TU WIEN Informatics

SQT: A Tool for the Automate Measurement of Respondent Behaviour and Response Quality in Online Surveys

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering/Internet Computing

eingereicht von

Stefan Biegler
00725380

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung
Betreuer: Thomas Grechenig

Wien, 10.09.2020

Unterschrift Verfasser | Unterschrift Betreuer

Technische Universität Wien
Karlsplatz 13 | 1040 Wien | +43-1-58801-0 | www.tuwien.at

TU WIEN Informatics

SQT: A Tool for the Automated Measurement of Respondent Behaviour and Response Quality in Online Surveys

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering/Internet Computing

by

Stefan Biegler
00725380

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Thomas Grechenig

Vienna, 10.09.2020

Unterschrift Verfasser | Unterschrift Betreuer

Technische Universität Wien
Karlsplatz 13 | 1040 Wien | +43-1-58801-0 | www.tuwien.at

SQT: A Tool for the Automated Measurement of Respondent Behaviour and Response Quality in Online Surveys

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering/Internet Computing

eingereicht von

Stefan Biegler

00725380

ausgeführt am

Institut für Information Systems Engineering

Forschungsbereich Business Informatics

Forschungsgruppe Industrielle Software

der Fakultät für Informatik der Technischen Universität Wien

Betreuung: Thomas Grechenig

Technische Universität Wien, Forschungsgruppe INSO

A-1040 Wien ▪ Wiedner Hauptstr. 76/2/2 ▪ Tel. +43-1-587 21 97 ▪ www.inso.tuwien.ac.at

Erklärung zur Verfassung der Arbeit

Stefan Biegler
Helenenstrasse 70a/Haus 4, 2500 Baden

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, am 10.09.2020 ----------
(Ort, Datum)

(Unterschrift Verfasser)

Danksagung

Zu allererst möchte ich mich bei meinem Betreuer Prof. Dr. Thomas Grechenig und bei Johannes Harms für die Unterstützung bei der Erstellung der Diplomarbeit bedanken. Weiters möchte ich auch meinen Dank Christoph Wimmer aussprechen, der mit treffendem Feedback die Arbeit verbessert hat. Auch bei meiner Familie möchte ich mich bedanken, angefangen bei meiner Frau Manuela die mir ermöglicht hat meine Arbeit in Ruhe fertigzustellen. Zusätzlich auch bei meinen Kindern Katharina und Konstantin, die mich bei mancher Schreibblockade aufgeheitert und mich motiviert haben. Schlussendlich möchte ich mich auch bei meiner Mutter Margit bedanken die öfter einmal auf die Kinder aufgepasst hat, um mir ein ruhiges Arbeiten zu ermöglichen.

Kurzfassung

Datenqualität ist essentiell, um valide und zuverlässige Ergebnisse aus einer Onlineumfrage zu erhalten. Verwandte Arbeiten zeigen, dass Nutzerantworten, welche eine schlechte Qualität aufweisen, die Datenqualität sowie auch das Ergebnis einer Onlineumfrage negativ beeinflussen können. Dieser Umstand ist für wissenschaftliche wie auch für kommerzielle Zwecke sehr problematisch. Ein Grund für schlechte Datenqualität in Onlineumfragen ist negatives Nutzerverhalten. Dieses unerwünschte Verhalten wurde bereits in vergangenen Arbeiten definiert und gemessen. Aktuell herrscht ein Mangel an einfach zu handhabbaren Programmen, welche negatives Nutzerverhalten automatisch und valide messen können. Mit Hilfe solcher Programme könnte der Umfragen-Analyst, Nutzerantworten mit geringer Datenqualität adaptieren oder diese aus dem Ergebnis der Onlineumfrage entfernen.

Diese Diplomarbeit beschreibt die Planung, technische Umsetzung und die empirische Evaluierung eines Survey Quality Tools (SQT), welches das negative Nutzerverhalten automatisch misst. SQT wurde in zwei Fallstudien empirisch evaluiert. Fallstudie eins hatte das Ziel SQT zu validieren. Fallstudie zwei hatte das Ziel, Feedback über die Usability, die praktische Anwendbarkeit und den Nutzen von SQT zu evaluieren.

Die Ergebnisse zeigten, dass SQT ein valides Tool, für das automatische Messen des negativen Nutzerverhaltens ist und in der praktischen Anwendung sich als nützlich erwies. Der Beitrag dieser Diplomarbeit zur Forschung an Online Umfragen ist somit ein erwiesenermaßen valides Programm (SQT), welches automatisch das negative Nutzerverhalten basierend auf verschiedenen Careless Response Patterns misst.

Zukünftige Arbeiten können dieses Tool verwenden. Für die weitere Forschung wäre es interessant zu vergleichen, ob Artificial Intelligence basierte Ansätze negatives Nutzerverhalten eventuell noch genauer detektieren können und ob neue Careless Response Patterns, welche noch nicht in verwandten Arbeiten aufscheinen, dadurch definiert werden können.

Keywords: *Negatives Benutzerverhalten, Careless Response Patterns, schlechte Datenqualität, Online Umfragen*

Abstract

Data quality is essential for valid and reliable results in online surveys. It has been shown that low-quality response data can not only bias results but also makes spurious effects significant, which clearly is highly problematic for both scientific and marketing research. One reason for low-quality response data in online surveys is negative respondent behaviour. Several measures for this unwanted behaviour had been defined in previous work. Nonetheless, there is currently a lack of validated and easy-to-use software tools that make it easy to detect unwanted behaviour and allow survey analysts to correct or exclude low-quality responses.

This thesis describes the development, implementation and empirical evaluation of a survey quality tool (SQT), which automatically detects negative respondent behaviour. The tool was evaluated in two empirical studies. The first study sought to validate SQT. The second study had the goal to collect qualitative and quantitative feedback and lessons learned regarding the tool's practical applicability, usability and usefulness.

Results confirmed the validity of SQT. Results also confirmed the tool's practical applicability and usefulness, i.e., it was found to be useful and easy-to-use. The contribution of this thesis hence is a validated, useful and practical survey quality tool (SQT), which automatically detects careless response patterns of negative respondent behaviour and warns survey analysts about low-quality responses.

Future work can use this tool and investigate new methods for measuring negative respondent behaviour and compare those new measures with the established ones that are already implemented in the tool. Specifically, it would be interesting to use machine learning to classify respondent behaviour and to potentially identify new careless response patterns that are not yet covered in related work.

Keywords: *careless response, negative respondent behaviour, low-quality, online surveys*

Table of Contents

Table of Figures

List of Tables

Table of Formulas

1 Introduction

Surveys play a major role in collecting data (e.g., for products, in health care etc.) and their resulting data is used for further statistical evaluation. One reason why so many surveys exist is, they are cheap to create and easy to manage. Due to the fact that the World Wide Web has grown very fast in recent years, “old style” surveys, which were distributed per phone or mail, have lost importance and online surveys have gained popularity.

One problem of online surveys is that they are dull to fill out and boring (KROSNICK, 1991). Because of this circumstance respondents have low motivation to answer online surveys and therefore provide low-quality response data (BIEMER, 2010). Low-quality response data has a negative effect on the survey result and increases total survey error (BIEMER, 2010). In order to prevent such unwanted effects, it is essential for survey analysts to measure respondent behaviour and response quality.

This thesis describes the development, implementation and empirical evaluation of a new software tool named “survey quality tool” (SQT). The goal of this tool is to automatically detect negative respondent behaviour and warns survey analysts about low-quality responses. With the help of these warnings the survey analysts can exclude low-quality responses from the final survey result in order to improve the overall data quality.

1.1 Problem Statement

Being able to detect quality problems in survey responses is very important in at least two situations: when conducting surveys, as well as for research on innovative survey design.

From the perspective of those who conduct surveys, respondents may display undesired behaviour that can bias survey results. Such behaviour can be classified using the total error framework (GROVES, ET AL., 2011) based on the type of statistical error that it produces. Broadly speaking, behaviour that leads to inaccurate answers produces measurement error, whereas missing responses produce nonresponse error if they lead to a certain part of the target population being systematically underrepresented. In order to take corrective action and reduce statistical error in survey results, it is very important for survey analysts to detect problematic behaviour. This need

is currently not well met because software tools that detect negative respondent behavior are not readily available – more information in **Chapter 2.2.3**.

Negative respondent behaviour describes various careless response patterns, including for example straightlining (when a respondent selected a significant number of same answers in series per question matrix) (KROSNICK, 1991), missing answers, when a respondent of an online survey does not respond or only partially completes the current online survey (BIEMER, 2010), (GUIN, BAKER, MECHLING, & RUYLE, 2012) and many more (HUANG, CURRAN, KEENEY, POPOSKI, & DESHON, 2012), (AKBULUT, 2015), (WIMMER, BIEGLER, HARMS, KAPPEL, & GRECHENIG, 2018), (FRANCAVILLA, MEADE, & YOUNG, 2018). Negative respondent behaviour has a negative effect on the overall data quality (BIEMER, 2010) and on total survey error of the current survey (GROVES, ET AL., 2011). This is especially problematic for scientific or marketing research, which clearly expects and requires high-quality responses (KROSNICK, 1991). Most survey analysts ignore the possible occurrence of negative respondent behaviour (HUANG, LIU, & BOWLING, 2015) and therefore, work with low-quality response data. Using such low-quality response data can lead to false research findings (AKBULUT, 2015).

Innovative survey designs were created to improve respondent behaviour and data quality. Researchers in this area must be cautious, because their designs may also produce unwanted effects; for this reason, it is essential to take detailed behavioural measures. One example of innovative survey design are gamified online surveys, which are characterized by the use of game design elements in the specific non-game context of online surveys. Gamified online surveys have the potential to increase participation and data quality (HAMARI, KOIVISTO, & SARSA, 2014), (CECHANOWICZ, GUTWIN, BROWNELL, & GOODFELLOW, 2013), (DOLNICAR, GRÜN, & YANAMANDRAM, 2013), (PULESTON, 2011). Despite demonstrated positive effects, not all gamified surveys have achieved better data quality. For example, HAMARI, ET AL., (2014) experienced that positive effects of gamification depend on the context in which gamification is used and may not be long-term. Furthermore, gamification depends on several external factors like the motivation of the respondent, which is not in the sphere of the survey creator (HAMARI, ET AL., 2014). In another scientific work (KEUSCH & ZHANG, 2017) it was highlighted that gamification does not increase the completion rate and had only small positive impact on the enjoyment of the online survey.

The need for detailed measures of respondent behaviour has also been experienced by the author of this thesis in a previous work, where a gamified survey was evaluated regarding data quality (HARMS, BIEGLER, WIMMER, KAPPEL, & GRECHENIG, 2015). In this work, the gamified design produced slightly different survey results, as compared to a conventional design. There were two possible explanations for this situation. The gamified design could have worked as intended to reduce negative respondent behaviour, but it could also have introduced additional undesired behaviour. Detailed behavioural measures were needed in order to tell the difference and identify the reason for observed differences.

1.2 Research Goals

The main research goal of this thesis is to develop, implement and empirically evaluate a survey quality tool (SQT), which automatically detects negative respondent behaviour with the help of careless response patterns. Based on the research goal the author of this thesis defined two research questions:

Research Question one: “Does SQT provide valid output?”.
Research Question two: “What is the usefulness, practicability and usability of SQT?”.

1.3 Methodical Approach

Methodically, this thesis (see **Figure 1**) started with a literature review to define careless response patterns, which describe negative respondent behaviour. Thereafter, a requirement analysis based on the literature review and the experience of the author was executed. The literature review and the requirement analysis resulted in a mockup prototype for the GUI and thereafter in a functional prototype of SQT. Empirical study one had the goal to provide answers for research question one: "Does SQT provide valid output?". Empirical study two had the goal to provide answers to research question two: "What is the usefulness, practicability and usability of SQT?". To fulfil the goal of empirical study two, SQT was applied to real-world data during Maximilian Stöchle's scientific work (STÖRCHLE, 2020), generating lessons to learn for future work and to show the usability, practicability and usefulness of SQT.

