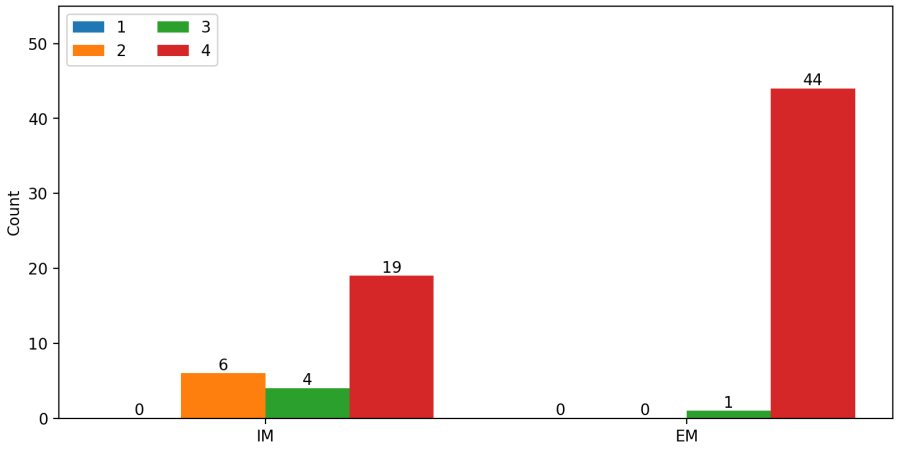
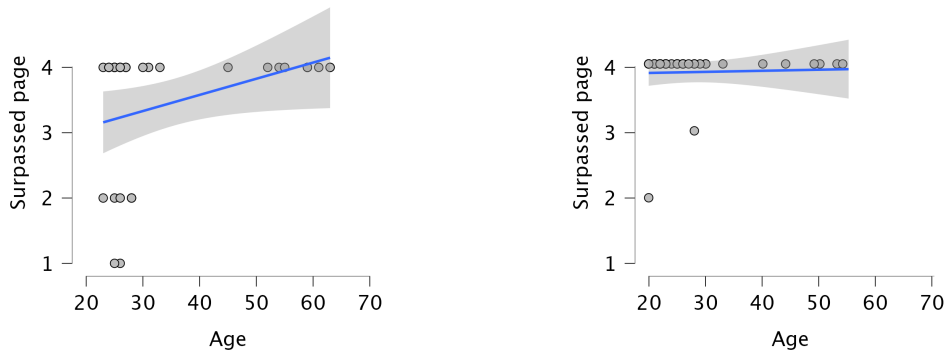


Figure 4.16: Bar plot illustrating the last page surpassed by respondents who were shown the passive achievement prime variant of the questionnaire.

The bar in the legend corresponding to the number 1 portrays the number of respondents having surpassed the first page before dropping out, 2 on the second page, 3 on the third page, and 4 portrays the number of respondents who did not drop out of the main questionnaire.



(a) Flex plot for the page surpassed by intrinsically motivated respondents shows the passive achievement prime variant controlled for age.

(b) Flex plot for the page surpassed by extrinsically motivated respondents shows the passive achievement prime variant controlled for age.

Figure 4.17: Flex plots illustrating the last page surpassed by respondents who were shown the passive achievement prime variant of the questionnaire controlled for age.

Perceived Workload

The perceived workload was evaluated using the TLX score ranging from 6 to 120, the data for which was normally distributed. The Independent Samples t-Test did not reveal a significant difference between IM ($M = 39.11$, $SD = 19.49$) and EM ($M = 27.85$, $SD = 18.29$) respondents, with $t(63) = 2.21$, $p' = 0.146$. However, the estimation plot in Figure 4.18 shows that IM respondents perceived the workload to be more than 10 points (out of 120) higher than EM respondents.

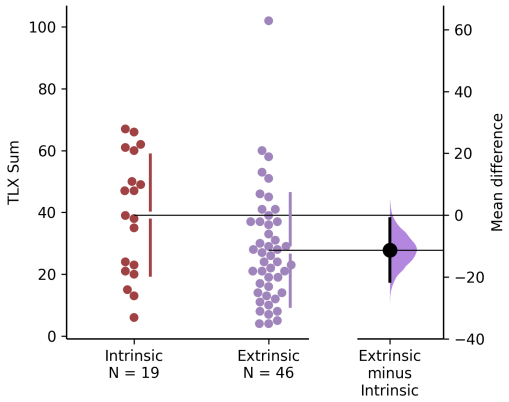


Figure 4.18: Estimation plot illustrating the perceived workload amongst respondents who were shown the passive achievement prime variant of the questionnaire.

User Engagement

Again, user engagement was measured by means of the UES-SF, and we analysed UES Mean, FA, PU, AE, and RW.

UES Mean was normally distributed, and the Independent Samples t-Test demonstrated a significant difference in overall user engagement between IM ($M = 3.06$, $SD = 0.46$) and EM ($M = 3.62$, $SD = 0.53$) respondents, $t(63) = -4.02$, $p' = 0.003$. As shown in Figure 4.19, IM respondents are significantly less engaged than EM respondents, which might be due to them being less susceptible to the primes.

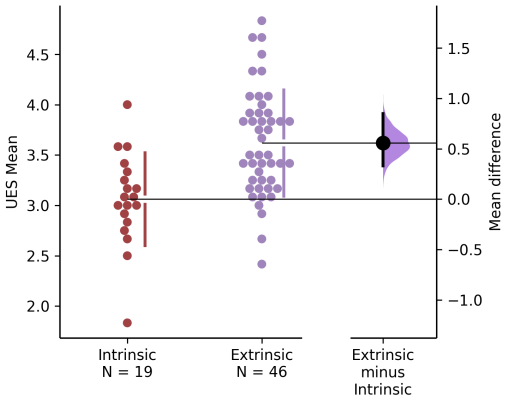
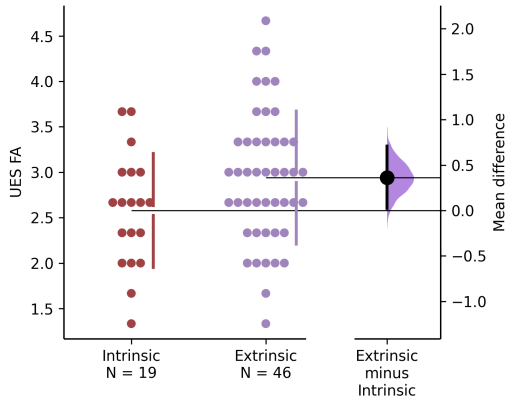
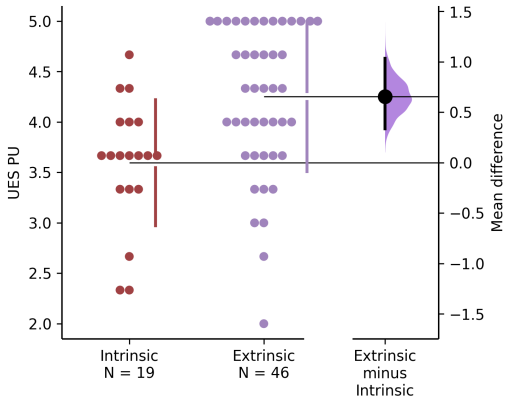


Figure 4.19: Estimation plot illustrating the mean user engagement amongst respondents who were shown the passive achievement prime variant of the questionnaire.

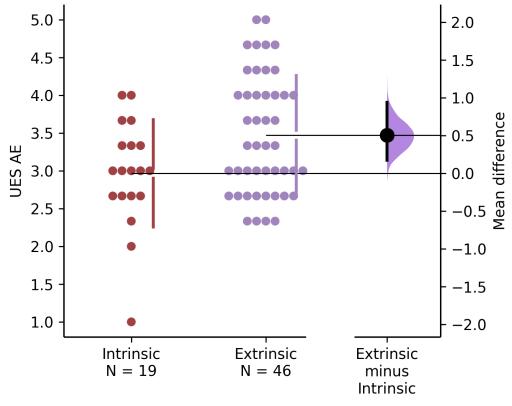
FA and RW data were normally distributed, whereas PU and AE were not. The Independent Samples t-Test for FA did not show a significant effect between IM ($M = 2.58$, $SD = 0.63$) and EM ($M = 2.94$, $SD = 0.73$) respondents, with $t(63) = -1.89$, $p' = 0.25$. The Mann-Whitney U test indicated a significant effect for PU between IM ($M = 3.60$, $SD = 0.62$) and EM ($M = 4.25$, $SD = 0.74$) respondents, with $U = 207.50$, $p' = 0.007$. However, the Mann-Whitney U test did not reveal a significant effect for AE between IM ($M = 2.96$, $SD = 0.71$) and EM ($M = 3.47$, $SD = 0.79$) respondents, with $U = 304.50$, $p' = 0.271$. Finally, the Independent Samples t-Test for RW showed a significant effect between IM ($M = 3.11$, $SD = 0.79$) and EM ($M = 3.12$, $SD = 0.65$) respondents, with $t(63) = -3.76$, $p' = 0.005$. Figure 4.20 shows the estimation plots for the UES subscales. Despite only PU and RW being significant, all subscales showed a difference between IM and EM respondents, with IM respondents scoring lower than EM respondents.



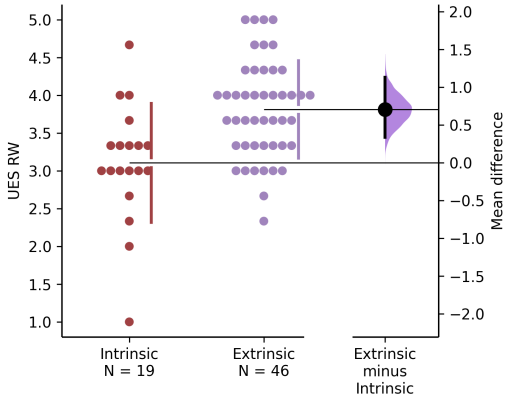
(a) Estimation plot illustrating the focused attention subscale.



(b) Estimation plot illustrating the perceived usability subscale.



(c) Estimation plot illustrating the aesthetic appeal subscale.



(d) Estimation plot illustrating the reward subscale.

Figure 4.20: Estimation plots illustrating the user engagement subscales amongst respondents who were shown the passive achievement prime variant of the questionnaire.

4.4.2. Active Prime

Out of the 74 participants who were shown the AP−A variant, 13 participants did not complete the TLX, and one additional participant dropped out of UES-SF post-questionnaires. The number of users considered for each dependent variable can be found in Table 4.8.

Table 4.9: Amount of respondents shown the active achievement prime questionnaire variation per dependent variable: dropout, perceived workload, and user engagement.

	Dependent variables		
Respondent group	Dropout	Perceived workload	User engagement
IM	29	17	16
EM	45	44	44

Dropout

As with the AP−P variant, we looked at finished, duration of filling out the survey, and the page surpassed before leaving the survey to capture dropout.

Fisher's Exact test was used to analyse whether the motivation of a respondent affected whether they completed the survey or not. The results showed a significant difference between respondent groups, with $p' < 0.001$. The distribution plot is provided in Figure 4.21. EM respondents were far more likely to finish a questionnaire when active primes were introduced than IM respondents.

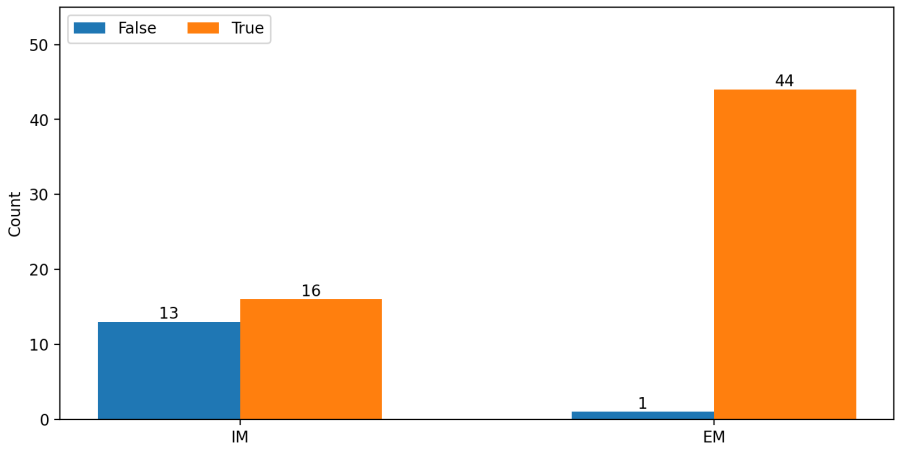


Figure 4.21: Bar plot illustrating the number of respondents who finished the active achievement prime variant of the questionnaire.

The bar in the legend corresponding to the number 1 portrays the number of respondents having surpassed the first page before dropping out, 2 on the second page, 3 on the third page, and 4 portrays the number of respondents who did not drop out of the main questionnaire.

Duration (in seconds) was not normally distributed, so the Mann-Whitney U test was used for analysis. The results indicated no significant effect of the respondent group on the duration for IM ($M = 1055$, $SD = 789$) and EM ($M = 1137$, $SD = 530$) respondent groups, $U = 516$, $p' = 0.396$. This is visualized in the estimation plot in Figure 4.22. For both EM and IM respondents, the average duration of the survey was slightly above 17 minutes. The latter group did have a larger standard deviation, which could be attributed to the smaller number of participants in that group.

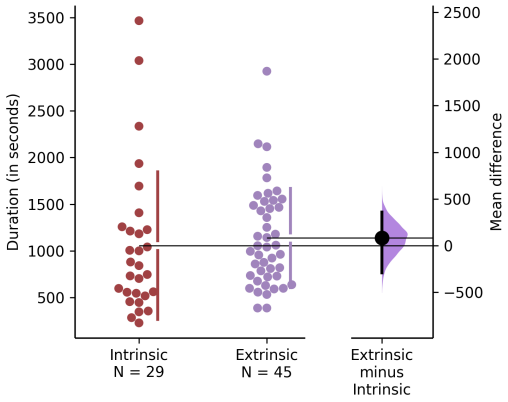


Figure 4.22: Estimation plot illustrating the time spent (in seconds) amongst respondents who were shown the active achievement prime variant of the questionnaire.

The data for the surpassed page was also not normally distributed. The Mann-Whitney U test reveals a significant difference between IM ($M = 3.45$, $SD = 0.83$) and EM ($M = 3.98$, $SD = 0.15$) respondent groups, with $U = 439.00$, $p' = 0.001$. The estimation plot in Figure 4.23 shows that the relative difference in the number of participants who reached pages 2, 3, and 4 was smaller for IM respondents than for EM respondents. Almost all EM respondents managed to surpass page 4.

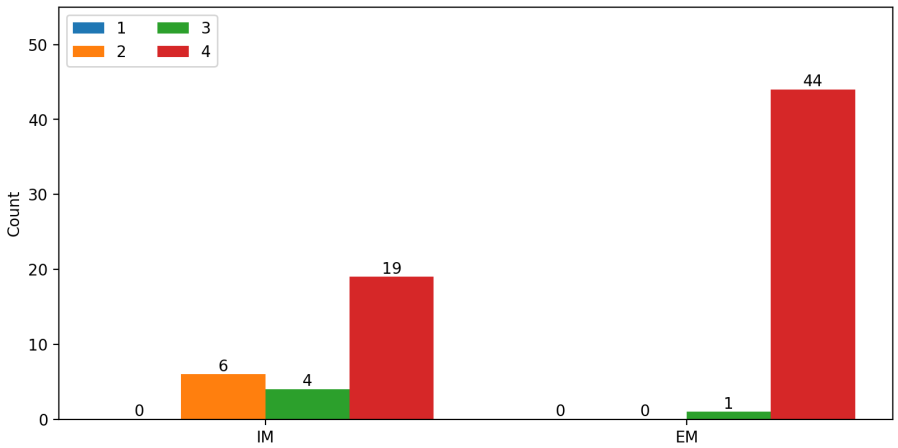


Figure 4.23: Bar plot illustrating the last page surpassed by respondents who were shown the active achievement prime variant of the questionnaire.

The bar in the legend corresponding to the number 1 portrays the number of respondents having surpassed the first page before dropping out, 2 on the second page, 3 on the third page, and 4 portrays the number of respondents who did not drop out of the main questionnaire.

Perceived Workload

The perceived workload, as determined by the TLX score ranging from 6 to 120, was not normally distributed. The Mann-Whitney U test revealed no significant difference $U = 439.00$, $p' = 0.598$, between the intrinsically motivated (IM , $M = 36.12$, $SD = 27.48$) and extrinsically motivated (EM , $M = 28.80$, $SD = 20.73$) respondent groups. However, the estimation plot in Figure 4.24 suggests that IM respondents perceived the workload to be roughly 10 points (out of 120) higher than the EM respondents.

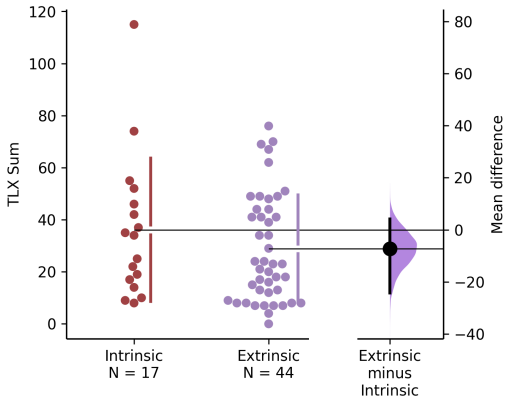


Figure 4.24: Estimation plot illustrating the perceived workload amongst respondents who were shown the active achievement prime variant of the questionnaire.

User Engagement

Again, user engagement was measured by the answers of the UES-SF, where we analysed UES Mean, FA, PU, AE, and RW.

UES Mean was normally distributed, and the Independent Samples t-Test showed a significant difference in overall user engagement between IM ($M = 2.86$, $SD = 0.64$) and EM ($M = 3.73$, $SD = 0.45$) respondent groups, $t(58) = -5.90$, $p' < 0.001$. Figure 4.25 indicates that IM respondents were significantly less engaged than EM respondents, which might be attributed to the prime variant.

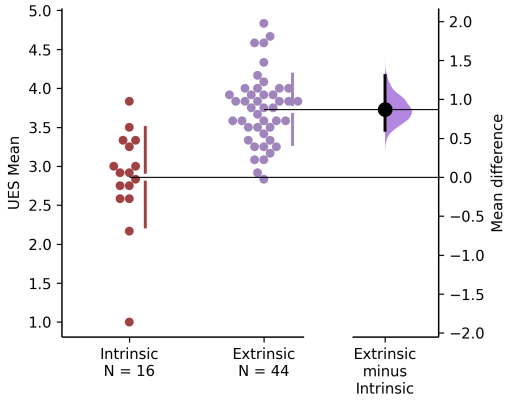
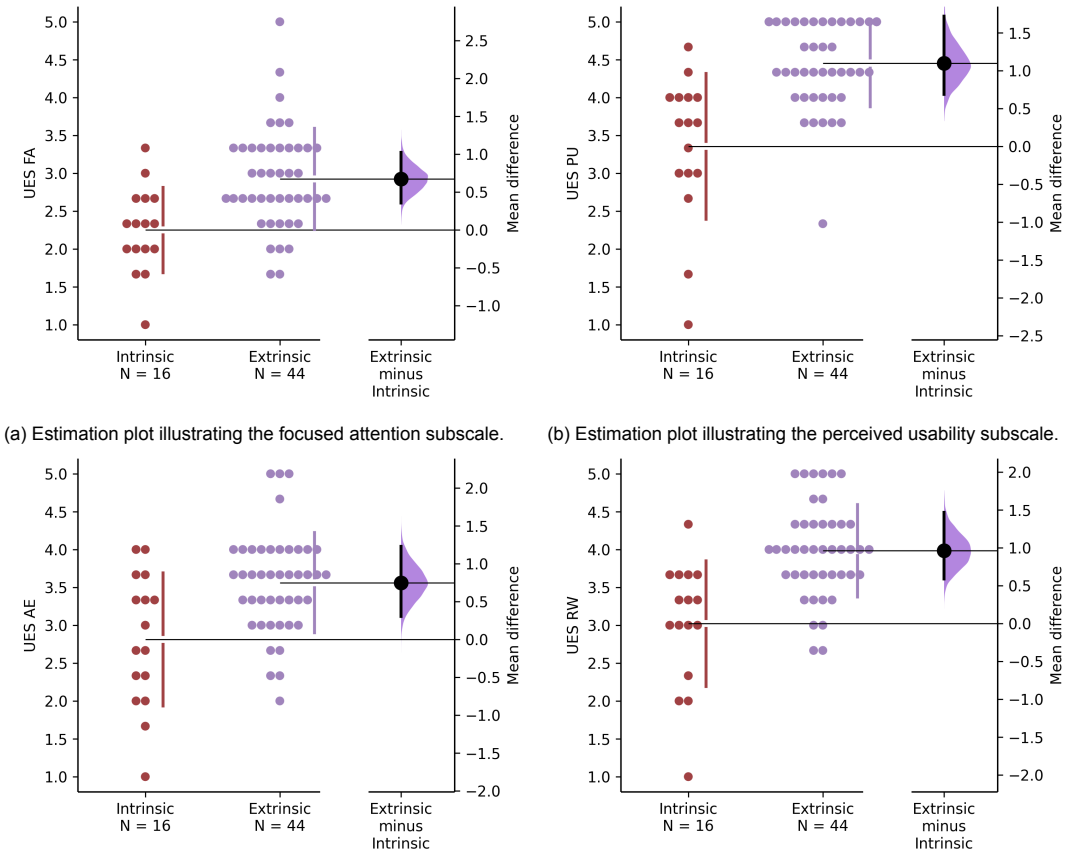


Figure 4.25: Estimation plot illustrating the mean user engagement amongst respondents who were shown the active achievement prime variant of the questionnaire.

All subscales except for PU are normally distributed. The Independent Samples t-Test applied to FA showed a significant effect, $t(58) = -3.60$, $p' = 0.008$, between IM ($M = 2.25$, $SD = 0.56$) and EM ($M = 2.92$, $SD = 0.67$) respondent groups. The Mann-Whitney U test suggested a significant effect for PU between IM ($M = 3.35$, $SD = 0.96$) and EM ($M = 4.45$, $SD = 0.58$) respondent groups with $U = 207.50$, $p' = 0.007$. Furthermore, the Independent Samples t-Test demonstrated a significant effect for AE, $t(58) = -3.54$, $p' = 0.007$, between IM ($M = 2.81$, $SD = 0.88$) and EM ($M = 3.56$, $SD = 0.66$) respondent groups. Finally, the Independent Samples t-Test also indicated a significant effect for RW, $t(58) = -4.90$, $p' < 0.001$, between IM ($M = 3.02$, $SD = 0.83$) and EM ($M = 3.98$, $SD = 0.61$) respondent groups. Figure 4.26 shows the estimation plots for the UES subscales. All subscales revealed a substantial difference between IM and EM respondents, with the IM group scoring lower than the EM group.



(a) Estimation plot illustrating the focused attention subscale.

(b) Estimation plot illustrating the perceived usability subscale.

(c) Estimation plot illustrating the aesthetic appeal subscale.

(d) Estimation plot illustrating the reward subscale.

Figure 4.26: Estimation plots illustrating the user engagement subscales amongst respondents who were shown the active achievement prime variant of the questionnaire.

4.4.3. Summary

The results of our data analysis revealed a significant effect on dropout rates and engagement levels between IM and EM respondents. The data clearly suggests that the EM respondents had a lower dropout rate and a higher engagement level than the IM respondents. While there was a higher perceived workload amongst IM respondents, this effect was not statistically significant. These results allow us to accept *H2* and to provide an answer to *RQ2*.

Regarding dropout, almost all EM respondents completed the entire survey. On the other hand, many of the IM respondents dropped out, regardless of the prime variant. It was also noted that there was only a slight and insignificant difference in the time spent on the survey between the two groups, with IM respondents spending slightly less time. This could potentially be attributed to their higher dropout rate, which would naturally result in less time spent on the survey. Looking at the pages surpassed, it is clear that a significantly smaller proportion of IM respondents managed to get past the fourth page compared to their EM counterparts.

Although the effect was not significant in terms of perceived workload, EM respondents reported a lower perceived workload than the IM respondents for both active and passive prime variants.

When examining user engagement, the data clearly shows that extrinsically motivated respondents had higher levels of engagement. Notably, significant effects were observed in PU and RW for both respondent groups. However, the effects on FA and AE were only significant for the active prime variant and not for the passive variant.

4.5. Exploratory Findings

These findings explore the observations we made that we did not account for in the experimental setup. The observations might provide insightful information to be used in future research.

Amongst both IM and EM respondents, there were several instances of incorrect responses to the prime questions in the AP-P variant of the questionnaire. Specifically, one person from the EM participants answered three questions incorrectly, and five people provided one incorrect answer each. From the IM group, three participants each gave one incorrect answer.

This analysis excluded responses from the EM participant who incorrectly answered three questions. We attributed these errors to potential factors such as inattentiveness or limited proficiency in English rather than a flaw in the theory on achievement primes or the implementation of the experiment.

The incorrect answers were observed across several quotes:

PP5: "I never said it would be *easy*, I only said it would be worth it." — Mae West

One participant out of 12 incorrectly replaced *easy* with *hesitant* instead of the correct synonym, *effortless*.

PP7: "The way I see it, if you want the rainbow, you gotta put up with the *rain*!" — Dolly Parton

Two participants out of 12 incorrectly replaced *rain* with *distribution* instead of the correct synonym, *mist*.

PP11: "You are not obligated to win. You're obligated to keep trying. To the *best* you can do every day." — Jason Mraz

There was one participant out of 13 who replaced *best* with *situation* instead of the correct synonym, *finest*.

PP20: "*Life* isn't about finding yourself. *Life* is about creating yourself." — George Bernard Shaw

One participant out of 11 replaced *life* with *battle* instead of the correct synonym, *being*.

PP25: "Always do what you are *afraid* to do." — Ralph Waldo Emerson

This quote had the most incorrect answers, with three out of 10 participants replacing *afraid* with either *idiotic* or *new* instead of the correct synonym, *scared*.

The occurrence of incorrect responses may hint at several factors. Misinterpretation of the instructions could be one such reason, though this seems less likely given the high accuracy rate of responses, with 279 out of 290 being correct.

Vocabulary limitations could be another reason for the errors. We attempted to control for this by recruiting only those participants who self-reported as fluent English speakers, thereby minimizing the chance that proficiency was a significant issue.

Attention-related factors, including time constraints or cognitive load, could also have contributed to the incorrect answers. However, based on our findings, this, too, seems unlikely. We found no significant differences in perceived workload, which measures cognitive workload, across the NP , AP-P , and AP-A variants. Additionally, the time spent on the survey and the last page reached showed little and insignificant difference between the AP-P and AP-A variants, despite the AP-P variant not posing questions to respondents.

Following this reasoning, we assume that the primary cause of the incorrect responses is the ambiguity of the alternative word options. Future studies may prevent this issue from occurring by pilot testing the quotes and alternative words on a smaller group of participants before implementing them in a full-scale survey, which we did not do. This would help ensure the correct answers are clearly correct and the alternative answers are clearly incorrect.

5

Discussion and Implications

This study aimed to decrease dropout rates in long surveys using engaging both intrinsically motivated (**IM**) and extrinsically motivated (**EM**) respondents through achievement primes. The effect of the achievement primes was observed by assigning respondents to different questionnaire variants, either containing no primes (**NP**), passive achievement primes (**AP-P**), or active achievement primes (**AP-A**).

During recruitment, it was observed that obtaining participants from the IM group was more challenging compared to the EM group. As IM and EM respondents were contacted through different platforms, the difference may be accounted for by the reach of the platforms. Where there were 100,000 EM respondents eligible to be recruited from Prolific, the IM respondents were recruited via public posts on social media platforms, with a combined view count of less than 1000. Additionally, direct inquiry via private social media channels was limited, resulting in fewer possible IM candidates. The desired sample size of 288 associated with a Power of 0.8 was therefore not reached, instead having obtained 228 responses with a resulting Power of 0.67. The discussion of significant results should therefore be generalised with caution.