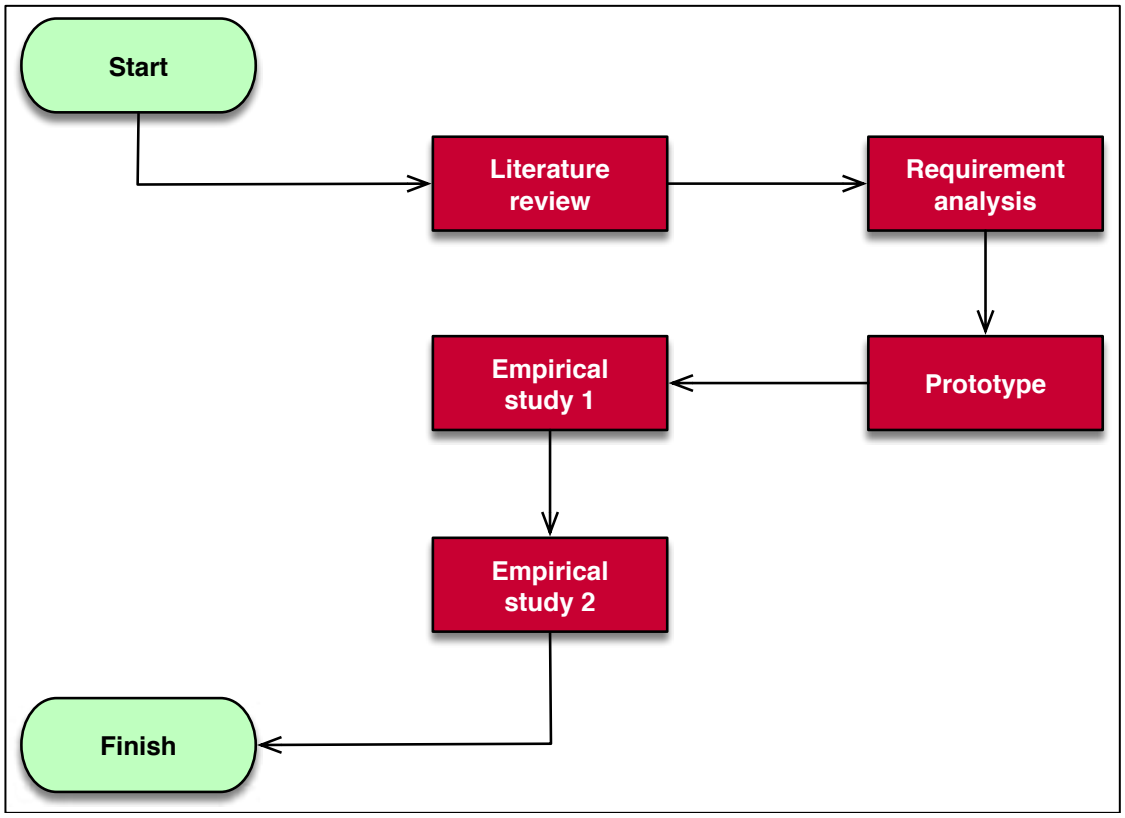


Figure 1. Methodical approach.Methodically this thesis started with a literature review to define careless response patterns, which describe negative respondent behaviour. With the help of the literature review and the requirement analysis the functional prototype of SQT was developed. Empirical study one validated SQT with unit tests and the output of SQT was validated against manual judgment. Empirical study two applied real-world data to SQT to receive lessons to learn for further improvement and to show the usefulness, usability and practicability of SQT.

The author of this thesis started the methodical approach by defining careless response patterns. These definitions were derived from a literature review (see **Appendix I**) and from the knowledge of Johannes Harms (senior designer), Christoph Wimmer (senior designer) and the author of this paper (junior designer), who lay the groundwork for this thesis (WIMMER, ET AL., 2018). In the next step of the methodical approach, the different technical and data protection requirements – like easy to use, open source license etc. – for a tool like SQT were defined.

After the definition of the careless response patterns and the requirements, SQT was developed as a web application in an iterative design and engineering process. At the beginning, a mockup prototype for the GUI was created (see **Appendix II**). Thereafter, a functional prototype was developed, which covered the previously defined requirements and careless response patterns.

The functional prototype was evaluated in two empirical studies.

Empirical study one was designed to formally and empirically investigate the validity of SQT (see **Figure 2**). The validity of SQT can be split in two parts.

Firstly, the definition of careless response patterns (see **Chapter 2.2.2**). The definition can be assumed to be valid because the careless response patterns were derived from scientific work.

The second part was the internal validation of SQT. This validation was performed by a two-step validation process. The first step of this process had the goal to validate the detection of the different careless response patterns by SQT. This was done by defining unit tests for each careless response pattern. In the second step, of the validation process, the output of SQT – detection of negative respondent behaviour per response – was validated. This was done by validating a pre-existing online survey (HARMS, ET AL., 2015) with SQT and by manual judgment. The pre-existing online survey was about sports and health related behaviour among teenagers and young adults. This survey was chosen because the survey's questions are easy to understand and an answer without requiring a domain specific expert knowledge can be given. Two designs exist for this survey, one conventional and one gamified, as created in a previous study (HARMS, ET AL., 2015).

Empirical study two was designed to gather experience and lessons learned regarding the validity of postulated requirements, possible missing requirements, and practical applicability, usefulness and usability of SQT. Therefore, SQT was provided to Maximilian Störchle, who used SQT for an online survey during his scientific work (STÖRCHLE, 2020). During the second empirical study, response data was processed by SQT, allowing Maximilian Störchle to interpret the response data in a deep way, i.e., to get information about each response and to assess the quality of the given

data set. SQT allows to dismiss low-quality data from the overall evaluation and to export data into a .CSV file for further statistical investigation (e.g. SPSS[1]). After Maximilian Störchle had finished his scientific work, he was interviewed by the author of this thesis. During this interview Maximilian Störchle gave feedback and rated SQT in terms of usefulness, practicability and usability.

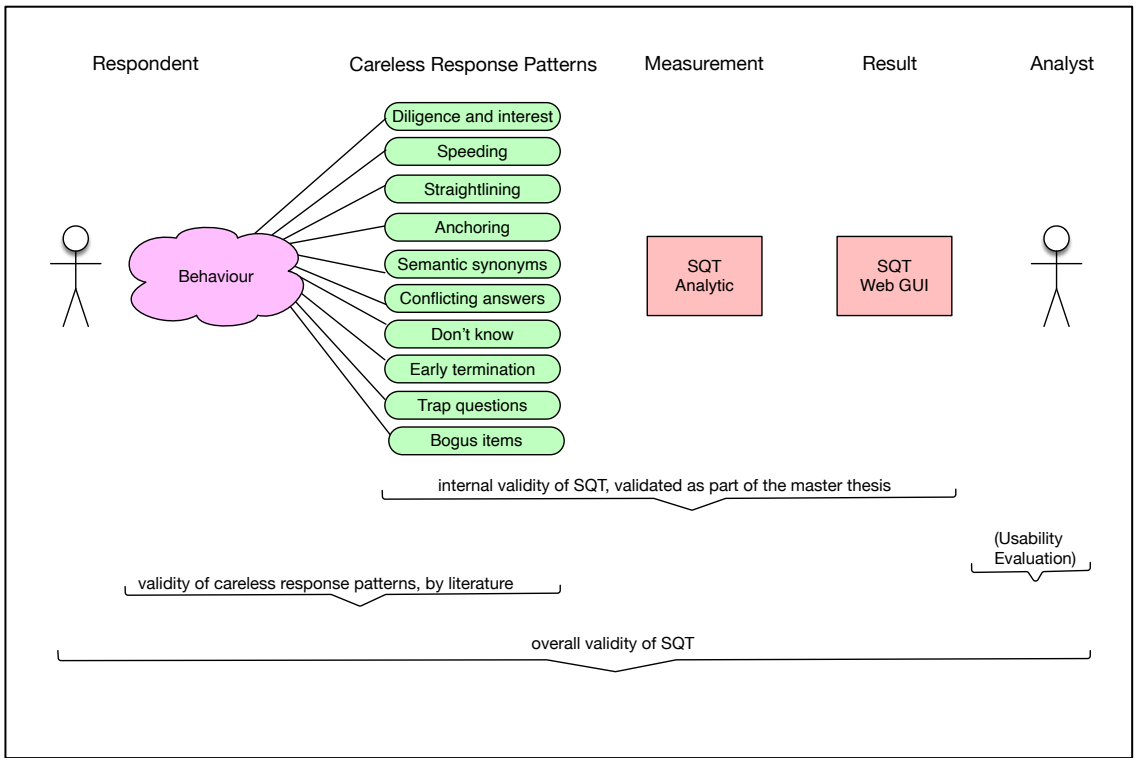


Figure 2. Validity of SQT. The validity of SQT was shown by the internal validity – unit tests of the implemented careless response patterns in SQT and an output validation of SQT by manual judgment – and the validity of negative respondent behaviour – supported by the literature resulting from the literature review. Furthermore, the usability of SQT was shown by distributing SQT to Maximilian Störchle, who used SQT for his scientific work (STÖRCHLE, 2020).

[1] https://www.ibm.com/analytics/at/de/technology/spss/ (12.05.2020)

1.4 Structure of Work

The theoretical part of this thesis will be explained in **Chapter 2**. It consists of an overview about data quality in online surveys – including total survey error, the total error framework, some background information about negative respondent behaviour, state-of-the art tools to detect careless response patterns and innovations in survey design, e.g. gamification.

Chapter 3 contains the practical part of this thesis starting with an introduction to SQT. Furthermore, the literature review, the definition of the requirements and the careless response patterns will be explained. The chapter ends with a detailed overview about the use cases of SQT.

Two different empirical studies will be presented. In **Chapter 4**, empirical study one describes the process how SQT was validated. In **Chapter 5** empirical study two describes the application of real-world data on SQT generating lessons to learn for future work and to show the usability, practicability and usefulness of SQT.

The results of empirical study one and two will be presented and discussed in **Chapter 6** and **7**. **Chapter 8** will give the reader an outlook and possible future work for SQT.

2 Data Quality in Online Surveys

The goal of the total error framework is to detect the sources of statistical error in online surveys. One source, which affects statistical error and has negative effects on the overall statistical result of the survey, is negative respondent behaviour (GROVES, ET AL., 2011). Negative respondent behaviour can lead, in the worst case, to false result findings (AKBULUT, 2015). To weaken the influence of the sum of all possible sources of statistical error that can arise in an online survey – the so called total survey error (see **Chapter 2.1**) –two techniques will be introduced in this chapter. **Chapter 2.2** explains the detection of different careless response patterns, including tools, which are already available to perform this task. This technique can be used to asses and dismiss low-quality survey responses from the statistical evaluation. Another technique to weaken total survey error is to improve the appearance of the survey. This can be done by adding game elements to a non-game context like an online survey (gamification), which will be explained in **Chapter 2.3** (SCHACHT, KEUSCH, BERGMANN, & MORANA, 2017).

2.1 Total Survey Error

Total survey error is defined as the sum of all possible sources of statistical error that can arise in an online survey (compare **Figure 3**). Total survey error describes the deviation of the survey result from its underlying true value (BIEMER, 2010) and contains two error classes, sampling and non-sampling errors. Both of them can occur anytime during the lifecycle of an online survey. There are many factors promoting both error classes. One factor can be the bad design of the online survey (BIEMER, 2010). Sampling and non-sampling errors increase total survey error and therefore effect the overall outcome of the survey in a negative way. This can finally lead to false interpretations of the survey result (GROVES, ET AL., 2011).

The main goal for survey designers is to reduce total survey error and as a consequence get high-quality survey results. Creating a survey without any statistical error is practically impossible (BIEMER, 2010), because too much effort – time and money – must be spent to achieve this goal. Therefore, the main goal for the survey designer is to make a trade-off between reducing total survey error to an acceptable value and effort in time respectively money. To help the survey designer performing such a trade off the total survey error paradigm - also called the total error framework - (BIEMER, 2010), (GROVES, ET AL., 2011) can be used.

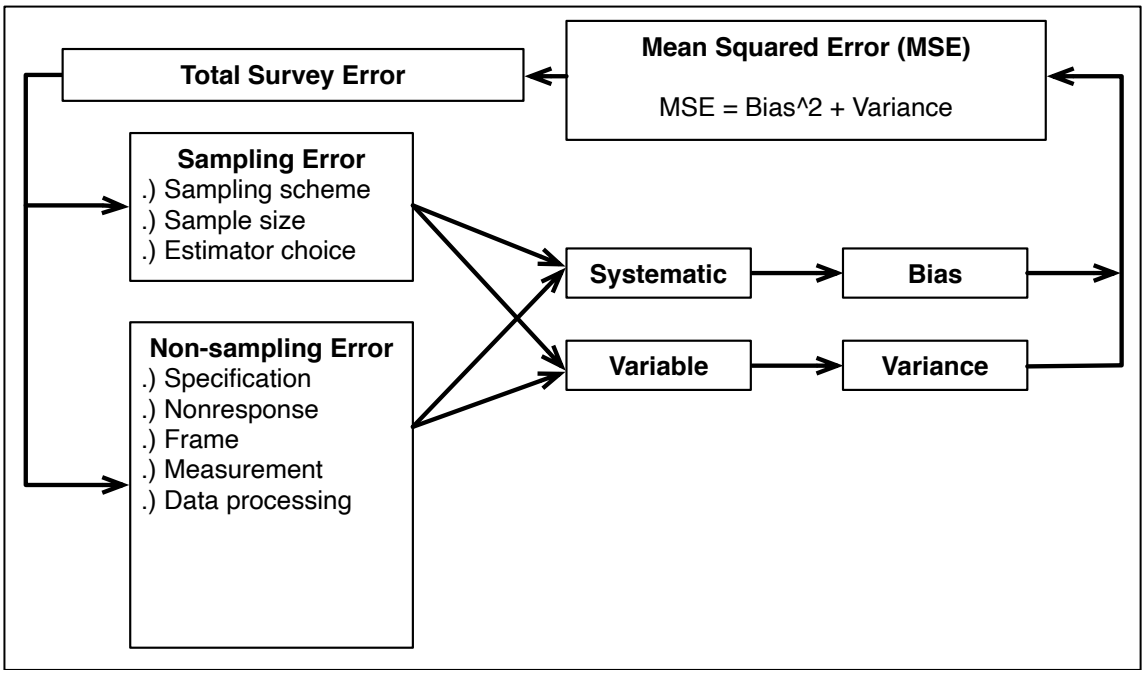


Figure 3. Total survey error. (BIEMER, 2010). Definition of the different components – sampling and non-sampling error – of total survey error.

2.1.1 Total Error Framework

The total error framework has the goal to help the survey designer to understand the different survey steps and to indicate where the different statistical errors can appear. This insight enables survey designers to counteract the different statistical errors and as a result reduce total survey error for a given survey. The total error framework including its components are shown in **Figure 4.**

First of all, the total error framework consists of two components. Ovals – defined as quality concepts – and rectangles – which describes the various survey steps every survey designer has to perform (GROVES, ET AL., 2011).The different survey steps are divided in a measurement and a representation category.

The measurement category starts with defining the *construct* of the survey. The construct is an abstract element describing the topic and the expected result of the survey. For example, the construct of an online survey can be defined as “Health status of Austrian citizens”. After defining the construct, the *measurement step* will be performed. During this step the survey designer defines, how the survey will get (measure) the needed data from the respondents to answer the construct. In the given example, about the health status of Austrian citizens, there exist many measurements – also called survey questions – e.g. the age of the respondent, blood pressure, etc. After

defining the different measurements, the *respondent* of the survey will answer the different questions and generate therefore a result. The answering of the survey questions, by the respondents, can be done intuitive or from previous experience. It is also possible that respondents give no answer at all, which will make a further evaluation even harder. The last step in the measurement part is the *edited response step*. During this step, validity checks will be performed to dismiss responses from the statistical evaluation before the final result is used for further evaluation (GROVES, ET AL., 2011).

The first step in the representation category defines the *target population* for the current survey. A target population is defined as a set of finite size, which will be used in the survey (GROVES, ET AL., 2011). In the example about the "Health status of Austrian citizens", the target population can be described as all Austrian citizens. After the definition of the target population, the *sampling frame* will be defined. The most common methods for defining the sampling frame are, by telephone numbers or by a country map (GROVES, ET AL., 2011). In the *sample step*, a sample will be selected from the sampling frame. This sample will be selected to participate in a survey. The *respondents step* describes the result of the different respondents during the survey. There are many possibilities, for example one respondent stops the survey before it actually ends and therefore left some questions unanswered (nonresponse). It must be defined how to handle such missing data or how to categorize response and nonresponse. In the *postsurvey adjustment step* some estimates, like weighting or imputation (GROVES, ET AL., 2011), are made to improve the quality, either to dismiss nonresponses or to resolve some coverage problems.

In contrast to the rectangles the ovals are situated between each of the successive steps of a survey (GROVES, ET AL., 2011). The gap between two steps correspond to statistical error. The next part of this chapter will explain the different error components in more detail.

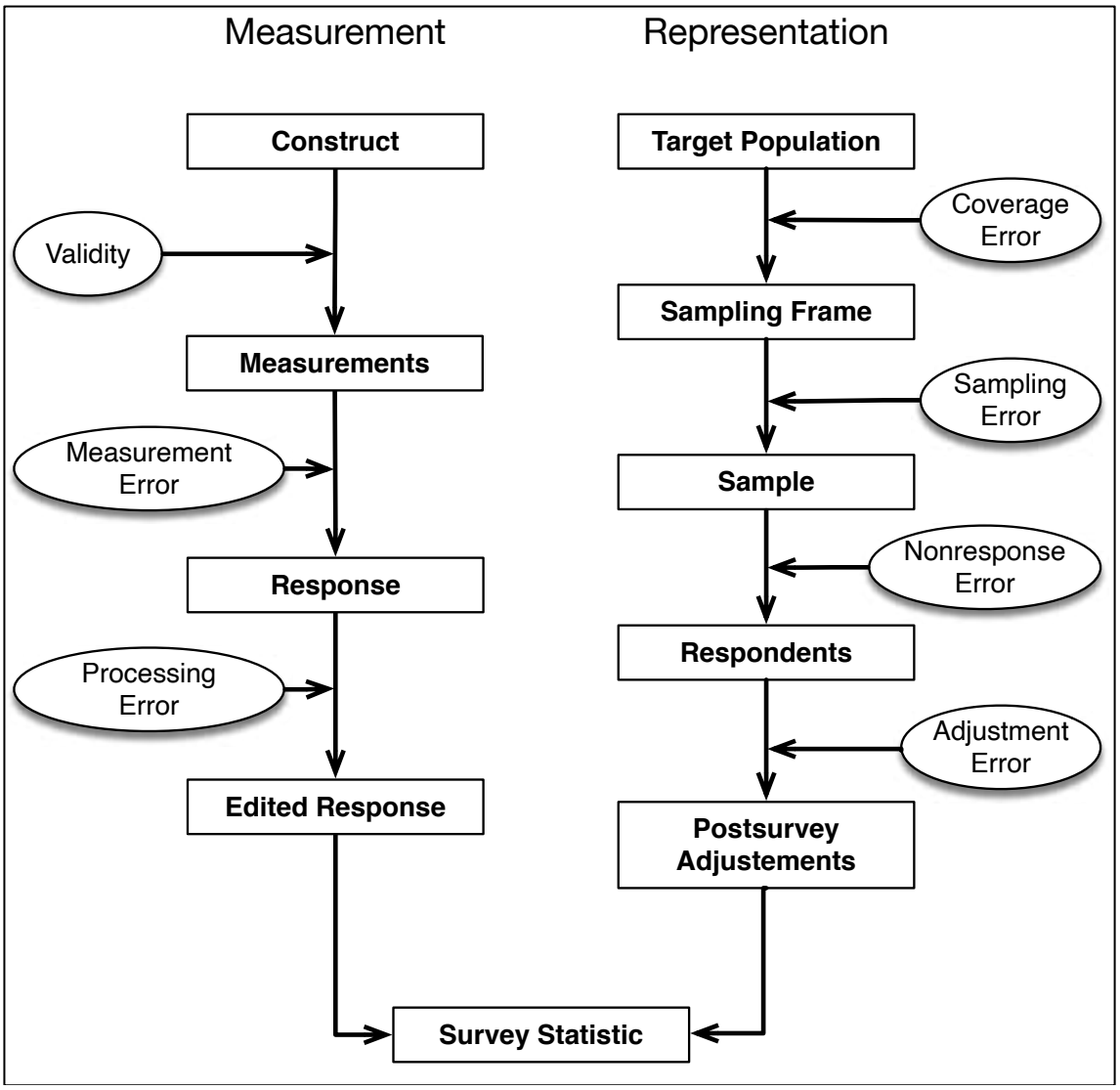


Figure 4. Total error framework. (GROVES, ET AL., 2011). Shows the measurement and representation category of the total error framework. Possible errors between two survey steps are shown in the ovals, which influence total survey error.

2.1.2 Validity of Survey Results

The validity highlights all deviations from a given true value and describes, how the measures of a survey (questions) reflect the underlaying construct (see **Formula 1**) (GROVES, ET AL., 2011).

$$Y_i = \mu_i + \in_i$$

Formula 1. Validity of survey results. (GROVES, ET AL., 2011). The error ($\in_i$) describes the difference between the value of measurement (Y_i) and the value of construct (μ_i).

2.1.3 Measurement Error

Many factors may influence measurement error. Some of these factors are, survey questions, interviewer performance (can be poorly prepared or misunderstood), respondent behaviour (their behaviour can purposely or unintentionally lead to incorrect data (BIEMER, 2010)), survey design (poorly designed survey), etc. (KÜHNE & KROH, 2018). According to BIEMER (2010) measurement error has a major influence on total survey error. Therefore, the survey designer has to keep this error as small as possible.

2.1.4 Processing Error

Processing error is situated after the data collection and before the estimation step (GROVES, ET AL., 2011). Factors influencing the processing error are, response editing, coding, assigning survey weights etc. (BIEMER, 2010). A common example for possible processing error is the problem of evaluating free text answers. To evaluate a free text answer the survey designer has to code the answer of the respondent into different classes. If a "wrong" coding was applied during the transformation step, processing error will negatively influence the survey results (GROVES, ET AL., 2011).

2.1.5 Coverage Error

Coverage error may be part of statistical error, if there is not a one to one relation between the target population and the following sample frame (GROVES, ET AL., 2011). This can be the case, if the survey designer, for example would invite only respondents, who have a valid email address, to participate in the online survey. Therefore, all other respondents, without a valid mail address, will be dismissed from the survey and won't get a chance to be picked as a respondent (GROVES, ET AL., 2011).

2.1.6 Sampling Error

Over or under representation of a specific respondent group with specific attributes can lead to sampling error (GROVES, ET AL., 2011).

2.1.7 Nonresponse Error

There exist two types of nonresponse errors –unit and item nonresponse error. On the one hand, unit nonresponse error is defined, if the respondent didn't answer any question of the given survey. On the other hand, item nonresponse error is defined, if the respondent skipped one or more questions (BIEMER, 2010). Both categories of nonresponse error have a negative impact on the survey result quality.

2.1.8 Adjustment Error

Adjustment error may be part of total statistical error during post survey adjustments. Such adjustments will be performed by the survey designer to counteract the coverage, sampling and nonresponse error. Performing post survey adjustments may be needed, if underrepresented respondents are in the data set. For example, if the survey designer detects a mismatch between the distribution of the age of the respondents the survey designer has the possibility to counteract this fact by introducing survey weights. These survey weights may reduce sampling error but may increase adjustment error (GROVES, ET AL., 2011).

2.2 Respondent Behaviour

According to BIEMER (2010) measurement error from the total error framework has a significant impact on total survey error. Many factors like, survey questions, interviewer performance (an interviewer may be poorly prepared or misunderstood), respondent behaviour, etc. may influence measurement error. This thesis highlights the factor respondent behaviour, which results in low quality answers (BIEMER, 2010), (GUIN, ET AL., 2012). Respondent behaviour, which results in low-quality answers, can be summarized as negative respondent behaviour and will be explained in more detail in the following chapters. In addition to the described behaviour, there is also the possibility that respondents give no answer at all or answer only some questions of a survey. Such a behaviour is related to nonresponse error from the total survey framework (BIEMER, 2010).

2.2.1 Negative Respondent Behaviour

Negative respondent behaviour can be summarized as careless responding or insufficient effort responding (WIMMER, ET AL., 2018). Careless responding is a risk for data quality (FRANCAVILLA, ET AL., 2018) by providing incorrect responses (WARD & MEADE, 2018). Providing such responses can be unintended or not (BOWLING, ET AL., 2016). Some examples for negative respondent behaviour are speeding (rushing through the survey), early termination, etc. Why do respondents perform careless or insufficient effort responding and therefore show negative respondent behaviour? This question can be answered with the help of the respondent burden (BRADBURN, 1978). This burden is a combination of the length of the interview, the amount of the effort taken to answer the survey, the stress and feeling during the survey and how often the survey was performed by the same respondent (GUIN, ET AL., 2012). After the respondent started the survey, this burden will arise as it becomes clear, that this survey is dull and boring to fill out and it will take a long time to complete it. Instantly the respondent will think to terminate the survey, but then the conscientiousness appears to complete this survey. Therefore, to satisfy the burden, the respondent may use another strategy beside termination, namely using short cuts (e.g. take always the first answer option) to reduce the effort to answer the survey questions (DÉCIEUX, MERGENER, NEUFANG, & SISCHKA, 2015), (WIMMER, ET AL., 2018), (KROSNICK, 1991), (KROSNICK, NARAYAN, & SMITH, 1996). The result of using such a strategy may be a low-quality response data – because the respondent had low motivation answering the questions (WARD, ET AL., 2018), (AKBULUT, 2015), (JONES, HOUSE, & GAO, 2015). The attitude of the responder to using short cuts (like careless response patterns, etc) and therefore reduce the effort to answer survey questions can be defined as satisficing (KROSNICK, NARAYAN, & SMITH, 1996) – human behaviour – (MALHOTRA, 2008), (BARGE & GEHLBACH, 2012) and is related to low response quality (GANASSALI, 2008).