Our research failed to replicate the findings of Gadiraju and Dietze (2017), which suggested that the exposure of EM respondents to AP-P and AP-A could reduce dropout. Amongst the 140 EM respondents, only three dropped out across the three survey variants in our study. The limited dropout amongst all EM respondents does not prove that repeated exposure to stimuli reinforces the achievement priming effect. We speculate that the fair wage provided to the EM respondents was the cause of the small number of dropouts. As a reward is a motivator for respondents to participate (Hall et al., 2019; Howell, 2020; Ziegenfuss et al., 2013), a large reward may increase the threshold at which point people drop out. The study also aimed to investigate the effect of achievement primes on IM respondents. Similar to the EM group, no significant effect on dropout was observed for IM respondents when controlling for the questionnaire variants. Consequently, we cannot confirm hypotheses *H1A*, *H1B*, and *H1C* and therefore cannot confirm *RQ1* conclusively.

Interestingly, a significant difference in dropout was noted between IM and EM , with 41% of IM respondents dropping out compared to 2% of EM respondents. Despite no significant difference within each respondent group across the questionnaire variants, the significant difference across the respondent groups suggests factors other than achievement primes could be at play. Nonetheless, we can confirm *H2* and answer *RQ2* by stating that EM respondents exposed to achievement primes exhibit lower dropout rates than their IM counterparts.

Regarding survey duration, both IM and EM respondents tended to spend more time on the survey when presented with AP-P or AP-A variants than the NP variant, albeit without statistical significance. This extended duration could be attributed to the additional time required to read and process the quotes or answer associated questions. Contrary to the findings of Gadiraju and Dietze (2017), our data indicated no significant difference between the AP-P and AP-A variants. Although respondents required more time to complete the survey when presented with achievement primes, an increase in engagement of EM respondents was observed, albeit insignificantly. The primes may have disrupted the accumulation of burden associated with the questionnaire itself (Howell, 2020), serving a similar purpose as enjoyable dummy questions used in other research (Elmalech et al., 2016). This

engagement trend was, however, not observed amongst IM respondents. The significant difference between the IM and EM might be explained by the heightened achievement motivation characteristic of EM respondents (Gadiraju & Dietze, 2017). Consistent with the recommendations of Manfreda and Vehovar (2002), our observations stress the importance of continuous engagement to reduce survey dropout rates. Even though a trend of increased engagement was observed, the changes were not statistically significant. Alternative methods with significant improvements might be more effective in preventing dropout than including achievement primes.

Regarding perceived workload, no significant difference was observed across the prime variants within each respondent group. However, between the groups, EM respondents reported a lower perceived workload than IM , which could be attributed to the promise of a reward upon completing the survey. This difference was, however, also insignificant.

Based on these findings, we recommend the following strategies for researchers intending to conduct long surveys. For those with a budget to offer financial incentives, it is advisable to target EM respondents instead of spending non-financial resources, such as time, to recruit IM respondents. This is because EM respondents, when paid a fair wage, demonstrate lower dropout rates and higher engagement levels, even without adding primes. Primes further increase the engagement of EM respondents but do so at the cost of a longer time to fill out the survey, which increases the cost of running a survey. Therefore, the primes should be included based on whether the increase in engagement justifies the additional cost. For those with budget constraints, it may be more practical to avoid including primes in long surveys.

Limitations and Future Research

This study had several limitations that should be considered when interpreting the results. These limitations may also provide directions for future research.

As mentioned before, the difference in dropout between the IM and EM respondents might have been caused by the high reward offered to the EM respondents upon survey completion. This incentive may have increased the burden threshold for these respondents, resulting in lower dropout rates. The difference observed might therefore reflect the impact that these wages have. Future research could adjust the wage used in this study to determine whether the observed differences were due to the combination of prime and respondent motivation or the reward provided to extrinsically motivated respondents. The wage, however, should always be fair for the target demographic (L. Han et al., 2019).

Moreover, the difference in duration may have also been caused by the differences in reward between the groups as opposed to the inclusion of the achievement primes. EM respondents, motivated by the fixed monetary reward, may have been incentivised to complete the survey quickly to maximise their hourly earnings by moving on to subsequent paid surveys, called satisficing (Kapelner & Chandler, 2010). Attention checks were put into place to mitigate this issue, which, when wrongfully answered, indicate which respondents were inattentive. IM respondents, on the other hand, may not have felt the same time pressure, as their income is not affected by the number of surveys they complete. As a result, they may have been more likely to pause in the middle of the survey and resume it later, resulting in an increased recorded duration. This difference could explain why there was no observable difference in time spent on the survey between respondent groups, despite the higher dropout rates among IM respondents. Future research may further investigate this relationship.

Furthermore, a possible personal connection between the IM respondents and the surveyor, caused by the partial recruitment via private social media channels, may have influenced their responses and behaviour during the survey. The respondents might have felt an inherent pressure to stay engaged with the survey for longer, potentially answering more questions than they would have otherwise. To mitigate this potential bias, we explicitly stated in the Informed Consent Form that respondents were free to quit the survey at any time. The details of this form can be found in Appendix B. Moreover, this perceived pressure could have led to more positive responses from IM respondents, who may have been inclined to please the surveyor. Similarly, EM respondents may have answered more positively than they usually would have, driven by self-interest bias (Draws et al., 2021). By answering more positively, they might have hoped to positively influence the surveyor's mood, increasing the likelihood of receiving a higher bonus for excellent performance.

Additionally, included primes feature the original author's name, which might cause an affect heuristic (Morris et al., 2013). Achievement primes are designed to evoke a respondent's drive for achievement. In our implementation, this effect may not occur if the respondent has a negative connotation with the name of the author associated with the quote. In contrast to the intended effect, the prime could reduce the respondent's motivation.

Also, the design of the achievement primes differs from that of regular questions, potentially leading to a salience bias (Draws et al., 2021). We made the design of the primes in the `AP-P` and `AP-A` variants equal to ensure a potential difference in behaviour resulted from the interactive nature rather than the design. However, to differentiate the primes in the `AP-P` variant from the preceding and succeeding questions, we added a design element in the form of a large quotation mark. Consequently, the observed effect in our study may not solely be caused by the inspirational value of the quote. Instead, it might be influenced by its visually distinctive presentation.

An unexpected observation from our study was the number of incorrectly answered `AP-A` questions. We attribute this issue to the ambiguity of the alternative words presented. To address this issue in future studies, ensuring that this ambiguity does not add to the respondent's cognitive load, we suggest pilot testing the alternative words before including them in a full-scale survey. This strategy could help identify and improve potential ambiguities, improving the reliability of the achievement priming theory.

The design of our study also results in a knowledge gap which might lead to promising results. The research by Gadiraju and Dietze (2017) applied achievement primes to short surveys targeting `EM` respondents, whereas we applied achievement primes to long surveys targeting both `IM` and `EM` respondents. Even though we observed no significant effects of achievement primes when applied to long surveys, we have not researched whether `IM` participants respond the same when the achievement primes are applied to short surveys.

Moreover, the frequency and placement of the achievement primes might have influenced the outcome. Our research ensured that one prime was placed pseudo-randomly on each of the four pages. Future studies might explore the impact of placing multiple primes on a single page. Alternatively, a dynamic system may be implemented to present primes to participants whose behaviour indicates they might drop out, similar to research by Kobren et al. (2015).

As explained in section 3.3, we used context- and domain-independent quotes as achievement primes. Future work may explore adapting the achievement primes to align more closely with the questionnaire context. This could potentially reduce any suspicion from the user regarding the unrelated subject matter of the primes, thereby potentially strengthening their impact.

The current implementation of primes allows them to be seen only once per page, making them easy to scroll past. Using a simple CSS command, `position: sticky`, we could ensure that the prime remains visible on the respondent's screen at all times, potentially reinforcing the prime's effect similar to research by Lewis et al. (2011) and Morris et al. (2012). However, this might only be applicable for the `AP-P` variants, as it can serve as decoration, while the current design of the `AP-A` variant requires it to be placed amongst the questions. As such, in addition to the interactivity of the prime, visibility becomes a new variable to be considered when analysing the results.

6

Conclusion

The use of online surveys to collect data from respondents has increased over the years due to the advantages they offer over other means of gathering information, including geographical independence, easy data management, and the ability to attract a large number of participants. However, online surveys suffer from dropout, defined as respondents who prematurely leave a survey, which causes a reduction in statistical power. Especially long surveys suffer from this issue as the dropout rate increases with the length of the survey.

In this study, we explored the potential of using achievement primes to combat this issue, focusing on both intrinsically (**IM**) and extrinsically motivated (**EM**) respondents.

Our study provided the user with one of three different survey variants, either without primes (**NP**), with passive achievement primes (**AP-P**), or with active achievement primes (**AP-A**). The effect of the achievement primes on the dropout of `EM` respondents when applied to *short* surveys as observed by Gadiraju and Dietze (2017), was not observed in our study when applied to *long* surveys. Furthermore, no significant difference in dropout amongst `IM` respondents was observed. As this implies that we cannot accept *H1A*, *H1B*, and *H1C*, we also cannot answer *RQ1: "How does achievement priming affect the dropout rate on long surveys?"* conclusively.

A difference in dropout between `IM` and `EM` respondents was observed, where `EM` respondents demonstrated statistically significant lower dropout rates than their `IM` counterparts when presented with either the `AP-P` or `AP-A` variant of the survey. However, this difference may have resulted from the fair wage provided as an incentive to the `EM` respondents rather than including achievement primes. Nevertheless, we can accept *H2*, and thereby can answer *RQ2: "How do achievement primes influence respondents with different motivations to participate?"* by stating that `AP-P` and `AP-A` result in lower dropout when presented to `EM` respondents than `IM` respondents. Moreover, both prime variants result in higher engagement in `EM` respondents compared to `IM` respondents.

However, several limitations of our study were identified, such as the imbalance in the number of `IM` and `EM` respondents, self-interest bias, or salience bias. These limitations may inspire future research to measure further the effects of strategies to reduce the dropout of surveys.

In conclusion, this study provided insights into the impact of achievement primes on the behaviour of respondents in long online surveys. We showed that the initial motivation of starting a survey impacts the dropout rate and user engagement. While achievement primes did not reduce dropout rates, their potential to influence respondent behaviour, such as increasing engagement despite the extended time spent on a survey, may be cause for further research. By means of exploring the knowledge gap presented in previous research, this study contributes to the continuous efforts to improve the effectiveness of online surveys.

Bibliography

10 tips to shorten your survey length. (n.d.). https://www.dynata.com/10-tips-to-shorten-your-survey-length/

5 ways your survey tool can help to shorten your questionnaire. (2016). https://www.confirmit.com/Resources/Voice-of-the-Customer/5-Ways-Your-Survey-Tool-Can-Help-to-Shorten-Your-Q/

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, *53-56*, 63–70. https://doi.org/10.1016/j.jom.2017.06.001

Abdi, H. (2010). Holm's sequential Bonferroni procedure. http://www.utd.edu/%E2%88%BCherve

Allen, P., & Roberts, L. (2016). The impact of academic sponsorship on Web survey dropout and item non-response. *First Monday, Volume 21, Number 2 - 1 February 2016*, *21*(2). https://doi.org/10.2501/IJMR-53-5-687-703

Antarika, S. (2021). Optimal survey length: How long survey lengths can affect data quality. https://medium.com/think-cult/optimal-survey-length-how-long-survey-lengths-can-affect-data-quality-f0d6398d25ee

Bargh, J., & Chartrand, T. (2000). The mind in the middle: A practical guide to priming and automaticity research. *Handbook of research methods in social and personality psychology*, 253–285.

Bargh, J. A., Lee-Chai, A., Barndollar, K., Gollwitzer, P. M., & Trötschel, R. (2001). *The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals* (tech. rep.).

Bell, S. W., Kong, J. C., Clark, D. A., Carne, P., Skinner, S., Pillinger, S., Burton, P., & Brown, W. (2022). The National Aeronautics and Space Administration-task load index: NASA-TLX: evaluation of its use in surgery. *ANZ Journal of Surgery*, *92*(11), 3022–3028. https://doi.org/10.1111/ans.17830

Bosch, O. J., Revilla, M., Qureshi, D. D., & Höhne, J. K. (2022). A new experiment on the use of images to answer web survey questions. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *185*(3), 955–980. https://doi.org/10.1111/rssa.12856

Bosnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., & Kaczmirek, L. (2008). Prenotification in web-based access panel surveys: The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review*, *26*(2), 213–223. https://doi.org/10.1177/0894439307305895

Brooke, J. (1995). *SUS: A quick and dirty usability scale Fault diagnosis training View project Usable systems View project* (tech. rep.). https://www.researchgate.net/publication/228593520

Byers, J. C., Bittner, A., & Hill, S. G. (1989). Traditional and raw task load index (tlx) correlations: Are paired comparisons necessary. *Advances in industrial ergonomics and safety*, *1*, 481–485.

Chinn, A. (2023). How long should a survey be? the ideal survey length. https://blog.hubspot.com/service/ideal-survey-length

Christensen, A. I., Ekholm, O., Kristensen, P. L., Larsen, F. B., Vinding, A. L., Glümer, C., & Juel, K. (2015). The effect of multiple reminders on response patterns in a Danish health survey. https://doi.org/10.1093/eurpub/cku057

Chudoba, B. (n.d.). How much time are respondents willing to spend on your survey? https://www.surveymonkey.com/curiosity/survey_completion_times/

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Conrad, F., Couper, M., Tourangeau, R., & Peytchev, A. (2005). *Impact of Progress Feedback on Task Completion: First Impressions Matter* (tech. rep.).

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2010). The impact of progress indicators on task completion. *Interacting with Computers*, *22*(5), 417–427. https://doi.org/10.1016/j.intcom.2010.03.001

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in Web- or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836. https://doi.org/10.1177/00131640021970934

Couper, M. P., & Mavletova, A. (2014). Mobile Web Survey Design: Scrolling versus Paging, SMS versus E-mail Invitations. *Journal of Survey Statistics and Methodology*, *2*(4), 498–518.