This circumstance has influence on total survey error (BIEMER, 2010) and will increase measurement and non-response error and therefore lower the overall data quality of the survey (GROVES, ET AL., 2011)

Related work has recommended to remove or ignore low-quality response, which were afflicted by negative respondent behaviour, from subsequent evaluation (WARD, ET AL.,2018), (FRANCAVILLA, ET AL., 2018). This method can result in an improved statistical power of the survey data (MANIACI & ROGGE, 2014).

2.2.2 Detectable Negative Respondent Behaviour

Negative respondent behaviour can be detected by careless response patterns. Numerous careless response patterns, which were used in SQT, are explained in this chapter.

The "*don't know*" careless response pattern will be used by the respondent for skipping questions in a fast way (DECIEUX, ET AL., 2015), (WIMMER, ET AL.,2018), (ZHANG & CONRAD, 2014), (DEUTSKENS, RUYTER, WETZELS, & OOSTERVELD, 2004). In related work the "don't know" careless response pattern is also called the skipping item pattern (BARGE, ET AL., 2012). This pattern will be detected, if the respondent selects the "don't know" answering option in a significant way instead of selecting a quality answer.

The *straightlining* or maximum long string or non-differential careless response pattern (GUIN, ET AL., 2012), (KROSNICK, 1991), (WIMMER, ET AL.,2018), (CIBELLI, 2017), (BARGE, ET AL., 2012), (WARD ,ET AL., 2018), (MANIACI, ET AL., 2014), (ZHANG, ET AL., 2014), (DESIMONE, HARMS, & DESIMONE, 2015), (MEADE & CRAIG, 2012) can be described as selecting a significant number of same answers in series per question matrix (CIBELLI, 2017), (FRANCAVILLA, ET AL., 2018), (MCGRATH, MITCHELL, & HOUGH, 2010).

The *anchoring* careless response pattern describes the effect that the respondent uses an anchor – selecting the first answer option of a question matrix – and does not vary the other answers in a significant way (BARGE, ET AL., 2012). For example, the respondent answers the current question of the matrix with the answer option 4. If the respondent is prone to the anchoring pattern, the other answers for this question matrix will vary near the anchor (selecting the answer option 3 or 5 or even 4). Because of this fact, if a respondent uses the anchoring pattern only a little effort is needed to answer the survey questions (BARGE, ET AL., 2012).

The *early termination* careless response pattern describes the effect, if a respondent terminates an online survey before it actually ends (WIMMER, ET AL.,2018). This behaviour can be an indicator for negative respondent behaviour (BARGE, ET AL., 2012), (HALBHERR, 2017).

The *speeding* careless response pattern is related to the completion time of an online survey. If the respondent completes the survey, but selects only low-quality answers, this strategy will lead to a significantly lower completion time than applying high-quality answers (ZHANG, ET AL., 2014), (WIMMER, ET AL.,2018), (BARGE, ET AL., 2012), (DESIMONE, ET AL., 2015), (MEADE, ET AL., 2012). If the completion time of one response is significantly lower than the average completion time of all survey responses, then the response with a lower completion time will be prone to speeding (HUANG, ET AL., 2012), (WARD, ET AL., 2018), (FRANCAVILLA, ET AL., 2018).

The idea behind the *semantic synonyms*, even-odd consistency or inter-item standard deviation careless response pattern is, that a respondent should give similar responses to similar questions. If this is not the case, this behaviour may be an indication for negative respondent behaviour (DESIMONE, ET AL., 2015), (MEADE, ET AL., 2012), (FRANCAVILLA, ET AL., 2018), (WARD, ET AL., 2018), (HUANG, ET AL., 2012). For example, if a question in an online survey asked the respondent “Do you like your job?” and the respondent answered “yes”, and another question arose “Are you happy with the job” and the user answered “no” then there were different responses to the same question.

The *conflicting answers* or semantic antonyms careless response pattern can be detected, if the respondent gives the same answer to dissimilar questions. For example, the question “Do you like meat?” and the question “Are you vegetarian?” will be asked during an online survey. Both questions will be answered by the respondent with “yes” resulting that dissimilar questions received the same answers. Because of this fact, a conflict arises, which is related to negative respondent behaviour (DESIMONE, ET AL., 2015), (WIMMER, ET AL.,2018).

The *diligence and interest* careless response pattern asks the respondent about his level of effort (MEADE, ET. AL., 2012), (WARD, ET AL., 2018). This pattern can be detected if a survey question like: “I carefully read the instructions of this survey!” or “I answered the questions of the survey in a qualitative way” will be asked at the beginning or at the end of an online SURVEY (FRANCAVILLA, ET AL., 2018), (WARD, ET AL., 2018), (DESIMONE, ET AL., 2015).

The *open-ended* careless response pattern detects negative respondent behaviour by counting the words of open-ended questions per response and compares this word count with the average word count of all responses. If the word count of a response for a question, is significantly lower than

the average, this can be an indicator for negative respondent behaviour (BRÜGGEN & DHOLAKIA, 2010), (MUÑOZ-LEIVA, SÁNCHEZ-FERNÁNDEZ, MONTORO-RÍOS, & IBÁÑEZ-ZAPATA, 2010).

The *Mahalanobis Distance* is another careless response pattern for detecting negative respondent behaviour (see **Formula 2**). Before this careless response pattern can be used, an average response pattern must be defined – so-called gold standard. Thereafter the current response pattern will be compared against this gold standard. The Mahalanobis Distance will be calculated as follows:

$$d_i = [(x_i - \bar{x})^t * C^1 * (x_i - \bar{x})]^{0.5}$$

Formula 2. Mahalanobis DistanceMahalanobis distance[2] (di) will be calculated by comparing current response (xi) with a gold standard ($\bar{x}$). High di values indicate negative respondent behaviour.

Extreme or higher values of the Mahalanobis Distance are an indicator for negative respondent behaviour (FRANCAVILLA, ET AL., 2018), (WARD, ET AL., 2018), (DESIMONE, ET AL., 2015).

Trap questions (MILLER & BAKER-PREWITT, 2009) or instructed items (DESIMONE, ET AL., 2015) instruct the respondent to perform an action. For example, "Before starting with the question, answer the first sub question with answering option 2". If the respondent does not comply with the given instruction this can be an indicator for negative respondent behaviour, because the respondent didn't read the instructions before answering.

Bogus items are questions, which offer only one clear and correct answer. If the respondent will give an incorrect answer to such a question this can be an indicator for negative respondent behaviour (FRANCAVILLA, ET AL., 2018), (WARD, ET AL., 2018), (HUANG, ET AL., 2012), (BEACH, 1989). To make full use of this careless response pattern bogus items should be placed on different places of the online survey (DESIMONE, ET AL., 2015), (MEADE, ET AL., 2012).

2.2.3 State – of – the Art

Negative respondent behaviour results in low-quality answers. Therefore, related work has recommended to remove or ignore low-quality response, which were afflicted by negative respondent behaviour, from subsequent evaluation (WARD, ET AL.,2018), (FRANCAVILLA, ET AL., 2018). To fulfil this task, it is crucial to detect negative respondent behaviour in an automatic and easy way. Therefore, the author of this thesis summarized the seven most used survey tools, according

[2] https://www.statisticshowto.com/mahalanobis-distance/ (07.06.2020)

to tech radar[3], according their ability to detect negative respondent behaviour and compare them against SQT (see **Table 1**):

- Survey Monkey
- Typeform
- JotForm
- AskNicley
- Formstack
- Surveygizmo
- Google Forms

Survey Monkey, JotForm, Formstack, Surveygizmo and Google Forms have the ability to detect some basic parts of negative respondent behaviour, like speeding or early termination. In contrast, SQT will detect 10 careless response patterns like speeding, straightlining, early termination, etc. In addition, the careless response patterns detected by SQT are well defined based on different scientific work, whereas the detection method of careless response patterns of the mentioned survey tools is not known to the author. Furthermore, all mentioned survey tools are not open source. This circumstance makes it relatively hard to move data from one survey tool to another. SQT in contrast is open source and therefore, any data from any survey analytic tool can be imported to SQT (JSON format of answers and questions of the online survey is needed). In addition, SQT features one common user interface for different types of devices, like smartphones, tablets and personal computers, where all functions are available. From a commercial point of view, all mentioned survey tools typically share the same disadvantage, only basic options are provided for free, but advanced analysis of domain-specific behaviour patterns (such as respondent behaviour) is poorly supported. For example, survey monkey has some features like page randomization or question randomization, but for these features the survey analyst has to pay.

Summarizing this chapter, there is a lack of validated and easy-to-use software tools, which detect negative respondent behaviour and allow survey analysts to correct or exclude low-quality responses.

[3] https://www.techradar.com/best/best-survey-tools (17.05.2020)

Feature	Survey Mon-key	Type-form	Jot-Form	Ask-Nicley	Form-stack	Survey-gizmo	Google Forms	SQT
Don't know	x	x	x	x	x	x	x	✓
Anchoring	x	x	x	x	x	x	x	✓
Straightlin-ing	x	x	x	x	x	x		✓
Early termination	✓	x	✓	x	✓	✓	✓	✓
Speeding	x	x	✓	x	x	✓	x	✓
Semantic synonyms	x	x	x	x	x	x	x	✓
Conflicting answers	x	x	x	x	x	x	x	✓
Diligence and interest	x	x	x	x	x	x	x	✓
Open-ended questions	x	x	x	x	x	x	x	x
Mahalanobis distance	x	x	x	x	x	x	x	x
Trap questions	x	x	x	x	x	x	x	✓
Bogus items	x	x	x	x	x	x	x	✓
A/B testing	✓	x	x	x	✓	✓	x	x

Table 1. Feature set of survey tools.Listing of the most used survey tools split according to their ability to detect negative respondent behaviour.

2.3 Innovation in Survey Designs

Besides the described method from **Chapter 2.2** (detecting negative respondent behaviour) another technique, to decrease measurement error and therefore total survey error in online surveys, exists. This technique is called gamification and can be described as improving the appearance of the online surveys by combining game elements with a non-game context (SCHACHT, ET AL., 2017). Gamification is a relative new method for combining online surveys – non-game context – with game elements like score boards or mini games – game elements – to improve the appearance of the survey and to motivate the respondents to complete the survey. This has the consequence that total survey error will decrease (DETERDING, DIXON, KHALED, & NACKE, 2011). Applying gamification to online surveys has not always a positive effect (HAMARI, ET AL., 2014), (KEUSCH, ET AL., 2017). The application of the method of gamification in a wrong way, can in the worst case, increase total survey error in contrast to the conventional – non-gamified – online survey. To minimise this problem and to help the survey designer to create a gamified survey Johannes Harms had developed a unified process to convert a conventional – non-gamified – survey in a gamified ONE (HARMS, WIMMER, KAPPEL, & GRECHENIG, 2014).

2.3.1 Gamified Survey Designs

Gamification is a relatively new topic and therefore many definitions about gamification exist. One general definition for gamification is provided by DETERDING, ET AL., (2011): "*a process (so called gamification process) combines game elements with a non-game context to motivate users to perform their tasks more efficiently*".

Another definition about gamification is given by (HUOTARI & HAMARI, 2012): "*Gamification refers to: a process of enhancing a service with affordances for gameful experiences in order to support user's overall value*."

Basically, both definitions reflect on the fact, that gamification in online surveys is done by using game design elements in a non-game context (SCHACHT, ET AL., 2017). This method is defined as gamification and can be found nowadays in many online surveys. Gamification can range from simple achievements – e.g. badges (HARMS, SEITZ, WIMMER, KAPPEL, & GRECHENIG, 2015) – to fully developed games in an online survey context (CECHANOWICZ, ET.AL., 2013). The following chapters will give the reader more information about gamification and the process of creating a gamified online survey.

2.3.2 Iterative Process for Gamified Online Surveys

Gamification is a new topic in the area of human interaction and therefore there exist no standardized process how to convert a conventional online survey into a gamified one. To counteract this problem Johannes Harms (HARMS, ET AL., 2014) defined a unified process for this problem (see **Figure 5**). This process is based on the MDA framework (mechanics-dynamics-aesthetics) (HUNICKE, LEBLANC, & ZUBEK, 2004). The different steps of the process will be explained in detail below (BIEGLER, 2015).