Crawford, S. D., & Couper Mark J Lamias, M. P. (2001). *Web Surveys Perceptions of Burden* (tech. rep.).

Dai, P., Rzeszotarski, J. M., Paritosh, P., & Chi, E. H. (2015). And now for something completely different: Improving crowdsourcing workflows with micro-diversions. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, 628–638. https://doi.org/10.1145/2675133.2675260

DeFranzo, S. (n.d.). Shorten your online survey for a better response rate. https://www.snapsurveys.com/blog/shorten-survey-response-rates/

Deshpande, H. (n.d.). Survey length: Keep it short and simple. https://www.questionpro.com/blog/survey-length-keep-it-short-and-simple/

Desmet, P. M. A., Vastenburg, M. H., & Romero, N. (2016). *Pick-A-Mood manual: Pictorial self-report scale for measuring mood states* (tech. rep.). Delft University of Technology. Delft. www.diopd.org

Desmet, P. M. A., Vastenburg, M. H., Romero, N., & Pieter, D. M. A. (2016). *Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale* (tech. rep. No. 3).

Deutskens, E., De Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). *Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study* (tech. rep. No. 1).

Draws, T. (2021). Statistical Test Finder. https://timdraws.net/files/StatisticalTestFinder.pdf

Draws, T., Rieger, A., Inel, O., Gadiraju, U., & Tintarev, N. (2021). *A Checklist to Combat Cognitive Biases in Crowdsourcing* (tech. rep.). www.aaai.org

Dykema, J., Stevenson, J., Day, B., Sellers, S. L., & Bonham, V. L. (2011). Effects of incentives and prenotification on response rates and costs in a national web survey of physicians. *Evaluation & the health professions*, *34*(4), 434–447.

Elmalech, A., Sarne, D., David, E., & Hajaj, C. (2016). *Extending Workers' Attention Span Through Dummy Events* (tech. rep.). www.aaai.org

Engeser, S. (2009). Nonconscious activation of achievement goals. Moderated by word class and the explicit achievement motive? *Swiss Journal of Psychology*, *68*(4), 193–200. https://doi.org/10.1024/1421-0185.68.4.193

Fang, J., Wen, C., & Pavur, R. (2012). Participation willingness in web surveys: Exploring effect of sponsoring corporation's and survey provider's reputation. https://doi.org/10.1089/cyber.2011.0411

Feyisetan, O., Simperl, E., Kleek, M. V., & Shadbolt, N. (2015). Improving Paid Microtasks through Gamification and Adaptive Furtherance Incentives. *Proceedings of the 24th International Conference on World Wide Web*, 333–343. http://fold.it

Frick, A., Bächtiger, M.-T., & Reips, U.-D. (2001). *Dimensions of Internet Science Financial Incentives, Personal Information, and Drop Out in Online Studies* (tech. rep.). Pabst Science Publishers. http://www.genpsy.unizh.ch/reips/dis/

Gadiraju, U., & Dietze, S. (2017). Improving Learning through Achievement Priming in Crowdsourced Information Finding Microtasks. *Proceedings of ACM LAK Conference*, *2022-March*. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics*, *22*(2), 313–328.

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

Goritz, A. S. (2014). *Determinants of the starting rate and the completion rate in online panel studies* (tech. rep.).

Guéguen, N., & Jacob, C. (2002). *Social Presence Reinforcement and Computer-Mediated Communication: The Effect of the Solicitor's Photography on Compliance to a Survey Request Made by E-Mail* (tech. rep. No. 2). www.liebertpub.com

Hall, E., Sanchez, T., Stephenson, R., Stein, A. D., Sineath, R. C., Zlotorzynska, M., & Sullivan, P. (2019). Randomised controlled trial of incentives to improve online survey completion among

internet-using men who have sex with men. *Journal of Epidemiology and Community Health*, *73*(2), 156–161. https://doi.org/10.1136/jech-2018-211166

Han, L., Roitero, K., Gadiraju, U., Sarasua, C., Checco, A., Maddalena, E., & Demartini, G. (2019). All Those Wasted Hours: On Task Abandonment in Crowdsourcing. *The TwelfthACM International Conference on Web Search and Data Mining (WSDM '19),February 11–15, 2019, Melbourne, VIC, Australia.*, *2022-March*. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Han, V., Albaum, G., Wiley, J. B., & Thirkell, P. (2009). Applying theory to structure respondents' stated motivations for participating in web surveys. *Qualitative Market Research*, *12*(4), 428–442. https://doi.org/10.1108/13522750910993338

Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Using Worker Avatars to Improve Microtask Crowdsourcing. *IEEE International Conference on Program Comprehension*, *2022-March*, 36–47. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Harisson, C., Yeo, Z., & Hudson, S. E. (2010). Faster Progress Bars: Manipulating Perceived Duration with Visual Augmentations. *CHI*.

Harrison, L., Skau, D., Franconeri, S., Lu, A., & Chang, R. (2013). *Influencing Visual Judgment through Affective Priming*.

Hart, S. (1986). NASA task load index (TLX).

Hart, W., & Albarracín, D. (2009). The Effects of Chronic Achievement Motivation and Achievement Primes on the Activation of Achievement and Fun Goals. *Journal of Personality and Social Psychology*, *97*(6), 1129–1141. https://doi.org/10.1037/a0017146

Heerwegh, D., & Loosveldt, G. (2006). *An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey Sponsor Logos in Web Surveys* (tech. rep.).

Henning, J. (2023). The ideal survey length. https://researchscape.com/blog/is-the-ideal-survey-length-20-minutes

Hermans, D., De Houwer, J., & Eelen, P. (2001). A time course analysis of the affective priming effect. *Cognition and Emotion*, *15*(2), 143–165. https://doi.org/10.1080/02699930125768

Ho, J., Tumkaya, T., Aryal, S., Choi, H., & Claridge-Chang, A. (2019). Moving beyond P values: data analysis with estimation graphics. https://doi.org/10.1038/s41592-019-0470-3

Hochheimer, C. J., Sabo, R. T., Krist, A. H., Day, T., Cyrus, J., & Woolf, S. H. (2016). Methods for evaluating respondent attrition in web-based surveys. *Journal of Medical Internet Research*, *18*(11). https://doi.org/10.2196/jmir.6342

Hoerger, M. (2010). Participant dropout as a function of survey length in internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behavior, and Social Networking*, *13*(6), 697–700. https://doi.org/10.1089/cyber.2009.0445

Howell, B. (2020). Why do participants drop out of online surveys and experiments? https://www.psychstudio.com/articles/dropout/

JASP Team. (2023). JASP (Version 0.17.2)[Computer software]. https://jasp-stats.org/

Jin, L. (2011). Improving response rates in web surveys with default setting: The effects of default on web survey participation and permission. *International Journal of Market Research*, *53*(1), 75–94. https://doi.org/10.2501/IJMR-53-1-075-094

Joinson, A. N., & Reips, U. D. (2007). Personalized salutation, power of sender and response rates to Web-based surveys. *Computers in Human Behavior*, *23*(3), 1372–1383. https://doi.org/10.1016/j.chb.2004.12.011

Kaczmirek, L. (2008). *Human-Survey Interaction Usability and Nonresponse in Online Surveys* (tech. rep.). http://www.kaczmirek.de/book2008

Kalleitner, F., Mühlböck, M., & Kittel, B. (2022). What's the benefit of a video? the effect of nonmaterial incentives on response rate and bias in web surveys. *Social Science Computer Review*, *40*(3), 700–716.

Kapelner, A., & Chandler, D. (2010). *Preventing Satisficing in Online Surveys: A "Kapcha" to Ensure Higher Quality Data* * (tech. rep.). http://danachandler.com/kapchastudy.html

Keep your survey length to 20 minutes or less. (n.d.). https://www.themarketechgroup.com/en/media-center/Press-releases/survey-length

Keusch, F. (2012). How to Increase Response Rates in List-Based Web Survey Samples. *Social Science Computer Review*, *30*(3), 380–388. https://doi.org/10.1177/0894439311409709

Kim, S., Lee, J., & Gweon, G. (2019). Comparing data from chatbot and web surveys effects of platform and conversational style on survey response quality. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300316

Kobren, A., Tan, C. H., Ipeirotis, P., & Gabrilovich, E. (2015). Getting more for less: Optimized Crowdsourcing with dynamic tasks and goals. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, 592–602. https://doi.org/10.1145/2736277.2741681

Lavidas, K., Petropoulou, A., Papadakis, S., Apostolou, Z., Komis, V., Jimoyiannis, A., & Gialamas, V. (2022). Factors Affecting Response Rates of the Web Survey with Teachers. *Computers*, *11*(9). https://doi.org/10.3390/computers11090127

Law, E., Yin, M., Goh, J., Chen, K., Terry, M., & Gajos, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. *Conference on Human Factors in Computing Systems - Proceedings*, 4098–4110. https://doi.org/10.1145/2858036.2858144

Lewis, S., Dontcheva, M., & Gerber, E. (2011). Affective Computational Priming and Creativity. *CHI*.

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

Liu, M., & Wronski, L. (2018b). Examining Completion Rates in Web Surveys via Over 25,000 Real-World Surveys. *Social Science Computer Review*, *36*(1), 116–124. https://doi.org/10.1177/0894439317695581

Maddalena, E., Basaldella, M., De Nart, D., Degl'innocenti, D., Mizzaro, S., & Demartini, G. (2016). *Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge* (tech. rep.). www.aaai.org

Maliković, M., & Tončić, M. (2021). *Using Individually Adapted Progress Indicator in Web surveys* (tech. rep.).

Manfreda, K. L., Batagelj, Z., & Vehovar, V. (2002). Design of web survey questionnaires: Three basic experiments. https://doi.org/10.1111/j.1083-6101.2002.tb00149.x

Manfreda, K. L., & Vehovar, V. (2002). Survey Design Features Influencing Response Rates in Web Surveys. In *The international conference on improving surveys proceedings* (pp. 25–28).

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys: A field experiment on nonresponse in web surveys. *Social Science Computer Review*, *25*(3), 372–383. https://doi.org/10.1177/0894439307297606

Marlow, J., & Dabbish, L. (2015). The effects of visualizing activity history on attitudes and behaviors in a peer production context. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, 757–764. https://doi.org/10.1145/2675133.2675250

Matzat, U., Snijders, C., & Van Horst, W. D. (2009). Effects of different types of progress indicators on drop-out rates in Web surveys. *Social Psychology*, *40*(1), 43–52. https://doi.org/10.1027/1864-9335.40.1.43

Menon, V., & Muraleedharan, A. (2020). Internet-based surveys: Relevance, methodological considerations and troubleshooting strategies. https://doi.org/10.1136/gpsych-2020-100264

Morris, R. R., Dontcheva, M., Finkelstein, A., & Gerber, E. (2013). Affect and creative performance on crowdsourcing platforms. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 67–72. https://doi.org/10.1109/ACII.2013.18

Morris, R. R., Dontcheva, M., & Gerber, E. (2012). *Priming for Better Performance in Microtask Crowdsourcing Environments* (tech. rep.). www.computer.org/internet/

Napierala, M. A. (1995). *What Is the Bonferroni Correction?* (Tech. rep.). http://www.aaos.org/news/aaosnow/apr12/research7.asp

Nasa task load index (nasa tlx). (n.d.). https://www.skybrary.aero/articles/nasa-task-load-index-nasa-tlx

Nestler, S., Thielsch, M., Vasilev, E., & Back, M. D. (2015). Will They Stay or Will They Go? Personality Predictors of Dropout in an Online Study. *International Journal of Internet Science*, *10*(1), 37–48.

Nielsen, J. (2004). Keep online surveys short. https://www.nngroup.com/articles/keep-online-surveys-short/

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human factors*, *33*(1), 17–33.

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human Computer Studies*, *112*, 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

Oikawa, M. (2004). *Moderation of Automatic Achievement Goals by Conscious Monitoring* (tech. rep.).

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

O'neil, K. M., Jay, J., Penrod, S. D., & Bornstein, B. H. (2003). *Web-based research: Methodological variables' effects on Web-based research: Methodological variables' effects on dropout and sample characteristics* (tech. rep.). https://digitalcommons.unl.edu/psychfacpub/324

Pan, B., Woodside, A. G., & Meng, F. (2014). How Contextual Cues Impact Response and Conversion Rates of Online Surveys. *Journal of Travel Research*, *53*(1), 58–68. https://doi.org/10.1177/0047287513484195

Pedersen, M. J., & Nielsen, C. V. (2016). Improving Survey Response Rates in Online Panels: Effects of Low-Cost Incentives and Cost-Free Text Appeal Interventions. *Social Science Computer Review*, *34*(2), 229–243. https://doi.org/10.1177/0894439314563916

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74–97. https://doi.org/10.1093/poq/nfp014

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web survey design: Paging versus scrolling. *Public Opinion Quarterly*, *70*(4), 596–607. https://doi.org/10.1093/poq/nfl028

Prinz, T. M., Graefe, L., & Plötner, J. (2020). Learning from the Past - Do historical data help to improve progress indicators in web surveys? *UIST 2020 - Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 379–390. https://doi.org/10.1145/3379337.3415838

Qiu, S., Gadiraju, U., & Bozzon, A. (2020). Improving Worker Engagement Through Conversational Microtask Crowdsourcing. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376403

Quinn, P. (2010). Excessive survey length – eight tips to help you avoid. https://peoplepulse.com/resources/useful-articles/excessive-survey-length/

Ramírez, J., Baez, M., Casati, F., & Benatallah, B. (2019). *Understanding the Impact of Text Highlighting in Crowdsourcing Tasks* (tech. rep.). https://rajpurkar.github.io/SQuAD-explorer/

Revilla, M., & Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), 538–545. https://doi.org/10.1177/1470785320943049

Revilla, M., & Ochoa, C. (2017). Ideal and Maximum Length for a Web Survey. *International Journal of Market Research*, *59*(5), 557–565. https://doi.org/10.2501/IJMR-2017-039

Roberts, C., Herzing, J. M., Manjon, M. A., Abbet, P., & Gatica-Perez, D. (2022). Response Burden and Dropout in a Probability-Based Online Panel Study - A Comparison between an App and Browser-Based Design. *Journal of Official Statistics*, *38*(4), 987–1017. https://doi.org/10.2478/jos-2022-0043

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003

Rzeszotarski, J. M., Chi, E., Paritosh, P., & Dai, P. (2013). *Inserting Micro-Breaks into Crowdsourcing Workflows* (tech. rep.). www.aaai.org

Saleh, A., & Bista, K. (2017). *Examining Factors Impacting Online Survey Response Rates in Educational Research: Perceptions of Graduate Students* (tech. rep.). http://www.jmde.com

Sanchez, M. E. (1992). *Effects of Questionnaire Design on the Quality of Survey Data* (tech. rep. No. 2).

Sánchez-Fernández, J., Muñoz-Leiva, F., & Montoro-Ríos, F. J. (2012). Improving retention rate and response quality in Web-based surveys. *Computers in Human Behavior*, *28*(2), 507–514. https://doi.org/10.1016/j.chb.2011.10.023

Sauermann, H., & Roach, M. (2013). Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. *Research Policy*, *42*(1), 273–286. https://doi.org/10.1016/j.respol.2012.05.003

Shropshire, K. O., Hawdon, J. E., & Witte, J. C. (2009). Web survey design: Balancing measurement, response, and topical interest. *Sociological Methods and Research*, *37*(3), 344–370. https://doi.org/10.1177/0049124108327130

Six tips to shorten your survey. (2013). https://www.verint.com/blog/six-tips-to-shorten-your-survey/

Standard Definitions. (2016). Final Dispositions of Case Codes and Outcome Rates for Surveys.

Stanton, N., Salmon, P., Walker, G., Baber, C., & Jenkins, D. (2005). Mental Workload Assessment Method. In *Human factors methods: A practical guide for engineering and design* (pp. 301–64).

Survey length best practices: Are shorter surveys better? (n.d.). https://www.dynata.com/survey-length-best-practices-are-shorter-surveys-better/

Terentev, E., & Maloshonok, N. (2019). The impact of response options ordering on respondents' answers to rating questions: results of two experiments. *International Journal of Social Research Methodology*, *22*(2), 179–198. https://doi.org/10.1080/13645579.2018.1510660

Tijdens, K. (2014). Dropout rates and response times of an occupation search tree in a web survey. *Journal of Official Statistics*, *30*(1), 23–43. https://doi.org/10.2478/jos-2014-0002

Trouteaud, A. R. (2004). How you ask counts: A test of Internet-related components of response rates to a web-based survey. https://doi.org/10.1177/0894439304265650

Vanette, D. (n.d.). Survey Methodology and Compliance Best Practices. https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/survey-methodology-compliance-best-practices/

Veit, D., Kaufmann Germany, N., Schulze, T., & Schwarz Chair, D. (2011). *More than fun and money. Worker Motivation in Crowdsourcing - A Study on Mechanical Turk* (tech. rep.). https://www.researchgate.net/publication/216184483

Vicente, P., & Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review*, *28*(2), 251–267. https://doi.org/10.1177/0894439309340751

Villar, A., Callegaro, M., & Yang, Y. (2013). Where Am I? A Meta-Analysis of Experiments on the Effects of Progress Indicators for Web Surveys. *Social Science Computer Review*, *31*(6), 744–762. https://doi.org/10.1177/0894439313497468

Vriesema, C. C., & Gehlbach, H. (2021). Assessing Survey Satisficing: The Impact of Unmotivated Questionnaire Responding on Data Quality. *Educational Researcher*, *50*(9), 618–627. https://doi.org/10.3102/0013189X211040054

Walston, J. T., Lissitz, R. W., & Rudner, L. M. (n.d.). The Influence of Web-based Questionnaire Presentation Variations on Survey Cooperation and Perceptions of Survey Quality. *2006*.

Wen, C., & Fang, J. (2012). *Predicting potential respondents' decision to participate in web surveys* (tech. rep.). http://thekeep.eiu.edu/business_fac/10

Wigmore, S. (2022). What is a good survey length for online research? https://www.kantar.com/north-america/inspiration/research-services/what-is-a-good-survey-length-for-online-research-pf#:~:text=It%27s%20best%20to%20keep%20surveys,that%20is%20under%20five%20minutes.

Wu, M. J., Zhao, K., & Fils-Aime, F. (2022). Response rates of online surveys in published research: A meta-analysis. *Computers in Human Behavior Reports*, *7*. https://doi.org/10.1016/j.chbr.2022.100206

Xu, L., Zhou, X., & Gadiraju, U. (2019). Revealing the role of user moods in struggling search tasks. *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1249–1252. https://doi.org/10.1145/3331184.3331353

Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011a). Should i stay or should i go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research*, *23*(2), 131–147. https://doi.org/10.1093/ijpor/edq046

Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011b). Should i stay or should i go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research*, *23*(2), 131–147. https://doi.org/10.1093/ijpor/edq046

Zhang, C., Arbor, A., & Conrad, F. G. (2014). *Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining* (tech. rep. No. 2). http://www.surveymethods.org

Ziegenfuss, J. Y., Niederhauser, B. D., Kallmes, D., & Beebe, T. J. (2013). An assessment of incentive versus survey length trade-offs in a web survey of radiologists. https://doi.org/10.2196/jmir.2322

Zuell, C., Menold, N., & Körber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review*, *33*(1), 115–122.

A

Requirements for Questionnaires

Table A.1: Requirements used as a basis for describing best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Hoerger (2010)	When expecting X results, the survey should be held with X/(0,1+0,02*Y) people, with Y being the number of multiples of 100 items	-	-	Yes	-	-	N/A	243-535	N/A	1963	Yes
Law et al. (2016)	The survey should trigger curiosity	-	-	Yes	-	-	15	30	N/A	98	Yes
Galesic (2006)	The survey should be subjectively interesting	?	?	Yes	-	?	14, 26, or 38	180	N/A	2339	Yes
Galesic (2006)	The survey should cause little subjective burden	?	?	Yes	-	?	14, 26, or 38	180	N/A	2339	Yes
Galesic (2006)	The survey should not be long	?	?	Yes	-	?	14, 26, or 38	180	N/A	2339	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Galesic (2006)	Enjoyable questions should be alternated by burdensome questions	?	?	Yes	-	?	14, 26, or 38	180	N/A	2339	Yes
Galesic (2006)	The survey should contain as few open questions as possible	?	?	Yes	-	?	14, 26, or 38	180	N/A	2339	Yes
Hall et al. (2019)	The survey should offer incentives	?	Yes	Yes	-	-	15	N/A	N/A	1178	Yes
Hall et al. (2019)	The survey should include altruistic messaging	?	Yes	Yes	-	-	15	N/A	N/A	1178	Yes
Howell (2020)	The survey should offer feedback on individual performance	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Howell (2020)	The survey should offer incentives	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Howell (2020)	A survey with monetary incentive should not allow for multiple submissions	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Howell (2020)	The survey should be fun and engaging	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Howell (2020)	Demographic information should be acquired at the beginning rather than at the end	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Kaczmirek (2008)	The importance of a survey should be stated in the invitation or introduction of a survey	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Kaczmirek (2008)	The survey should have a personalised invitation	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Kaczmirek (2008)	The survey should be well designed and consistent	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Kaczmirek (2008)	The survey should include notes and instructions where confusion may occur	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Kaczmirek (2008)	The survey should only sparsely use matrix questions	-	-	-	-	-	N/A	47	13	1581	N/A
Kaczmirek (2008)	Progress should either be calculated dynamically, solving the problem imposed by filter questions, or not be shown at all	Yes	-	-	-	-	N/A	N/A	20	749	Yes
Matzat et al. (2009)	Progress indicators should not be placed on surveys which take 20 minutes or more to complete, regardless of the (artificial) speed of the progress indicator	?	?	-	-	-	22	N/A	N/A	2460	Yes
Matzat et al. (2009)	When shown only during a select time, progress indicators should only be shown early on in the survey	?	?	-	-	-	22	N/A	N/A	2460	Yes
Liu and Wronski (2018b)	The survey should not be long	?	?	?	?	?	N/A	N/A	N/A	N/A	Yes
Liu and Wronski (2018b)	The survey should contain as few open questions as possible	?	?	?	?	?	N/A	N/A	N/A	N/A	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Liu and Wronski (2018b)	A survey should start with multiple-choice questions over open-ended questions, and having the first item be non-question (such as a picture or text box) should be avoided	?	?	?	?	?	N/A	N/A	N/A	N/A	Yes
Liu and Wronski (2018b)	Preferably, there should be no progress indicator, after which a progress indicator at the top is preferred, and a progress indicator at the bottom should be avoided	?	?	?	?	?	N/A	N/A	N/A	N/A	Yes
Saleh and Bista (2017)	Invitations should be sent out via mail or phone surveys rather than via email	Yes	Yes	-	Yes	?	5 to 7	28	N/A	454	Yes
Villar et al. (2013)	When always shown, progress indicators should move fast-to-slow, rather than slow-to-fast, in order to reduce drop-off rates	?	?	?	?	?	8 to 30	N/A	N/A	N/A	Yes
Yan et al. (2011b)	The progress indicator should be shown when the expected and real duration of the survey are both short	-	-	Yes	-	-	16 or 25	101 or 155	10 or 12	2385	Yes
F. Conrad et al. (2005)	The progress indicator should move fast-to-slow	?	-	Yes	-	-	13 or 15	N/A	67	3179	Yes
F. Conrad et al. (2005)	The progress indicator should be placed intermittently	?	-	Yes	-	-	N/A	N/A	80	N/A	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Prinz et al. (2020)	Historical data should be used to show accurate progress	-	-	-	-	-	N/A	N/A	10,15, 20, 30, 40, 50 100, 200 or 500	1000	N/A
F. G. Conrad et al. (2010)	The progress indicator should move fast-to-slow	?	-	Yes	-	-	N/A	57	57	3179	Yes
F. G. Conrad et al. (2010)	The progress indicator should be placed intermittently	?	-	Yes	-	-	N/A	80	80	N/A	Yes
Harisson et al. (2010)	The progress indicator should show a ribbon effect which moves backwards to appear to be accelerating faster	-	-	-	-	-	N/A	N/A	N/A	20	N/A
Maliković and Tončić (2021)	The progress indicator should move fast-to-slow	?	?	?	?	?	18	49	N/A	328	Yes
Maliković and Tončić (2021)	The length of the progress indicator should be calculated by dividing the logarithm of the number of the current question by the logarithm of the total number of questions.	?	?	?	?	?	18	49	N/A	328	Yes
Heerwegh and Loosveldt (2006)	The user should be saluted personally	?	?	?	?	?	27	231 to 284	13	2520	Yes
Hochheimer et al. (2016)	Attrition rates should be analysed using visualisation of dropouts per question	-	?	-	?	-	N/A	17	N/A	2355	N/A
Hochheimer et al. (2016)	The visualisation of attrition patterns should be checked if they’re statistically significant	-	?	-	?	-	N/A	17	N/A	2355	N/A