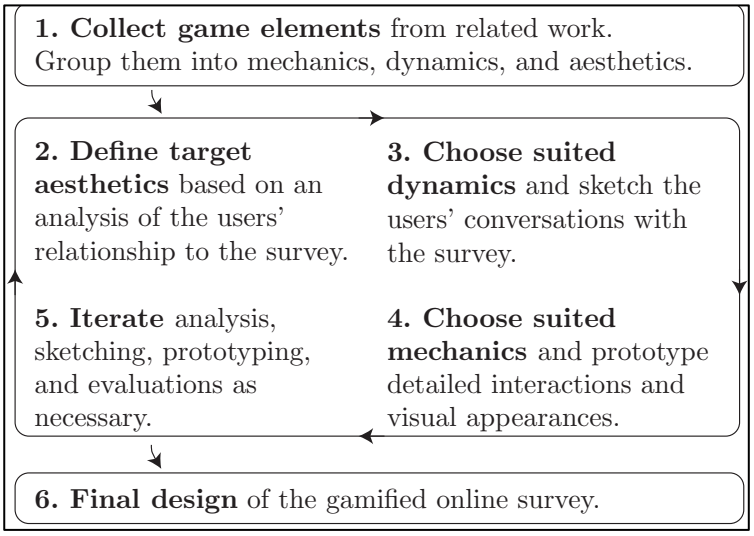


Figure 5. Gamification design process. (HARMS, ET AL., 2014). Iterative design process to convert a conventional into a gamified online survey.

Step 1: Collect Game Elements

In the first step of the unified process the main goal of the design team is to collect different game elements, which may positively influence the further design of the gamified survey (HARMS, ET AL., 2014). One drawback of the unified design process is, that step one "Collecting Game Elements" will be performed only once. All other steps of the process will be executed in an iterative way. Therefore, wrong decisions made by the design team in step one will be hard to fix in later steps. Due to this fact, the design team should plan enough time and resources to perform the first step of the process in an exact way to avoid problems later on. To help the design team selecting correct game elements and to weaken the disadvantage of the unified process some useful game elements can be found in the scientific work of PULESTON (2011) and HARMS, ET AL., (2014). For understanding all of the available game elements some background knowledge about the MDA framework (HUNICKE, ET AL., 2004) is needed, because the steps of the unified process rely on the MDA framework. Therefore, it will be briefly explained.

The MDA framework consists of three different components – game mechanics, dynamics and aesthetics (see **Figure 6**).

Mechanics are used to describe the basic building blocks (data representations, algorithms, rules, interactive elements, etc.) for creating a game.

Dynamics – as the name already implies – refer to the resulting run-time behaviour and will be influenced by the mechanics.
The last part is the *aesthetics* component, which will be triggered by dynamics and characterize a player's emotional response and experience.

Figure 6. MDA design counterparts.(HUNICKE, ET.AL., 2004). Mechanics are the basic components for creating a game (e.g. rules, algorithms, etc.) and will trigger the dynamics. Dynamics describe the run time behaviour and will trigger the aesthetics. Aesthetics characterize the player's emotional response.

Step 2: Define Target Aesthetics

In the second step of the unified process the target aesthetics for the online survey will be defined. Before the design team can perform this task, the intended target population and the context (e.g. sport) of the online survey must be analysed and defined. This is crucial, because the positive effect of aesthetics depends on the target population (HARMS, ET AL., 2014) and the context. After analysing the target population and the survey context the design team can select the aesthetics from PULESTON (2011) and HARMS, ET AL., (2014). To give the reader an insight about the characteristics of different aesthetics an example will be explained below:

In the most cases the aesthetics "challenge" and "sensation" will be selected by the design team of an online survey. This two aesthetics have the advantage that they almost fit for every online survey independently of the context. Fulfilling the aesthetic "sensation" is rather easy. An example would be, to use visual or audio effects for different questions. The other aesthetic "challenge" is a little bit harder to fulfil. One example for triggering the aesthetic "challenge" could be the introduction of mini games. The advantage of using this technique for answering different question is, the respondent has a challenge to answer different questions. The disadvantage is, if the selected mini games are too hard to solve, they can have a negative effect on respondent behaviour (KAMINSKA, MCCUTCHEON, & BILLIET, 2010). To counteract this possible negative effect the design team has to make a trade-off between satisfying the "challenge" aesthetic and don't frustrate the respondent.

Step 3: Choose Suited Dynamics

After the design team has selected the right aesthetics for the given online survey the next step is to choose dynamics, which trigger the selected aesthetics. Finding the right dynamics can be done either by using the MDA framework (HUNICKE ET. AL., 2004) or other scientific work. One problem of using the MDA framework is, the framework offers a lot of different dynamics, but not all of them fit in the context of an online survey. This can be the case, because they are too complex for the context of an online survey (HARMS ET. AL., 2014). One example for a dynamic, which fits the context of an online survey is "time pressure". This dynamic can easily trigger the aesthetic "challenge", because if the respondent has a time limit for answering some question this situation is new to him and therefore challenges him. Another useful application of the dynamic „time pressure" is given in the context of open question. Applying "time pressure" to this type of questions can increase the quality of the given answer (PULESTON, 2011). One drawback of this approach is, if the dynamic „time pressure" is used for many questions of the survey. This can lead to demotivated respondents and can result in non-response.

Step 4: Choose Suited Mechanics

In this step the design team selects the game mechanic, which will trigger the dynamics and aesthetics from the previous steps. Game mechanics are detailed building blocks and rules, which define a game (HUNICKE ET. AL. 2004) and can be found in the MDA framework. One example for a game mechanic is "points and badges". This game mechanic triggers the dynamic "feedback", which can produce the aesthetic "challenge". Another example for using a game mechanic is to visualize a stopwatch next to the open questions. This stopwatch triggers the dynamic "time pressure" and the aesthetic "challenge". Using an avatar is also a valid game mechanic which enables the respondent to move their avatar through all levels. Such a game mechanic can trigger the aesthetic "exploration".

Step 5: Iterate

The main goal of this step is to iterate through the previous defined steps of the design process – created by Johannes Harms (HARMS, ET AL., 2014) – to correct the aesthetics, dynamics and mechanics selected for the survey in such a way, that at the end a high-quality gamified online survey will be the result.

2.3.3 Real-World Example

The application of the unified design process from HARMS, ET AL., (2014) was performed by the author of this thesis in another scientific work (HARMS, ET AL., 2015), (BIEGLER, 2015) resulted in a gamified online survey. At the beginning, the author of this thesis selected a conventional online survey – contained different standard interaction possibilities like radio buttons, open text

questions, etc. (see **Figure 7**). This conventional online survey was about sport and health among teenagers. Therefore, the first idea of the author of this thesis was to build the gamified survey on top of a sports event. After defining the setting of the gamified online survey, the author applied the unified design process. The result of the gamified online survey can be seen in **Figure 8.**

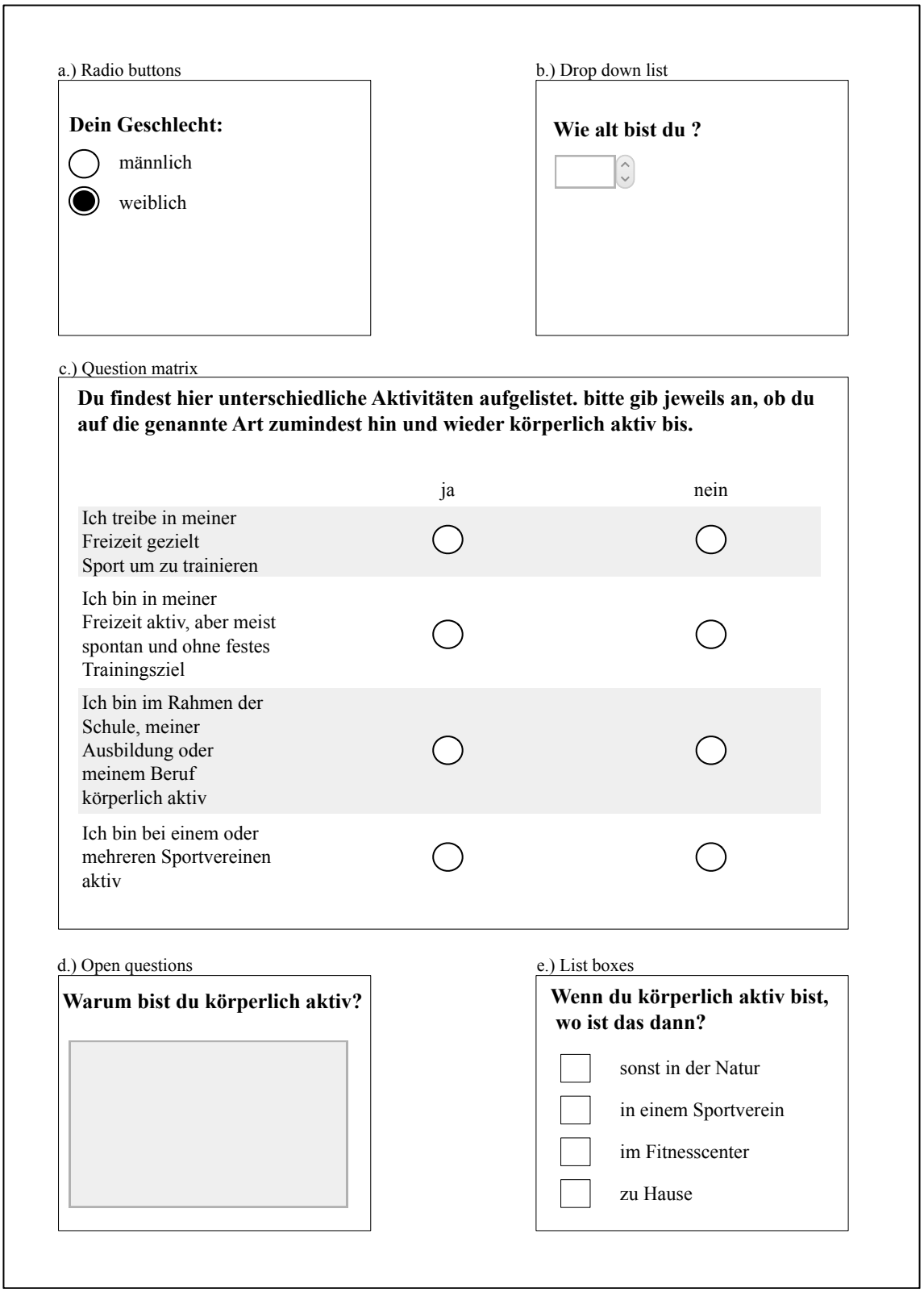


Figure 7. Conventional survey design.The survey contained different interaction possibilities (a-e).

At the first glance following improvements between the conventional survey (see **Figure 7**) and the gamified version (see **Figure 8**) after applying of the unified design process can be seen:

- *Creating a customized avatar* (see **Figure 8a**): At the beginning of the survey the respondent had to answer some demographic questions. To make this common step more enjoyable the author of this thesis introduced an avatar for the respondent. The avatar changed the appearance depending on the demographic answers given by the respondent. Later on, the respondent used the avatar to navigate through a map (see **Figure 8b**) and different mini games (see **Figure 8c-f**) to answer the survey questions. This feature challenged the aesthetic "exploration".
- *Introduction of different mini games* (see **Figure 8c-f**): All questions of the conventional online survey were grouped and split into four different mini games. Each mini game was developed in such a way, that the respondent needed to interact with the avatar in a different way. This circumstance triggers the aesthetic "challenge". To overcome the possible issue that a respondent would be overstrained with the controls of the different mini games a demo question at the start of each mini game was available. During the demo question the respondent could train the controls of the mini game as long as needed.
- *Coins and shop* (see **Figure 8g**): After the completion of a mini game (respondent answered all questions of a question matrix) the respondent will earn as a reward coins. These coins will motivate the respondent to answer the other questions of the remaining mini games. Furthermore, the coins can be spent in an avatar shop (see **Figure 8g**) to customise the avatar.
- *Feedback:* After solving the different mini games the avatar will return to the map (see **Figure 8b**). The solved mini games will be deactivated and marked.
- *Story telling* (see **Figure 8h**): Starting with the avatar creation and furthermore with the different mini games, in the context of different sports disciplines, the gamified survey will tell a story. This story is connected to the context of the survey (sports and health). Therefore, if the respondent answered all questions – solved all mini games – the online survey is completed, and the avatar will be part of a medal ceremony.