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Hochheimer et al. (2016)	Statistically significant visualised attrition patterns should be associated with factors such as user characteristics	-	?	-	?	-	N/A	17	N/A	2355	N/A
Crawford and Couper Mark J Lamias (2001)	The stated length of the survey should represent the actual length	?	?	-	?	?	20	46	N/A	4500	Yes
Crawford and Couper Mark J Lamias (2001)	A follow-up reminder should be sent two days (instead of 5 days) after the initial e-mail	?	?	-	?	?	20	46	N/A	4500	Yes
Crawford and Couper Mark J Lamias (2001)	To increase the likelihood of signing in, the ID and password should be automatically entered. To increase the honesty of reporting, the respondent should manually enter the ID and password.	?	?	-	?	?	20	46	N/A	4500	Yes
Crawford and Couper Mark J Lamias (2001)	The progress indicator should not be shown when there are many open-ended questions	?	?	-	?	?	20	46	N/A	4500	Yes
Ziegenfuss et al. (2013)	An incentive should be provided to have people fill in long surveys	-	-	Yes	-	-	5-7 or 10	N/A	N/A	254	Yes
Lavidas et al. (2022)	Sponsorship should have a positive reputation to the respondents	Yes	Yes	-	-	-	N/A	30	N/A	263	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Lavidas et al. (2022)	Altruistic invite, interest in topic and content, reference of explicit purpose and promise of informing about the results should be included	Yes	Yes	-	-	-	N/A	30	N/A	263	Yes
Lavidas et al. (2022)	The questions should be simple and closed-ended, time required to complete, and the number of questions should be disclosed. The length should preferably not exceed 10 minutes	Yes	Yes	-	-	-	N/A	30	N/A	263	Yes
Lavidas et al. (2022)	There should be approval of the research from an official ethics commitment service. The way of access to the respondent's account by the researcher should be disclosed in case the invitation is sent via e-mail.	Yes	Yes	-	-	-	N/A	30	N/A	263	Yes
Lavidas et al. (2022)	The e-mail address should not be asked for within the survey	Yes	Yes	-	-	-	N/A	30	N/A	263	Yes
Pedersen and Nielsen (2016)	The the invitation should attain to the egotistic need for approval	-	Yes	Yes	-	Yes	N/A	N/A	N/A	6162	Yes
Pedersen and Nielsen (2016)	The survey should offer low-cost cash prize lottery incentives rather than donations	-	Yes	Yes	-	Yes	N/A	N/A	N/A	6162	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Marlow and Dabbish (2015)	The survey should report positive news on previous participants' performance	-	-	Yes	-	-	N/A	N/A	N/A	284	Yes
Sauermann and Roach (2013)	The survey should salute by using the first name only	-	-	-	-	-	N/A	N/A	N/A	24000	Yes
Sauermann and Roach (2013)	The survey should offer lottery incentives rather than pre-paid incentives	-	-	-	-	-	N/A	N/A	N/A	24000	Yes
Sauermann and Roach (2013)	Use as many reminders as possible, but include an opt-out link	-	-	-	-	-	N/A	N/A	N/A	24000	Yes
Sauermann and Roach (2013)	Each reminder should use different wording	-	-	-	-	-	N/A	N/A	N/A	24000	Yes
Wen and Fang (2012)	Sponsorship should have a positive reputation to the respondents	Yes	Yes	-	-	?	N/A	N/A	3	248	Yes
Wen and Fang (2012)	The survey topic should be interesting to the respondent	Yes	Yes	-	-	?	N/A	N/A	3	248	Yes
Wen and Fang (2012)	The survey invitation should create a feeling of moral obligation	Yes	Yes	-	-	?	N/A	N/A	3	248	Yes
Wen and Fang (2012)	The survey should ensure anonymity	Yes	Yes	-	-	?	N/A	N/A	3	248	Yes
Jin (2011)	The survey's default length option should be the preferred one of the researcher	-	-	-	-	Yes	N/A	N/A	N/A	1800	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Jin (2011)	The survey's default option to participate in future studies should be the preferred one of the researcher	-	-	-	-	Yes	N/A	N/A	N/A	2400	Yes
Keusch (2012)	A male recipient should be contacted by a female survey conductor	?	?	-	-	-	13,5	70	32	1563	-
Keusch (2012)	The respondent should be notified beforehand that they will be asked to participate in a survey	?	?	-	-	-	13,5	70	32	1563	-
Keusch (2012)	The survey should be well defined and consistent	?	?	-	-	-	13,5	70	32	1563	-
V. Han et al. (2009)	The survey should limit the time and effort required	?	?	-	-	?	N/A	25 to 28 or 36 to 39	N/A	861	-
Bosnjak et al. (2008)	Pre-notification should be sent out via SMS instead of via e-mail	-	-	Yes	-	Yes	N/A	N/A	2	562	Yes
Bosnjak et al. (2008)	Invitations should be sent out via e-mail	-	-	Yes	-	Yes	N/A	N/A	2	562	Yes
Joinson and Reips (2007)	The invitation should salute the respondent by first name (as opposed to full name)	?	?	-	-	?	N/A	N/A	N/A	1405	Yes
Joinson and Reips (2007)	The invitation should come from someone higher in power than the recipient	?	?	-	-	?	N/A	N/A	14	1054	Yes
Manfreda and Vehovar (2002)	The survey should not be long	-	-	-	-	-	N/A	N/A	N/A	N/A	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Manfreda and Vehovar (2002)	The survey should remain interesting or prevent annoyance for the entire duration	-	-	-	-	-	N/A	N/A	N/A	N/A	Yes
Manfreda and Vehovar (2002)	The survey should not use many open-ended questions	-	-	-	-	-	N/A	N/A	N/A	N/A	Yes
Manfreda and Vehovar (2002)	The survey should provide monetary incentive after completion	-	-	-	-	-	N/A	N/A	N/A	N/A	Yes
Pan et al. (2014)	The invitation should contain an authoritative, trustworthy, and consistent solicitation message	-	-	Yes	-	-	N/A	N/A	N/A	1257	Yes
Pan et al. (2014)	The research team should use a researcher with a name in a similar cultural background to the respondents to be the main contact person.	-	-	Yes	-	-	N/A	N/A	N/A	1257	Yes
Pan et al. (2014)	Researchers should use and disclose a sponsor's information that the respondents are familiar with	-	-	Yes	-	-	N/A	N/A	N/A	1257	Yes
Guéguen and Jacob (2002)	The survey's invitation should include a picture of the reviewer when the subject and reviewer do not know one another to have the respondent feel a personal connection	?	?	-	-	?	15-20	N/A	N/A	160	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Guéguen and Jacob (2002)	Male recipients should be invited by a female reviewer	?	?	-	-	?	15-20	N/A	N/A	160	Yes
Sánchez-Fernánd et al. (2012)	The invitation email should be personalized	?	?	Yes	-	-	N/A	N/A	N/A	1182	Yes
Sánchez-Fernánd et al. (2012)	A higher frequency of remainders may be best when time is limited, whereas a lower one may be suitable without time constraints to obtain a better retention rate	?	?	Yes	-	-	N/A	N/A	N/A	1182	Yes
Manfreda et al. (2002)	Logotypes should not be used when asking questions	?	?	?	?	?	N/A	16	N/A	750	Yes
Couper and Mavletova (2014)	For mobile surveys, the invitation should be sent via SMS instead of via e-mail	-	-	Yes	-	-	7	17	17	2100	Yes
Cook et al. (2000)	The survey's invitation should be personalized	-	-	-	-	-	N/A	N/A	N/A	N/A	Yes
Cook et al. (2000)	The survey should send out pre-notifications	-	-	-	-	-	N/A	N/A	N/A	N/A	Yes
Christensen et al. (2015)	One reminder should be sent out	?	?	-	-	?	N/A	56	N/A	177639	Yes
Goritz (2014)	The survey should be offered in a panel in which fewer surveys have recently been conducted	?	?	Yes	-	?	N/A	N/A	N/A	N/A	Yes
Goritz (2014)	The survey should invite panellists who participated in previous studies	?	?	Yes	-	?	N/A	N/A	N/A	N/A	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Goritz (2014)	The survey should be conducted after a longer time lag to the previous study, and the survey should be kept short	?	?	Yes	-	?	N/A	N/A	N/A	N/A	Yes
Frick et al. (2001)	The survey should mention the inclusion of a lottery at the beginning of the survey	-	-	Yes	-	-	N/A	N/A	N/A	804	Yes
Dai et al. (2015)	The survey should include micro-diversions in the form of a game or short story	-	-	Yes	-	-	60 to 120	N/A	N/A	270	Yes
Elmalech et al. (2016)	The survey should include dummy events on the fly when a model predicts the attention span is reduced	-	-	Yes	-	-	40	N/A	N/A	570	Yes
Gadiraju and Dietze (2017)	The survey should include quotes from famous people as achievement primes	-	-	Yes	-	-	N/A	N/A	N/A	240	Yes
Haque et al. (2022)	When using a conversational interface, worker avatars should be included (including similarity and wishful avatar identification)	-	-	Yes	-	-	N/A	50	N/A	360	Yes
L. Han et al. (2019)	AMT should be preferred over F8	-	-	Yes	-	-	N/A	24	N/A	200	Yes
L. Han et al. (2019)	The hourly wage should be sufficient	-	-	Yes	-	-	N/A	24	N/A	200	Yes
L. Han et al. (2019)	The effort required to complete the task should be subjectively low	-	-	Yes	-	-	N/A	24	N/A	200	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Rzeszotarski et al. (2013)	The survey should introduce micro-breaks in the form of a comic or a game in which they can gamble for additional earnings	-	-	Yes	-	-	60+	N/A	N/A	270	Yes
Feyisetan et al. (2015)	The survey should include gamification elements in the form of a leaderboard, levels, badges, feedback alerts, bonus points, treasure points, activities widgets	Yes	-	Yes	-	-	N/A	N/A	N/A	600, 423	Yes
Shropshire et al. (2009)	Interest-related questions should be placed early in the survey	-	-	-	-	-	N/A	N/A	N/A	1679	Yes
Tijdens (2014)	The list of options in a question should not be too long	-	-	-	Yes	-	10	N/A	22	18824	Yes
Qiu et al. (2020)	The tasks should be provided to the respondent in the style of a conversation	-	-	Yes	-	-	N/A	50	N/A	800	Yes
Ramírez et al. (2019)	Survey's text should be highlighted (as long as the selection of the highlighted text is of good quality)	-	-	Yes	-	-	N/A	18 or 24 or 26	6 or 12	1337, 255, 1035	Yes
Lewis et al. (2011)	The survey should contain elements which cause either positive or negative affect (e.g. in the form of an image of a smiling baby)	-	-	Yes	-	-	6	10 to 20	10	240	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Lewis et al. (2011)	The survey should contain elements which cause either positive or neutral affect (e.g. in the form of an image of a smiling baby)	?	?	Yes	-	?	5	N/A	N/A	27	Yes
Kobren et al. (2015)	Tasks should be allocated based on their historical performance, survival-based	Yes	?	-	-	-	N/A	50	N/A	13123	Yes
Kobren et al. (2015)	When the model predicts that the user will drop out, the user should be presented with an explicit limit on the quiz length (which is not shown beforehand)	Yes	?	-	-	-	N/A	50	N/A	13123	Yes
Maddalena et al. (2016)	A relevance judgment worker should not be able to prematurely skip pages	-	-	Yes	-	-	N/A	N/A	N/A	3200	N/A
Maddalena et al. (2016)	A relevance judgement task should be limited to 25-30 seconds	-	-	Yes	-	-	N/A	N/A	N/A	3200	N/A
Kim et al. (2019)	The questions should be posed by a chatbot with informal chat style	-	-	-	-	-	17 or 26	10 to 24	N/A	117	N/A
Roberts et al. (2022)	On mobile, the survey should be administered via an app rather than via the browser	-	Yes	-	-	-	20	N/A	N/A	2175	N/A
Vicente and Reis (2010)	A long survey should intermittently display true progress information	-	-	-	-	-	N/A	N/A	N/A	N/A	N/A
Revilla and Ochoa (2017)	The survey should ideally be ten min. and max 20 min.	-	-	Yes	-	-	N/A	115	N/A	755	Yes

Continues on next page …

Table A.1: Requirements used to describe best practices.

Researcher	Requirement	Participants					Duration in min	#questions	#pages	#respondent	Significant
		EB	CB	IP	DP	SM					
Bosch et al. (2022)	Respondents should not be asked to answer with images instead of typing text	-	-	-	-	-	N/A	86	N/A	3043	Yes
Peytchev (2009)	Long, attitudinal, numeric, and open-ended questions should be avoided	Yes	Yes	-	-	-	18 or 21	N/A	N/A	2831, 3195	Yes
Peytchev (2009)	The progress indicator should not under-report feedback in the beginning	Yes	Yes	-	-	-	18 or 21	N/A	N/A	2831, 3195	Yes
Peytchev (2009)	The survey should start with simple questions which exert low demand	Yes	Yes	-	-	-	18 or 21	N/A	N/A	2831, 3195	Yes

B

Informed Consent

B.1. Achievement Prime Curation

You are being invited to participate in a research study titled ”Preventing dropout on long web surveys” (under review). This study is being done by T. van Tussenbroek (graduate student), U. Gadiraju (thesis advisor), and G. Allen (daily co-supervisor) from the TU Delft. T. van Tussenbroek performs this research as a thesis intern at KPMG.

The purpose of this research study is to measure the effect of engagement methods in web surveys, and will take you approximately 15 minutes to complete. We will be asking you to rate questions on how inspiring they are. Your participation in this study is entirely voluntary and you can withdraw at any time. We automatically record your Prolofic ID to make sure that you will receive compensation and will delete this information once the study has finished.

The data will be used for writing a Master’s Thesis as part of the graduation procedure of T. van Tussenbroek. We do not collect any data aside from the information described above. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by ensuring your answers are stored completely anonymously. As data will be stored anonymously, we cannot remove an individual’s data. Be advised that with any online activity the risk of a breach is always possible, even given the precautions described.

For further questions, you may contact T. van Tussenbroek by e-mail via T.A.R.vanTussenbroek @student.tudelft.nl.

By clicking through to the (anonymous) online survey, you agree to this Opening Statement.

B.2. Main Study

You are being invited to participate in a research study. This study is being done by T. van Tussenbroek (graduate student), U. Gadiraju (thesis advisor), and G. Allen (daily co-supervisor) from the TU Delft. T. van Tussenbroek performs this research as a thesis intern at KPMG.

The study will take you approximately 20 minutes to complete. We will be asking you to answer questions about your health while working behind a computer. Your participation in this study is entirely voluntary and you can withdraw at any time.

The data will be used for writing a Master’s Thesis as part of the graduation procedure of T. van Tussenbroek. We do not collect any data aside from the information described above. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by ensuring your answers are stored completely anonymously. As data will be stored anonymously, we cannot remove an individual’s data. Be advised that with any online activity the risk of a breach is always possible, even given the precautions described.

For further questions, you may contact T. van Tussenbroek by e-mail via T.A.R.vanTussenbroek @student.tudelft.nl.

By clicking through to the (anonymous) online survey, you agree to this Opening Statement.

C

Data Achievement Prime Curation

Table C.1: Top 25 rated inspirational quotes used for achievement primes and the alternative words used for the active achievement prime questions.

ID	Quote	Author	Score	Word to be replaced	Replacement	Alternative words		
1	Turn your wounds into wisdom.	Oprah Winfrey	4.71	wisdom	knowledge	dad	contract	cold
2	Don't be pushed around by the fears in your mind. Be led by the dreams in your heart.	Roy T. Bennett	4.43	heart	soul	inspector	uncle	employment
3	"But hoping," he said, "is how the impossible can be possible after all."	Marissa Meyer	4.43	possible	conceivable	two	stormy	female
4	Dreams are only dreams until you wake up and make them real.	Ned Vizzini	4.43	dreams	fantasies	stores	directions	breakfasts
5	I never said it would be easy, I only said it would be worth it.	Mae West	4.43	easy	effortless	hesitant	nonstop	giant
6	Whatever the mind can conceive and believe, it can achieve.	Napoleon Hill	4.43	mind	brain	shouting	pouring	belonging
7	The way I see it, if you want the rainbow, you gotta put up with the rain!	Dolly Parton	4.29	rain	mist	transportation	distribution	refrigerator

Continues on next page …

Table C.1: Top 25 rated inspirational quotes used for achievement primes and the alternative words used for the active achievement prime questions.