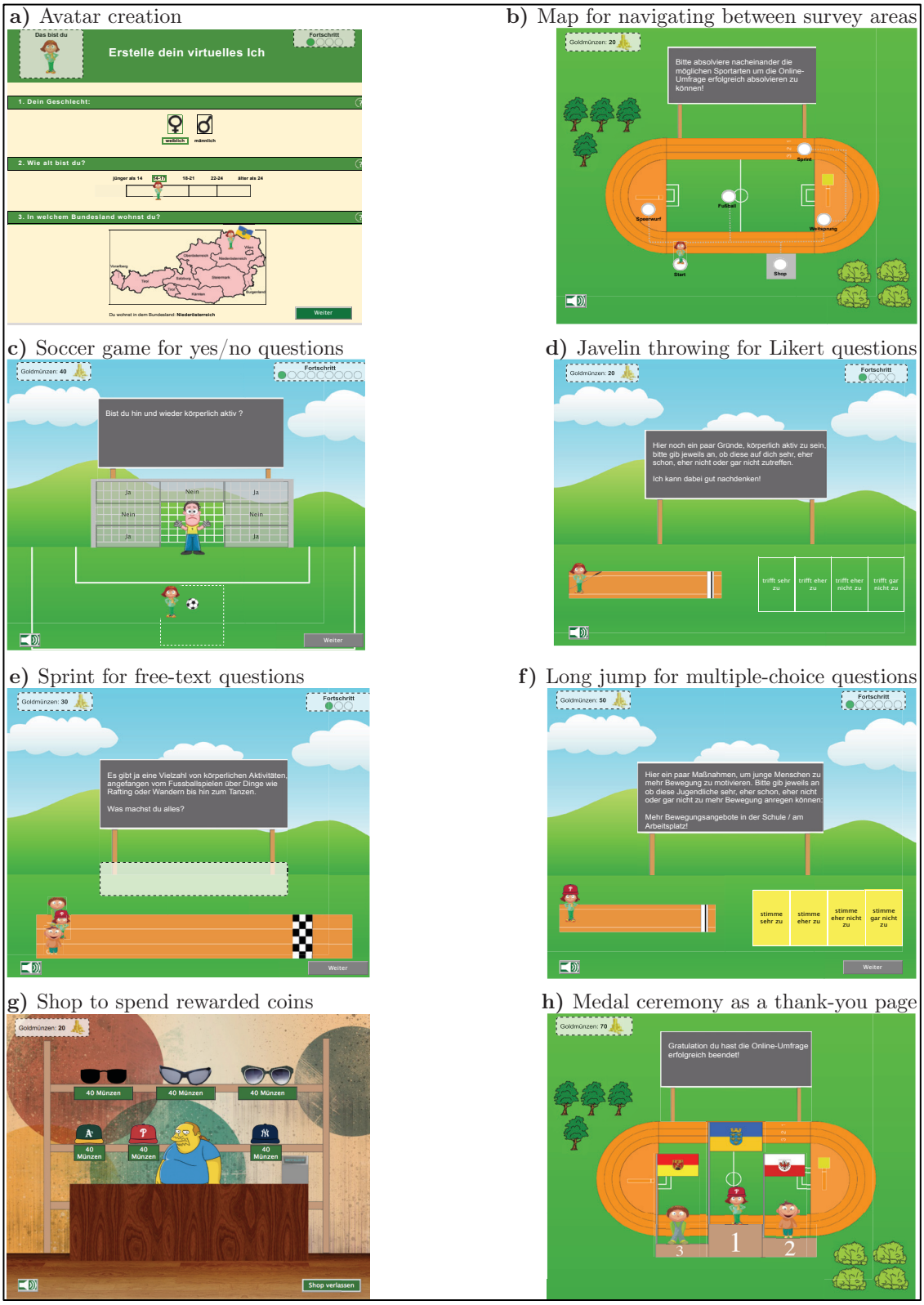


Figure 8. Gamified online survey. The appearance of the avatar (a) changes with the answers given for the demographic questions. To challenge the aesthetic "exploration" a map (b) was constructed to navigate to the different question pools. Different mini games (c-f) were introduced to cover the aesthetic "challenge". After the respondent completed a mini game the respondent will receive coins,

which can be spent for accessories for the avatar (g). This should motivate the respondent to complete all mini games. If the respondent answered all questions a medal ceremony (h) will end the online survey (HARMS, ET AL., 2015).

Summarizing the features of the real-world example, the main task of gamification is to motivate respondents to perform tasks they dislike instead of aborting the survey and to detain using careless response patterns (therefore decrease total survey error) (CECHANOWICZ, ET AL., 2013). To achieve this task the practical example uses design elements characteristic for games (DETERDING ET AL., 2011) – mini games, avatar, coins, etc. – and combines them with a non-game context – answering questions about the topic sport and health – to give the respondent the feeling of playing a game instead of answering a boring online survey.

2.3.4 Evaluating the Effects of Innovative Survey Designs

Despite the positive effects using gamification in online surveys – attracting more respondents (CECHANOWICZ, ET. AL., 2013), (DOLNICAR, ET AL., 2013), (GUIN, ET AL., 2012), (PULESTON, 2011), increases the user experience (DOLNICAR, ET. AL., 2013), (GUIN, ET AL., 2012), higher motivation and engagement of the respondents (CECHANOWICZ, ET AL., 2013), (DOLNICAR, ET AL., 2013) and a better data quality (DOLNICAR, ET AL., 2013) – some scientific papers highlight drawbacks using gamification (HAMARI, ET AL., 2014).

One drawback of gamification is, that the mentioned positive effects depend on the context in which gamification is used and may not be long-term (HAMARI, ET AL., 2014). Furthermore, gamification depends on several external factors like the motivation of the respondent (HAMARI, ET AL., 2014), which is not in the sphere of the survey creator. Another problem of gamifcation can be an increased competition between the respondents – e.g. major goal is to be the first on the leader board – disregarding the quality of given answers (HAMARI, ET AL., 2014) by answering questions without thinking (KEUSCH, ET AL., 2017). Besides HAMARI, ET. AL., (2014) another scientific work published from KEUSCH, ET. AL., (2017) highlights, that gamification does not increase the completion rate and had only small positive impact on the enjoyment of the online survey. The possible negative effects of gamification can introduce new errors or increase pre-existing errors in the online survey and therefore can negatively influence total survey error. Due to this fact, total survey error of a gamified survey may – in some cases –be greater than total survey error of a conventional survey (HARMS, ET AL., 2014). Johannes Harms (HARMS, ET AL., 2014) had therefore updated the total error framework to highlight, which statistical errors may be increased by gamification (see **Figure 9**). The first statistical error, which may be influenced by gamification is validity. Gamification can also influence measurement error because of negative respondent behaviour. This can be the case, if the respondent uses speeding, straightlining,

etc. because of low motivation using a gamified online survey. The last statistical error, which can be increased by gamification is nonresponse error. Higher nonresponse error can result, because some respondents could be "scared" from the appearance of a gamified online survey. All other statistical error from the total survey framework are the same for a gamified as for a conventional online survey (HARMS, ET AL., 2014).

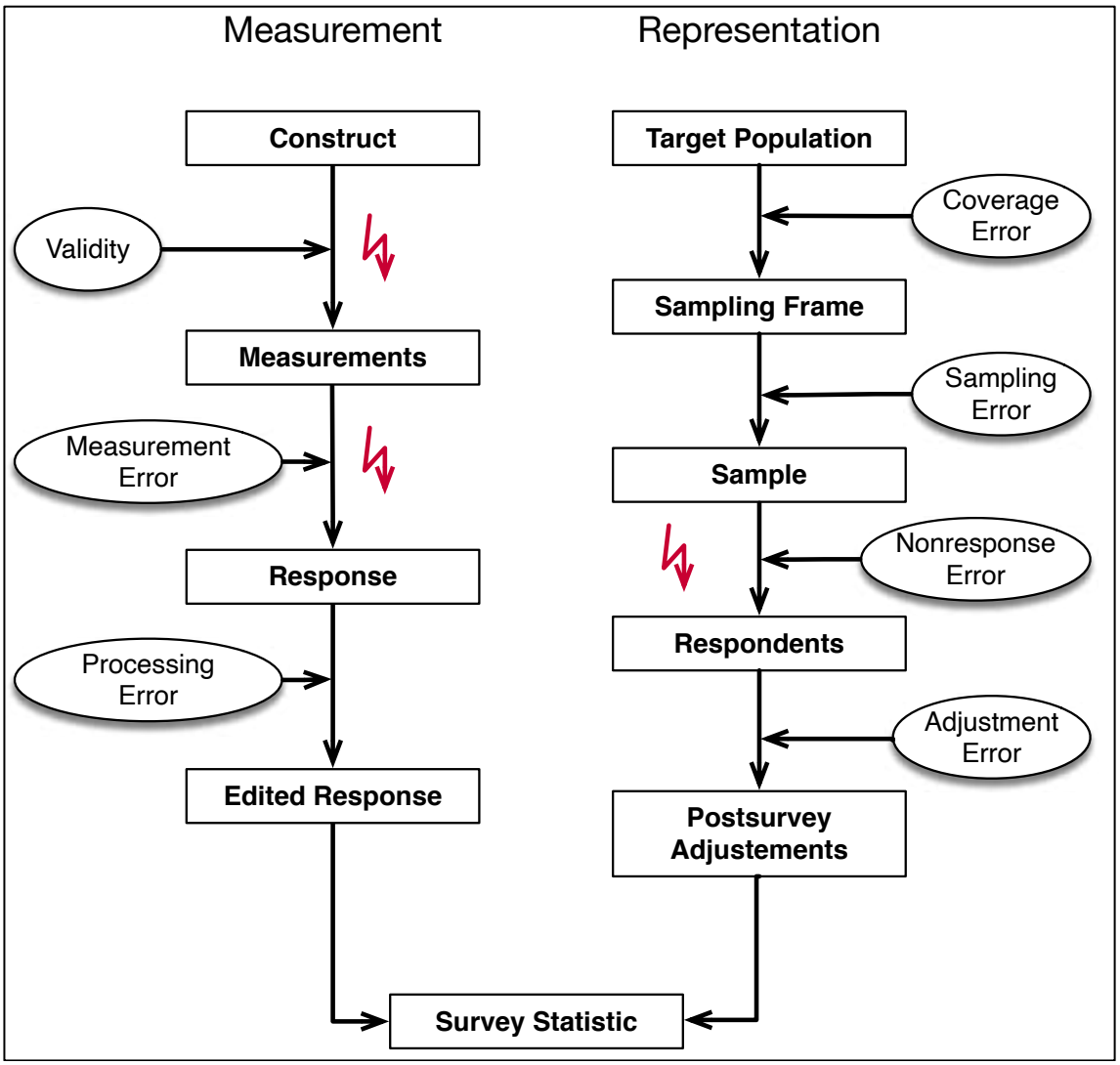


Figure 9. Possible negative effects of gamification on the total error framework. (HARMS, ET AL., 2014). Red arrows mark statistical error, which may be increased or will be newly introduced when using a gamified online survey.

3 Survey Quality Tool (SQT)

SQT was developed as a web application in an iterative design and engineering process. First of all, a literature review was performed to detect the different careless response patterns including their calculation method. In the next step, the different requirements for a tool like SQT were defined. The requirements for SQT were defined during meetings with Johannes Harms (senior designer) and the author of this thesis. In the next step a GUI mockup prototype was created (see **Appendix II**). Based on the literature review and requirements SQT was developed as a web-based tool for automatically detecting negative respondent behaviour in online surveys.

3.1 Literature Review

The main research goal of this thesis is to develop and implement a survey quality tool (SQT), which is able to automatically detect and measure negative respondent behaviour. Based on this research goal it is necessary to find scientific work related to negative respondent behaviour. Scientific work should contain a description of negative respondent behaviour and how negative respondent behaviour can be measured and detected. Performing a literature review for this topic is crucial, because, according to the experience of the author with this topic, many definitions and techniques, how to measure negative respondent behaviour, are ambiguously defined. Selecting a scientific work with an ambiguously defined negative respondent behaviour can lead to significant problems later on.

3.1.1 Methods

Methodically, the following systematic literature review is based on the PRISMA framework (LIBERATI, ET AL., 2009). This framework has its origin in the area of healthcare, where systematic literature reviews are essential and common. The main advantage of using this framework is, it provides a well-proven and detailed guideline for performing a literature review. The remainder of this chapter describes the procedure, starting from a defined research question, applying different rules of the systematic literature review and, as a result, getting high-quality scientific work.

3.1.2 Scoping

At the first sight the scoping of the systematic literature review is the already defined research goal to develop and implement a survey quality tool (SQT), which is able to automatically detect

and measure negative respondent behaviour. This research goal is a general problem description and in the opinion of the author, not suitable for finding scientific work about negative respondent behaviour. To overcome this issue, the author of this thesis decided to introduce two questions resulting in a more precise scoping for the literature review:

- What is negative respondent behaviour in online surveys?
- How can negative respondent behaviour be measured and detected in online surveys?

3.1.3 Planning

In this part of the systematic literature review, the execution of the scientific literature search will be explained.

The planning step started with the definition of search terms based on the scoping of the literature review. A search term can be simple – just one phrase – or more complex – multiple terms with different alternatives. Different alternatives could be synonyms, singular/plural forms, different spellings etc. (SIDDAWAY, 2014). At the beginning, the author of this thesis defined initial search terms, which covered the scoping of the literature review (see **Table 2**). Thereafter, search queries (with initial search terms) were applied to a literature database. After some queries were applied, the author of this thesis recognized, that the initial search terms didn't cover the whole search space. Therefore, he decided to make the search term more complex. The complexity was derived by generating different synonyms for the initial search terms with the webpage SYNONYM[4]. The result was a list, with different synonyms for each initial search term. The drawback of this approach was, that not all of the resulting synonyms were useful, because some of them had a different meaning – based on the context – and therefore did not fit the context of the scoping of the literature review. To overcome this drawback, the author of this thesis used two countermeasures.