ID	Quote	Author	Score	Word to be replaced	Replacement	Alternative words		
8	There is no dishonor in losing the race. There is only dishonor in not racing because you are afraid to lose.	Garth Stein	4.29	dishonor	shame	butter	selection	songs
9	What is now proved was once only imagined.	William Blake	4.29	imagined	thought	macho	male	lopsided
10	If my mind can conceive it, and my heart can believe it - then I can achieve it.	Muhammad Ali	4.29	heart	soul	quarter	cook	pan
11	You're not obligated to win. You're obligated to keep trying. To the best you can do everyday.	Jason Mraz	4.14	best	finest	situation	responsibility	estate
12	The grand essentials to happiness in this life are something to do, something to love, and something to hope for.	George Washington Burnap	4.14	something	stuff	party	comb	worm
13	What's meant to be will always find a way	Trisha Yearwood	4.14	way	path	chicken	paper	sheet
14	Everyone wants to live on top of the mountain, but all the happiness and growth occurs while you're climbing it.	Andy Rooney	4.14	growth	gain	environments	insects	gates
15	The things you do for yourself are gone when you are gone, but the things you do for others remain as your legacy.	Kalu Ndukwe Kalu	4.14	legacy	inheritance	department	emotion	two
16	Do you really want to be happy? You can begin by being appreciative of who you are and what you've got.	Benjamin Hoff	4.14	appreciative	grateful	thirsty	automatic	violet
17	Change the way you look at things and the things you look at change.	Wayne W. Dyer	4.14	things	situations	babies	looks	giraffes

Continues on next page …

Table C.1: Top 25 rated inspirational quotes used for achievement primes and the alternative words used for the active achievement prime questions.

ID	Quote	Author	Score	Word to be replaced	Replacement	Alternative words		
18	A man with outward courage dares to die; a man with inner courage dares to live.	Lao Tzu	4.14	courage	bravery	bits	actors	steps
19	Never be bullied into silence. Never allow yourself to be made a victim. Accept no one's definition of your life, but define yourself.	Harvey Fierstein	4.14	life	existence	snail	door	dog
20	Life isn't about finding yourself. Life is about creating yourself.	George Bernard Shaw	4.14	Life	being	kettle	battle	friend
21	Keep your face always toward the sunshine - and shadows will fall behind you.	Walt Whitman	4.00	shadows	darkness	pianos	vehicles	oatmeel
22	You have power over your mind - not outside events. Realize this, and you will find strength.	Marcus Aurelius	4.00	strength	power	apartment	complaint	office
23	Reputation is what other people know about you. Honor is what you know about yourself.	Lois McMaster Bujold	4.00	Honor	integrity	diamond	guest	writer
24	You never fail until you stop trying.	Albert Einstein	4.00	trying	attempting	featuring	shopping	greeting
25	Always do what you are afraid to do.	Ralph Waldo Emerson	4.00	afraid	scared	ablaze	new	idiotic

D

Main Study Questionnaire

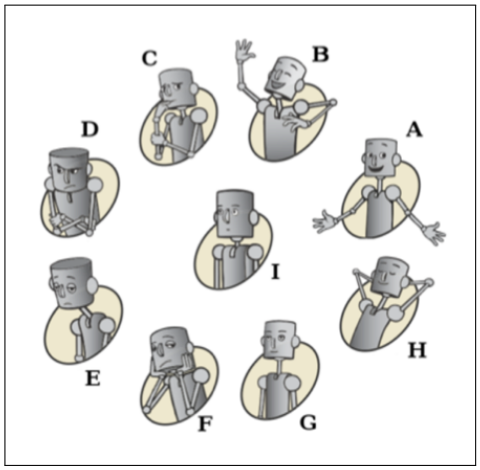


Figure D.1: Pick-A-Mood scale (Desmet, Vastenburg, & Romero, 2016).

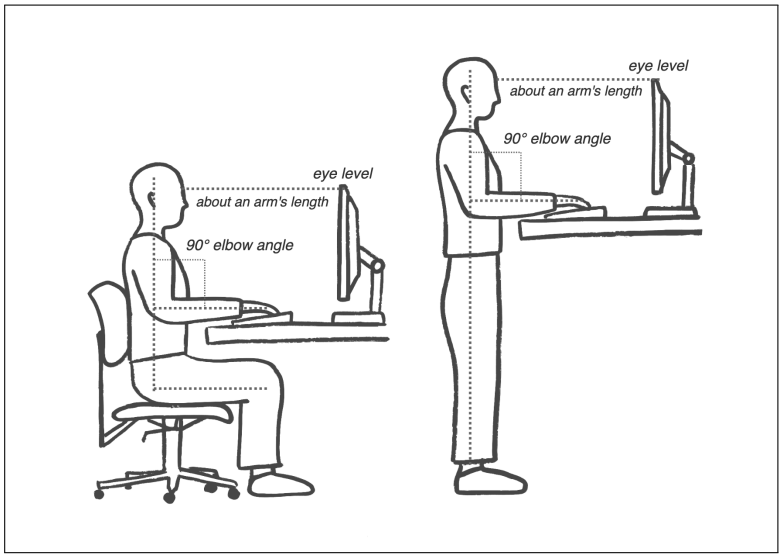


Figure D.2: Proper sitting and standing postures for working behind a desk.

Table D.1: Questions and options used in the first part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
1	May I know your gender?	-	Single-selection	Male Female Other Prefer not to say
2	How old are you?	-	Free text (number, 18+)	
3	In what mood are you today? *Showing Figure D.1 of Pick-A-Mood scale*	-	Single-selection	Cheerful Excited Tense Irritated Sad Bored Calm Relaxed Neutral
4	Which of the following describes the income you earn from **crowdsourced microtasks**?	Which of the following describes the income you earn from **working on a computer**?	Single-selection	Primary source of income Secondary source of income
5	How many hours do you work **on Prolific** each day on average?	How many hours do you work **on a computer** each day on average?	Single-selection	1 hour or less 1 to 3 hours 4 to 6 hours 7 to 9 hours 9 hours or more
6	Please indicate your usual working time **on Prolific** in a day.	Please indicate your usual working time **on a computer** in a day.	Multiple-selection	Early morning (until 9 AM) Mid-morning (after 9 AM, up to 12 noon) Early afternoon (after 12 noon, up to 3 PM) Late afternoon (after 3 PM, until 6 PM) Evening (after 6 PM, until 9 PM) Night (after 9 PM)

Continues on next page …

Table D.1: Questions and options used in the first part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
7	For how long have you been **working on Prolific**?	For how long have you been **doing computer work**?	Single-selection	1 year or less 1 to 3 years 4 to 6 years 7 to 9 years 9 years or more
8	To what extent do you think your current working environment is comfortable in terms of lighting, temperature, humidity, and noise?	-	7-pt Likert-scale	1: Very uncomfortable 2: Uncomfortable 3: Somewhat uncomfortable 4: Neither comfortable nor uncomfortable 5: Somewhat comfortable 6: Comfortable 7: Very comfortable
9	So your current working environment is comfortable/uncomfortable, then do you think it is healthy?	-	7-pt Likert-scale	1: Very unhealthy ⋮ 7: Very healthy
10	To what extent do you think your current working setup and devices are comfortable in terms of layout, seating, displays, and compatibility?	-	7-pt Likert-scale	1: Very uncomfortable ⋮ 7: Very comfortable
11	So your current working setup and devices are is comfortable/uncomfortable, then do you think it is healthy?	-	7-pt Likert-scale	1: Very unhealthy ⋮ 7: Very healthy
	Prime (Quote IDs 1, 5, 9, 13, 17, 21, 25)			
12	Do you consider that you have colleagues **(eg. other crowd workers)**?	Do you consider that you have colleagues?	Single-selection	No Yes
13	Do you share workspaces with your colleagues or work together in a shared work environment?	-	5-pt Likert-scale	1: Always alone 2: Sometimes alone 3: Neither together nor alone 4: Sometimes together 5: Always together

Continues on next page …

Table D.1: Questions and options used in the first part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
14	Do you take some measures to keep yourself healthy? (If so, what do you do?)	-	Free-text	-

Table D.2: Questions and options used in the second part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
15	What is your primary working posture?	-	Single-selection	Sitting Standing Other posture
16	*Showing Figure of proper sitting and standing postures* Looking at these examples of healthy working postures, to what extent do you think your working posture is healthy?	-	7-pt Likert-scale	1: Very unhealthy 2: Unhealthy 3: Somewhat unhealthy 4: Neither healthy nor unhealthy 5: Somewhat healthy 6: Healthy 7: Very healthy
17	*If the posture includes sitting* How often do you use armrests?	-	5-pt Likert-scale	1: Never 2: Sometimes 3: About half the time 4: Most of the time 5: Always
18	*If the posture includes sitting* Can you indicate your sitting position?	-	Single-selection	On the front edge of the chair On the middle of the chair On the back of the chair
	Prime (Quote IDs 2, 4, 10, 14, 18, 22)			
19	*If the posture includes sitting* How often do you use your backrest?	-	5-pt Likert-scale	1: Never ⋮ 5: Always

Continues on next page …

Table D.2: Questions and options used in the second part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
20	How often do you take a break?	-	Single-selection	More frequently than every 30 mins Every 30 mins to 1 hour Every 1 to 2 hours Every 2 to 3 hours Every 3 to 4 hours Less frequently than every 4 hours Never
21	What is the distance between you and your screen?	-	5-pt Likert-scale	1: Too close 2: Somewhat too close 3: Correctly distanced (about an armslength) 4: Somewhat too far 5: Too far
22	Can you indicate the position of the top of your screen?	-	5-pt Likert-scale	1: Too low 2: Somewhat too low 3: Correctly positioned (top of the screen at eye level) 4: Somewhat too high 5: Too high
23	Can you indicate your keyboard/mouse placement?	-	Multiple-selection	Arms less than 90 degrees with keyboard Arms at 90 degrees with keyboard Arms more than 90 degrees with keyboard Arms are supported when using the keyboard Arms less than 90 degrees with mouse Arms at 90 degrees with mouse Arms more than 90 degrees with mouse Arms are supported when using the mouse
	Please tell me how comfortable your different body parts feel on an average day working **on Prolific**.	Please tell me how comfortable your different body parts feel on an average day working **on a computer**.		

Continues on next page …

Table D.2: Questions and options used in the second part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
24	Your eyes?	-	7-pt Likert-scale	1: Very uncomfortable 2: Uncomfortable 3: Somewhat uncomfortable 4: Neither comfortable nor uncomfortable 5: Somewhat comfortable 6: Comfortable 7: Very comfortable
25	What about your head?	-	7-pt Likert-scale	1: Very uncomfortable ⋮ 7: Very comfortable
26	And your neck and shoulders?	-	7-pt Likert-scale	1: Very uncomfortable ⋮ 7: Very comfortable
27	How are your hands/wrists and arms?	-	7-pt Likert-scale	1: Very uncomfortable ⋮ 7: Very comfortable
28	How is your back?	-	7-pt Likert-scale	1: Very uncomfortable ⋮ 7: Very comfortable
29	What about your seat and thighs?	-	7-pt Likert-scale	1: Very uncomfortable ⋮ 7: Very comfortable
30	And your knees and feet?	-	7-pt Likert-scale	1: Very uncomfortable ⋮ 7: Very comfortable

Table D.3: Questions and options used in the third part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
	Type of production and tasks	-		

Continues on next page …

Table D.3: Questions and options used in the third part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
31	How often do you have enough time for **tasks on Prolific**?	How often do you have enough time for **computer work**?	5-pt Likert-scale	1: Never 2: Sometimes 3: About half the time 4: Most of the time 5: Always
32	Do you have to work very fast?	-	5-pt Likert-scale	1: Never ⋮ 5: Always
33	Is **completing tasks on Prolific** emotionally demanding?	Is completing tasks **on your computer** emotionally demanding?	5-pt Likert-scale	1: To a very large extent 2: To a large extent 3: To a moderate extent 4: To a small extent 5: To a very small extent
	Work organization and job content	-		
34	Do you have a large degree of influence on the decisions concerning **completing tasks on Prolific**?	Do you have a large degree of influence on the decisions concerning **computer work**?	5-pt Likert-scale	1: To a very small extent 2: To a small extent 3: To a moderate extent 4: To a large extent 5: To a very large extent
35	Do you have the possibility of learning new things through **completing tasks on Prolific**?	Do you have the possibility of learning new things through **computer work**?	5-pt Likert-scale	1: To a very small extent ⋮ 5: To a very large extent
	Prime (Quote IDs 3, 7, 11, 15, 19, 23)			
36	Do you feel that **completing tasks on Prolific** is meaningful?	Do you feel that **your work on a computer** is meaningful?	5-pt Likert-scale	1: Never ⋮ 5: Always
	Interpersonal relations	-		
37	How often do you get help and support from **Prolific or task requesters**, if needed?	How often do you get help and support from **your employer or IT**, if needed?	5-pt Likert-scale	1: Never ⋮ 5: Always
38	How often do you get help and support from **other workers**, if needed?	How often do you get help and support from **colleagues**, if needed?	5-pt Likert-scale	1: Never ⋮ 5: Always

Continues on next page …

Table D.3: Questions and options used in the third part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
39	How often **do task requesters bonus/message** you because how well you carry out your work?	How often **does your employer contact/bonus** you because how well you carry out your work?	5-pt Likert-scale	1: Never ⋮ 5: Always
40	Is there a good atmosphere between you and **other workers (on either crowdsourcing platforms or other worker forums e.g. Reddit)**?	Is there a good atmosphere between you and **your colleagues**?	5-pt Likert-scale	1: To a very small extent ⋮ 5: To a very large extent
41	In general, would you say your health is excellent, very good, good, fair or poor?	-	Single-selection	Excellent Very good Good Fair Poor
42	While **completing tasks on Prolific**, do you feel full of pep?	While **doing computer work**, do you feel full of pep?	6-pt Likert-scale	1: None of the time 2: A little of the time 3: Some of the time 4: A good bit of the time 5: Most of the time 6: All of the time
43	While **completing tasks on Prolific**, have you been a very nervous person?	While **doing computer work**, have you been a very nervous person?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time
44	While **completing tasks on Prolific**, have you felt so down in the dumps that nothingcould cheer you up?	While **doing computer work**, have you felt so down in the dumps that nothing could cheer you up?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time
45	While **completing tasks on Prolific**, have you felt calm and peaceful?	While **doing computer work**, have you felt calm and peaceful?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time
46	While **completing tasks on Prolific**, do you have a lot of energy?	While **doing computer work**, do you have a lot of energy?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time
47	While **completing tasks on Prolific**, have you felt downhearted and blue?	While **doing computer work**, have you felt downhearted and blue?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time

Continues on next page …

Table D.3: Questions and options used in the third part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
48	While **completing tasks on Prolific**, do you feel worn out?	While **doing computer work**, do you feel worn out?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time
49	While **completing tasks on Prolific**, have you been a happy person?	While **doing computer work**, have you been a happy person?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time
50	While **completing tasks on Prolific**, do you feel tired?	While **doing computer work**, do you feel tired?	6-pt Likert-scale	1: None of the time ⋮ 6: All of the time

Table D.4: Questions and options used in the fourth part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
51	For which part(s) of your body do you think you need some physical exercises?	-	Multiple-selection	Head Neck and shoulders Arms Hands Back Seat Thighs Knees and feet
52	For which aspect(s) of your psychosocial condition do you think you need improvements?	-	Multiple-selection	General health Mental health Vitality Behavioural stress Sense of community Cognitive demand Emotional demand

Continues on next page …

Table D.4: Questions and options used in the fourth part of the main questionnaire.

ID	Original question	Adapted question	Answer type	Options
53	To what extent will you be happy to use a tool that provides breaks/exercises/treatments to improve your overall health while **completing crowdsourcing tasks**? *Optional:* Can you tell me why?	To what extent will you be happy to use a tool that provides breaks/exercises/treatments to improve your overall health while **working on a computer**? *Optional:* Can you tell me why?	5-pt Likert-scale & Free-text	1: Very unhappy 2: Somewhat unhappy 3: Neither happy nor unhappy 4: Somewhat happy 5: Very happy
54	What features would you like to see in such a tool, considering that they are all backed by scientific evidence? *Optional:* Can you tell me why?	-	Multiple-selection & Free-text	Simple breaks Physical exercises Treatments for mental health
55	What type of working modes of this tool would you prefer? *Optional:* Can you tell me why?	-	Multiple-selection & Free-text	Pull (asking for interventions when you want to) Push (enabling a tool to actively sending interventions to you)
56	Do you think that you should get paid while you are using the tool to take some breaks/exercises/treatments? *Optional:* Can you tell me why?	-	Single-selection & Free-text	No Yes
	Prime (Quote IDs 4, 8, 12, 16, 20, 24)			
57	How would you like to receive interventions (breaks/exercises/treatments)?	-	Single-selection	Between task batches/Between tasks Within task batches/Within tasks Outside of the platform/Outside of the digital work platform
58	How long would you like the interventions (breaks/exercises/treatments) from the tool to be?	-	Single-selection	5 mins 10 mins 15 mins 20 mins 30 mins

Continues on next page …

Table D.4: Questions and options used in the fourth part of the main questionnaire.

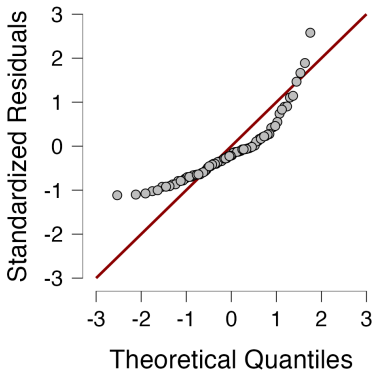
ID	Original question	Adapted question	Answer type	Options
59	How frequently would you like to take breaks/exercises/treatments from such a tool?	-	Single-selection	More frequently than every 30 mins Every 30 mins to 1 hour Every 1 to 2 hours Every 2 to 3 hours Every 3 to 4 hours Less frequently than every 4 hours Never
60	Who do you think should be responsible for developing the tool? Please check all that apply.	-	Multiple-selection	Crowdsourcing platforms/Digital work platforms Task requesters/Employers Academic researchers Crowd workers/Employees Third-parties
61	Do you have any other comments, remarks, or suggestions? Your thoughts are valuable to us.	-	Free-text	-

E

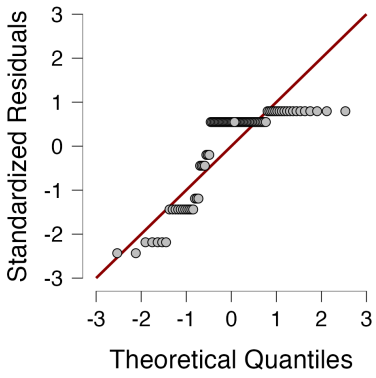
Tests for Normality

E.1. Questionnaire Variant

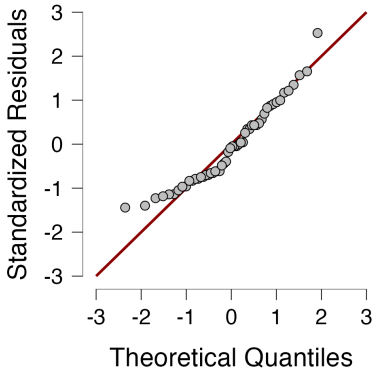
E.1.1. Intrinsic Motivation



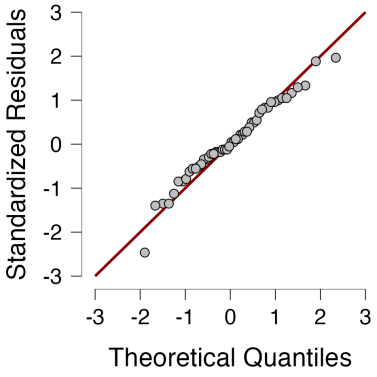
(a) Q-Q plot for duration.



(b) Q-Q plot for surpassed page.

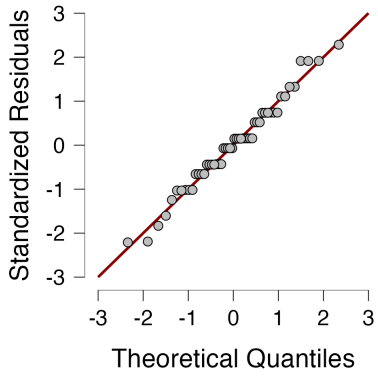


(c) Q-Q plot for perceived workload.

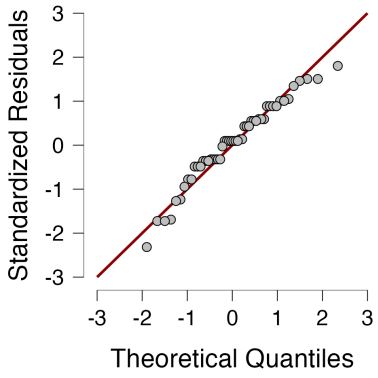


(d) Q-Q plot for user engagement mean.

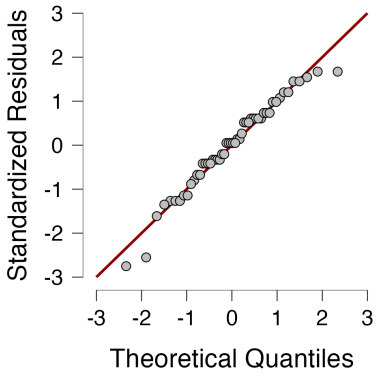
Figure E.1: Test of normality for intrinsically motivated respondents using Q-Q plots.



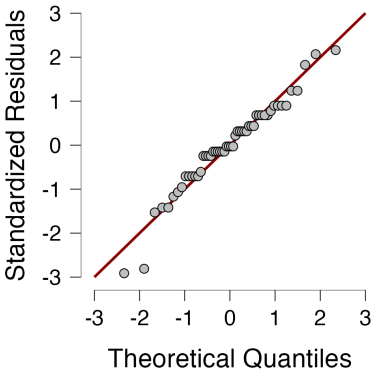
(e) Q-Q plot for focused attention.



(f) Q-Q plot for perceived usability.



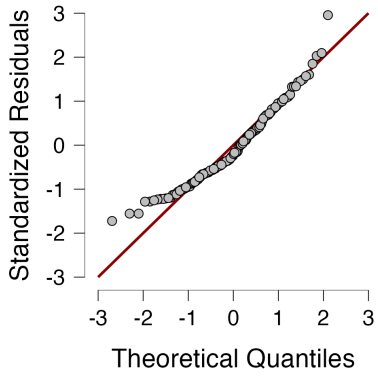
(g) Q-Q plot for aesthetic appeal.



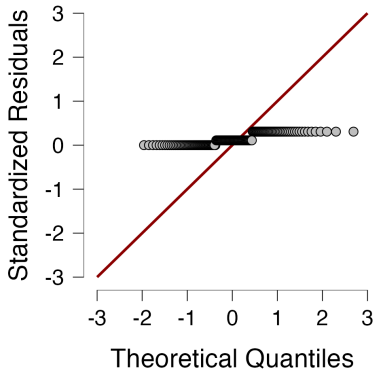
(h) Q-Q plot for reward.

Figure E.1: Test of normality for intrinsically motivated respondents using Q-Q plots.

E.1.2. Extrinsic Motivation

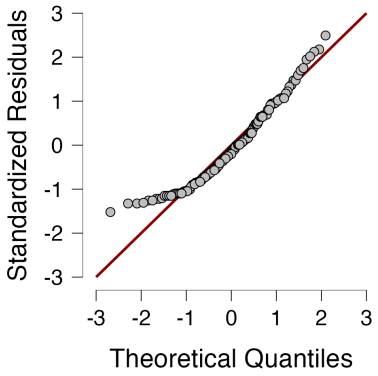


(a) Q-Q plot for duration.

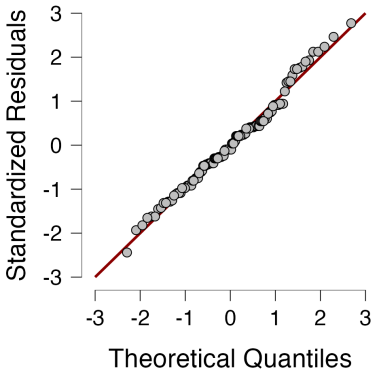


(b) Q-Q plot for surpassed page.

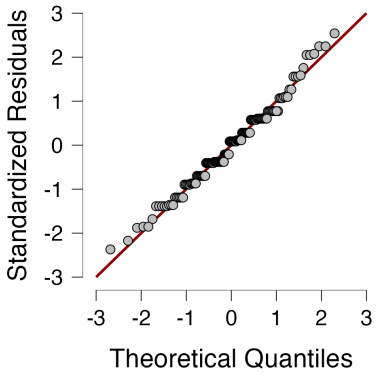
Figure E.2: Test of normality for extrinsically motivated respondents using Q-Q plots.



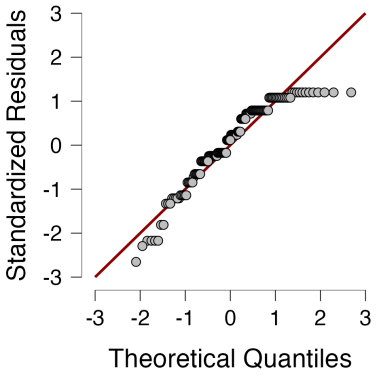
(c) Q-Q plot for perceived workload.



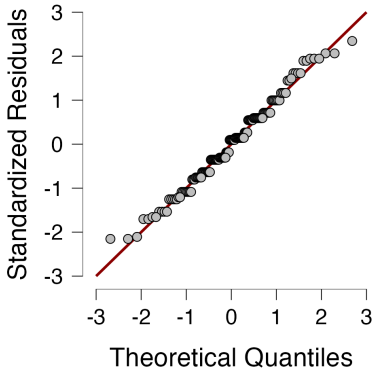
(d) Q-Q plot for user engagement mean.



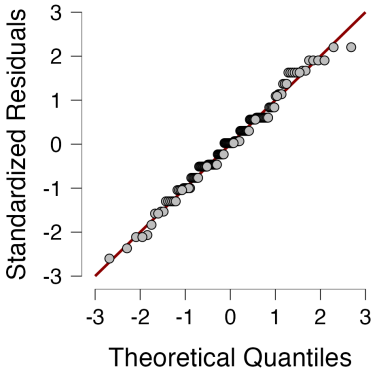
(e) Q-Q plot for focused attention.



(f) Q-Q plot for perceived usability.



(g) Q-Q plot for aesthetic appeal.



(h) Q-Q plot for reward.

Figure E.2: Test of normality for extrinsically motivated respondents using Q-Q plots.

E.2. Respondent Group

E.2.1. Passive Prime

Table E.1: Test of normality for respondents shown the passive prime questionnaire variant using Shapiro-Wilk.

Dependent variable	Respondent group	W	p
Duration (in seconds)	Extrinsic	0.8917	< 0.001
	Intrinsic	0.8554	0.0001
Surpassed page	Extrinsic	0.2343	< 0.001
	Intrinsic	0.7070	< 0.001
TLX Sum	Extrinsic	0.8876	< 0.001
	Intrinsic	0.9299	0.154
UES Mean	Extrinsic	0.9748	0.397
	Intrinsic	0.9684	0.721
UES FA	Extrinsic	0.9742	0.379
	Intrinsic	0.9568	0.481
UES PU	Extrinsic	0.8821	< 0.001
	Intrinsic	0.9034	0.048
UES AE	Extrinsic	0.9343	0.011
	Intrinsic	0.8967	0.036
UES RW	Extrinsic	0.9633	0.145
	Intrinsic	0.9230	0.113

E.2.2. Active Prime

Table E.2: Test of normality for respondents shown the active prime questionnaire variant using Shapiro-Wilk.

Dependent variable	Respondent group	W	p
Duration (in seconds)	Extrinsic	0.9320	0.009
	Intrinsic	0.8260	< 0.001
Surpassed page	Extrinsic	0.1311	< 0.001
	Intrinsic	0.7538	< 0.001
TLX Sum	Extrinsic	0.9218	0.004
	Intrinsic	0.8582	0.014
UES Mean	Extrinsic	0.9701	0.280
	Intrinsic	0.8742	0.032
UES FA	Extrinsic	0.9501	0.047
	Intrinsic	0.9671	0.790
UES PU	Extrinsic	0.8495	< 0.001
	Intrinsic	0.8960	0.069
UES AE	Extrinsic	0.9501	0.048
	Intrinsic	0.9541	0.557
UES RW	Extrinsic	0.9462	0.034
	Intrinsic	0.9118	0.124