The first countermeasure was to check the meaning of the synonyms with the online dictionary LEO[5].

The second countermeasure was to test the complex search terms in ad-hoc queries in a literature database.

The resulting alternatives – after eliminating search terms that did not fit into the context – are described in the table below:

[4] https://www.synonym.com/synonyms/(19.10.2019)

[5] https://www.leo.org/englisch-deutsch (19.10.2019)

Initial Search Term	Alternative Search Terms
Negative	Destructive, random, careless, quality.
Response	Result, feedback, answer, reply, respondent, responding.
Behaviour	Action, engagement, effort.
Survey	Only the word survey was used because no other synonym fits this word!
Measure	Scale, grade, evaluate, score.
Detect	Trace, discover, observe, find.

Table 2. Search terms for the scoping of the literature review.Search terms were used to find scientific work, which is related to the scoping of the literature review.

After the definition of the search terms, the next planning step started with the selection of the literature databases. For this work, Google Scholar[6] was chosen as single database for the literature search. Therefore, the above defined search terms were applied to queries, searching the literature database for suitable scientific work. Performing this task, resulted in many scientific work but not all of them were relevant for the scoping of the literature review. To dismiss scientific work, which were not relevant, the author of this work defined filters – exclusion and inclusion criteria. The most common exclusion and inclusion criteria were picked from (LIBERATI, ET AL., 2009) and some additional were defined by the author:

Type of publication: Scientific publications (papers, graduate thesis, etc.) from sources with a background from the related topic (ACM, IEEE, university, etc.) will be included. Excluded from the results will be manuals, popular scientific journals (e.g. GEO, P.M. Magazine, National Geographic, etc.), thesis and webpages.
Date of issue: The scientific publication should be issued in the year 1975 or newer.
Research design: There are no restrictions for the result.

When applying the defined exclusion and inclusion criteria on the query result this method may introduce a bias. One example for introducing a bias can be, if a scientific work may be deselected from the query result set because of a single exclusion criterion. It may be the case, that this criterion is no strictly relevant in the present context – and the decision would be better to include the scientific work in the result set. Furthermore, the judgement if an exclusion/inclusion criterion should be applied or not may be biased or tainted. To keep such a bias as small as possible, the

[6] https://scholar.google.at (24.01.2019)

author of this work used two countermeasures for applying and defining exclusion/inclusion criteria:

- A pre-defined inclusion, exclusion list from (LIBERATI, ET AL., 2009) will be used. This will avoid a bias through post-hoc modifications of inclusion and exclusion criteria.
- The author of this thesis has experience in the current topic, and this experience will be beneficial for his judgements for applying the exclusion/inclusion criteria.

In addition to the query database search, a second search strategy was performed during the literature review. This second strategy had the goal to investigate the reference chapter of the already included scientific work, finding additional scientific work. The additional method worked as follows, if a scientific work, which was found by the query result, had in the reference chapter a title, which was related to negative respondent behaviour, it was treated like a search result from a query. In the next step the inclusion and exclusion criteria were applied to decide whether to include the scientific work in the systematic literature review or not.

3.1.4 Performing the Review

The last part of the systematic literature review was the actual execution of the review. As defined in the PRISMA framework (LIBERATI, ET AL., 2009), this contains four phases:

Identification: Using the defined search terms from the planning phase to search at least one electronic database for relevant literature. Thereafter inspect the search results.
Screening: Export the relevant literature to a citation manager to collate the search results.
Eligibility: Specify study characteristics (e.g. PICOS, length of follow-up, etc.) and report characteristics (e.g. years considered, language, publication status, citation rate etc.)
Included: Scientific work selection according to the inclusion and exclusion criterions.

3.1.5 Resulting Scientific Work

Applying the search terms to the literature database resulted in a total of 70 scientific work (see **Appendix I**). In the next step, duplicates were removed from the 70 scientific work resulted in 66 different scientific work. After applying the defined inclusion and exclusion criteria and the full-text assessment a total of 31 scientific work were selected for answering the scoping of the literature review (see **Figure 10**). In addition, **Table 3** gives the reader of this thesis an overview about the resulted scientific work, including a comment about the information they provide to answer the two questions of the scoping part of the literature review:

- What is negative respondent behaviour in online surveys?
- How can negative respondent behaviour be measured and detected in online surveys?

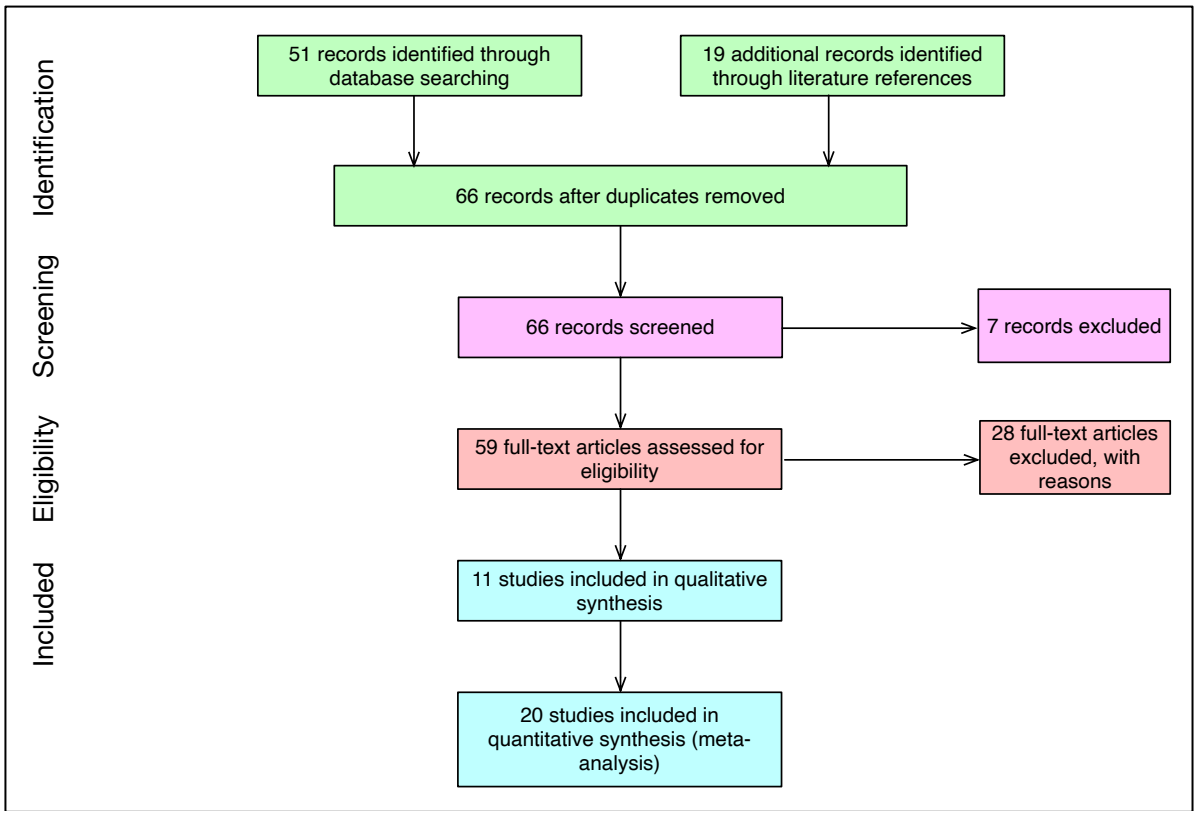


Figure 10. Resulting scientific work from the literature review.With the help of the four phases of the PRISMA framework (LIBERATI, ET AL., 2009) the total 70 scientific work found by the literature review were reduced to 31.

Scientific work	Information about negative respondent behaviour
Response rate and response quality of internet-based surveys: An experimental study. (Deutskens, Ruyter, Wetzels, & Oosterveld, 2004)	Definition of the careless response pattern "don't know".
Detecting and deterring insufficient effort responding to surveys. (Huang, Curran, Keeney, Poposki, & DeShon, 2012)	Describes the variety of careless response patterns (speeding, open questions, nonresponse, straightlining) including the term "insufficient effort responding".
Myths and realities of respondent engagement in online surveys. (Guin, Baker, Mechling, & Ruyle, 2012)	Definition of the term respondent burden and the definition of the term nonresponse.
Improving the response rate and quality in Web-based surveys through the personalization and frequency of reminder mailings.	General definition of the term response quality and the definition for detecting careless response in open-ended questions.

(Muñoz-Leiva, Sánchez-Fernández, Montoro-Ríos, & Ibáñez-Zapata, 2010)	
Determinants of participation and response effort in web panel surveys. (Brüggen & Dholakia, 2010)	Describes a measurement technique for open-ended questions, titled as survey response effort.
Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys. (Jones, House, & Gao, 2015)	Definition of different forms of careless response patterns (speeding, straightlining and trap questions).
Implementation of the forced answering option within online surveys: Do higher item response rates come at the expense of participation and answer quality? (Décieux, Mergener, Neufang, & Sischka, 2015)	Definition of the term random answering.
SQT: A tool for the automated measurement of respondent behaviour and response quality in health-related gamified online surveys. (Wimmer, Biegler, Harms, Kappel, & Grechenig, 2018)	Definition and calculation methods of different careless response patterns (premature termination, speeding, straightlining, "don't know" answers conflicting answers).
Data Quality in Cross-National Surveys. A Longitudinal and Cross-Cultural Analysis of the Quality Indicators Response Rate, Fieldwork Efforts, and Nonresponse Bias. (Halbherr, 2017)	Detailed description about nonresponse.
The Effects of Respondent Commitment and Feedback on Response Quality in Online Surveys. (Cibelli, 2017)	Details about the straightlining careless response pattern.
Personalized Feedback in Web Surveys: Does It Affect Respondents' Motivation and Data Quality? (Kühne & Kroh, 2018)	Definition of the term survey satisfaction and measurement error.
Social Interaction and Internet-Based Surveys: Examining the Effects of Virtual and In-Person Proctors on Careless Response. (Francavilla, Meade, & Young, 2018)	Definition of the term careless responding and random responses. Definition of different careless response patterns and their calculation method (Bogus Items, Instructed Response Items, Mahalanobis Distance,

	Even-Odd Consistency, Maximum LongString, Response Time, Diligence and Interest)
Identifying the random responder. (Beach, 1989)	Information about random responses and satisficing.
Respondent burden. (Bradburn, 1978)	Additional information about the respondent burden.
Completion time and response order effects in web surveys. (Malhotra, 2008)	Definition of the term satisficing.
Satisficing in surveys: Initial evidence. (Krosnick, Narayan, & Smith, 1996)	Additional information about satisficing.
The short-term campaign panel of the German longitudinal election study 2009: Design, implementation, data preparation, and archiving; version 5.0. 0. (Steinbrecher, Roßmann, & Bergmann, 2013)	Calculation method of the careless response pattern speeding.
Insufficient effort responding: Examining an insidious confound in survey data. (Huang, Liu, & Bowling, 2015)	Definition of the term careless responses and respondent burden.
Speeding in web surveys: The tendency to answer very fast and its association with straightlining. (Zhang & Conrad, 2014)	Definition of speeding and straightlining.
Insufficient effort responding as a reflection of respondent personality. (Bowling, et al., 2016)	Groundwork about careless responding
Applying social psychology to prevent careless responding during online surveys. (Ward & Meade, 2018)	Definition of different markers for careless responses.
Identifying careless responses in survey data. (Meade & Craig, 2012)	Definition of the term careless responses.
Predictors of inconsistent responding in web surveys. (Akbulut, 2015)	Definition of the term satisficing.
Using the theory of satisficing to evaluate the quality of survey data. (Barge & Gehlbach, 2012)	More information about satisficing and some examples of negative respondent behaviour. Describes the effect of low-quality responses on result findings.

Best practice recommendations for data screening. (DeSimone, Harms, & DeSimone, 2015)	Data screening methods to detect the straightlining careless response pattern.
Anchoring and Adjusting in Questionnaire Responses. (Gehlbach & Barge, 2012)	Definition of the careless response pattern anchoring and the calculation method.
Response strategies for coping with cognitive demands of attitude measures in surveys. (Krosnick, 1991)	Background information about satisficing.
Caring about carelessness: Participant inattention and its effects on research. (Maniaci & Rogge, 2014)	Definition of the term data quality and careless responses.
Evidence for response bias as source of error variance in applied assessment. (McGrath, Mitchell, & Hough, 2010)	Definition of the different types of responder and why there exists a respond bias.
Beyond 'trapping' the undesirable panelist: The use of red herrings to reduce satisficing. (Miller & Baker-Prewitt, 2009)	Definition of the term satisficing.
The Influence of the Design of Web Survey Questionnaires on the Quality of Responses. (Ganassali, 2008)	Different definitions about negative response behaviour.

Table 3. Scientific work included in the literature review.

The scientific work resulted from the literature review will be highlighted in terms of usefulness to answer the two questions of the scoping part about negative respondent behaviour.

3.2 Requirements

After the literature review, the author of this thesis defined the needed requirements for a tool like SQT. The needed requirements were derived from various meetings between Johannes Harms (senior designer) and the author of this thesis. The resulting requirements can be seen in **Table 4**.

NR	Requirements
1	Definition of clear requirements for SQT.
2	The reliability of SQT should be more than 99%.
3	The output of SQT (detection of negative respondent behaviour) should be valid.
4	The GUI of SQT should be simple to use.
5	All sensitive data should be stored in a secure way.
6	SQT should detect all available careless response patterns.
7	There should be an overview about an index value which shows whether a negative respondent behaviour is likely for this current response data or not.
8	Smooth execution for the survey analysts.
9	The design of the web application must be functional and simple.
10	A user manual for the important functions of SQT should be available.
11	The respondent data of any survey platform can be imported to SQT.
12	The imported respondent data can be easily exported to enable further statistical evaluation.
13	Respondent data should be filtered according specific attributes.
14	Before importing the respondent data different import options should be available.
15	At least an English version shall be available.
16	The import of the respondent data of an online survey and the calculation of the negative respondent behaviour should be fast.
17	The installation process of SQT should be easy and fast.
18	Detected negative respondent data can be removed from the result set.
19	SQT should be open source.
20	A user administration should be available.
21	SQT should be developed as a web application.
22	There should be no developing time limit.
23	Any available developer tool can be used to develop SQT.

Table 4. Final requirements of SQT.

3.3 Behavioural Measures

In this chapter the different careless response patterns, which resulted from the literature review and later on will be implemented in SQT – will be defined (see **Table 5** – **Table 16**). Each careless response pattern was described with additional information like, how SQT will detect and measure this careless response pattern, one real-live example and whether the careless response pattern will be included in SQT or not. The measurement of the careless response patterns was split in three stages.

First of all, a *careless response pattern value* (abbreviated with v) will be defined. This value calculates the careless response pattern for the actual answer of the current complete response. For example, the response of a survey question was, A1, A1, A1, A2, A3. The careless response pattern value (v) for the careless response pattern straightlining has the result three, because the answering option A1 was selected three times in a row. SQT will further calculate all careless response patterns based on their definitions in the same way.

In the second stage an *indicator variable* (abbreviated with ID) will be introduced. This indicator has exactly the value 0 (careless response pattern is not detected for the current question) or exactly 1 (careless response pattern is detected for the current question). In the previous example, the indicator value for the question with the response A1, A1, A1, A2, A3 and the careless response pattern value for straightlining = 3 will be 1. This is the case, because the careless response pattern value for straightlining was greater than 1 (automatic threshold for detecting this pattern).

The last stage defines the *index* (abbreviated with I), which will show an overall summary of the individual careless response pattern for the response. The value will be between 0 (careless response pattern is not detected for the response) and 1 (careless response pattern is detected for the response). For example, if a complete survey response contains two questions. Question one has an indicator variable ID of 1 (careless response pattern was detected) and for the second question the indicator variable ID is 0 (careless response pattern was not detected) the index of this response will be 0.50. This means half of the questions of the response show the given careless response pattern.

3.3.1 Anchoring

Included in SQT	YES
Description	This careless response pattern is defined as selecting the first answering option – of a question matrix – as an anchor. Thereafter, the difference of the other answering options relative to this anchor (e.g. if the anchor is set to answer option 4 and the current answer is 3, the difference will be 1) will be calculated. Each difference per question will be summed up to get an anchor value per question matrix (v_A). For the indicator (ID_A) all anchor values (v_A) will be summed up and compared to a predefined value ($\overline{v_A} = 1$). All respondents who have an anchor indicator below the predefined value ($\overline{v_A}$) of the survey are prone to anchoring.
Example	**Du findest hier unterschiedliche Aktivitäten aufgelistet. bitte gib jeweils an, ob du auf die genannte Art zumindest hin und wieder körperlich aktiv bis.** stimmt / stimmt eher / stimmt eher nicht / stimmt gar nicht Ich treibe in meiner Freizeit gezielt Sport um zu trainieren — Anchor Ich bin in meiner Freizeit aktiv, aber meist spontan und ohne festes Trainingsziel — Delta Ich bin im Rahmen der Schule, meiner Ausbildung oder meinem Beruf körperlich aktiv Ich bin bei einem oder mehreren Sportvereinen aktiv — Delta
Calculation	$Anchor\ value: v_A = \sum_{n=1}^{per\ question} \frac{\lvert(x_{anchor} - x_{selected})\rvert}{answers}$ $Anchor\ indicator: ID_A \left(\sum_{i=1}^{anchor\ values} v_{A\_i} \right) = \begin{cases} 0, & v_A > (\overline{v_A} = 1) \\ 1, & v_A \leq (\overline{v_A} = 1) \end{cases}$ $Anchor\ index: I_{A=} \frac{\sum_{n=0}^{questions} ID_A}{questions}$
Literature reference	• Barge, et al., 2012 • Gehlbach & Barge, 2012
Remark	

Table 5. Careless response pattern "Anchoring". Evaluates the difference between the anchor (first answer selection of the respondent) to the following answers. A small difference will indicate negative response behaviour.

3.3.2 Bogus Items

Included in SQT	YES
Description	Bogus items have exactly one correct answer and if the respondent answered wrong, this can be an indicator for negative respondent behaviour. Therefore, the bogus value (v_B) will save the answer of the respondent of a bogus question. The indicator (ID_B) compares the given answer to the correct answer. If the answer was right, the indicator will be assigned to 0 (careless response pattern not detected) otherwise to 1 (careless response pattern detected).
Example	If the respondent answers the question "1 + 1 = ?" with an answer other than 2 this is a violation to the bogus items careless response pattern.
Calculation	$$Bogus\ value: v_B = x_{answer}$$ $$Bogus\ indicator: ID_B(v_B) = \begin{cases} 0, & v_B = v_{correct_{answer}} \\ 1, & v_B \neq v_{correct_{answer}} \end{cases}$$ $$Bogus\ \ index: I_B = \frac{\sum_{n=0}^{questions} ID_B}{questions}$$
Literature reference	• DeSimone, et al., 2015 • Meade, et al., 2012 • Francavilla, et al., 2018
Remarks	

Table 6. Careless response pattern "Bogus Items". Some questions of a survey had a predefined right answer. If the respondent gave a false answer, this will indicate a negative respondent behaviour.

3.3.3 Conflicting Answers

Included in SQT	YES
Description	In this careless response pattern, the respondent will give the same answer to dissimilar questions. Therefore, the conflicting value (v_C) will be calculated by saving the first answer and calculate the delta value with the help of the current answer. If the first answer is the same as the current answer, then the delta value is 0 and the conflicting indicator (ID_C) will show the value 1 (careless response pattern is detected because there should be different answers for different questions). If the first answer is different to the current answer, then the delta value is unequal to 0 and the conflicting indicator (ID_C) will show the value 0 (careless response pattern not detected).
Example	The respondent answers the question "Are you a vegetarian?" with "yes" and another questions "Do you like meat" with "yes". Therefore, the respondent gives the same answers to different questions which are in conflict to each other.
Calculation	$$Conflicting\ value: v_C = \sum_{n=0}^{conflicting\_questions} |x_{first\_answer} - x_{current\_answer}|$$ $$Conflicting\ indicator: ID_C(v_C) = \begin{cases} 0, & v_C \neq 0 \\ 1, & v_C = 0 \end{cases}$$ $$Conflicting\ \ index: I_C = \frac{\sum_{n=0}^{questions} ID_C}{questions}$$
Literature reference	• Desimone, et al., 2015 • Wimmer, et al., 2018
Remarks	

Table 7. Careless response pattern "Conflicting Answers". A respondent gave the same answer to dissimilar questions.

3.3.4 Diligence and Interest

Included in SQT	YES
Description	Some questions will be designed by the survey analyst to check, whether the respondent is willing to answer the questions of the survey or not. Therefore, the survey analyst will mark these special questions with a correct answer. If the respondent gives not the "correct" answer there is the possibility that this response will be prone to the "diligence and interest" response pattern. Therefore, the diligence and interest value (v_{DS}) saves the given answer to such a question. The diligence and interest indicator (ID_{DS}) will show the value 1, if the respondent answered a question with a "forbidden" answer, otherwise the diligence and interest indicator (ID_{DS}) will show the value 0.
Example	Asking at the beginning of the survey questions like: "I carefully read the instructions of this survey!" or at the end of the survey "I answered the questions of the survey in a qualitative way".
Calculation	$$Diligence\ and\ Interest\ value: v_{DS} = x_{answer}$$ $$Diligence\ and\ Interest\ indicator: ID_{DS}(v_{DS}) = \begin{cases} 0, & v_{DS} \neq v_{forbidden_{answers}} \\ 1, & v_{DS} = v_{forbidden_{answers}} \end{cases}$$ $$Diligence\ and\ Interest\ \ index: I_{DS} = \frac{\sum_{n=0}^{questions} ID_{DS}}{questions}$$
Literature reference	• Francavilla, et al., 2018 • Ward, et al., 2018 • Desimone, et al., 2015 • Meade, et. al., 2012
Remarks	

Table 8. Careless response pattern "Diligence and Interest".At the start or at the end of a survey some questions will be designed to check, whether the respondent is motivated to answer the survey.

3.3.5 Don't Know

Included in SQT	YES
Description	The "don't know" response pattern will be detected, when the respondent significantly selects the answering option "don't know" instead of a high-quality answer. Therefore the "don't know" value (v_D) will be incremented, if the respondent selects the "don't know" answering option. If the "don't know" indicator (ID_d) is above a predefined value ($\overline{v_D} = 1$) the current response is prone to the careless response pattern "don't know".
Example	**Du findest hier unterschiedliche Aktivitäten aufgelistet. bitte gib jeweils an, ob du auf die genannte Art zumindest hin und wieder körperlich aktiv bis.** ja / nein / weiß nicht Ich treibe in meiner Freizeit gezielt Sport um zu trainieren Ich bin in meiner Freizeit aktiv, aber meist spontan und ohne festes Trainingsziel Ich bin im Rahmen der Schule, meiner Ausbildung oder meinem Beruf körperlich aktiv Ich bin bei einem oder mehreren Sportvereinen aktiv
Calculation	$$Don't\ Know\ value: v_D = \sum_{n=1}^{per\ question} x_{dont\ know}$$ $$Don't\ Know\ indicator: ID_D \left(\sum_{i=1}^{don't\ know\ values} v_{D\_i} \right) = \begin{cases} 0, & v_D < (\overline{v_D} = 1) \\ 1, & v_D \geq (\overline{v_D} = 1) \end{cases}$$ $$Don't\ Know\ index: I_{D=} \frac{\sum_{n=0}^{questions} ID_D}{questions}$$
Literature reference	• Barge, et al., 2012 • Decieux, et al., 2015 • Wimmer, et al.,2018

	• Zhang & Conrad, 2014
Remarks	

Table 9. Careless response pattern "Don't Know".If the respondent selected a high number of the "don't know" answering option, this can indicate negative respondent behaviour.

3.3.6 Early Termination

Included in SQT	YES
Description	The termination value (v_T) will save the question, where the respondent terminated the survey. The indicator value (ID_T) will be computed by comparing the stored question number in the termination value to the overall amount of available questions. If the termination value is lower than the overall question amount than the indicator will show 1 (early termination detected) else the indicator will show 0.
Example	
Calculation	$$Termination\ value: v_T = x_{question\_number}$$ $$Termination\ indicator: ID_T(v_T) = \begin{cases} 0, & v_T \geq v_{all\_questions} \\ 1, & v_T < v_{all\_questions} \end{cases}$$ $$Termination\ index: I_T = \mathrm{ID}_T$$
Literature reference	• Wimmer, et al., 2018
Remarks	

Table 10. Careless response pattern "Early Termination". The respondent didn't answer all questions and terminated the survey.

3.3.7 Mahalanobis Distance

Included in SQT	NO
Description	Before this careless response pattern can be detected, an average response pattern must be defined – the so-called gold standard. In the next step, the current response pattern will be compared against this gold standard and the Mahalanobis Distance will be calculated. If the result of the Mahalanobis Distance shows extreme respectively higher values, the current response is prone to negative respondent behaviour.
Example	
Calculation	
Literature reference	• Francavilla, et al., 2018 • Ward, et al., 2018 • DeSimone, et al., 2015
Remarks	The Mahalanobis distance was dismissed from the implementation of SQT, because some scientific work indicate, that using this technique will deliver inaccurate results (Egan & Morgan, 1998).

Table 11. Careless response pattern "Mahalanobis Distance".Comparing a predefined gold standard response pattern to the actual pattern of the respondent.

3.3.8 Open-ended Questions

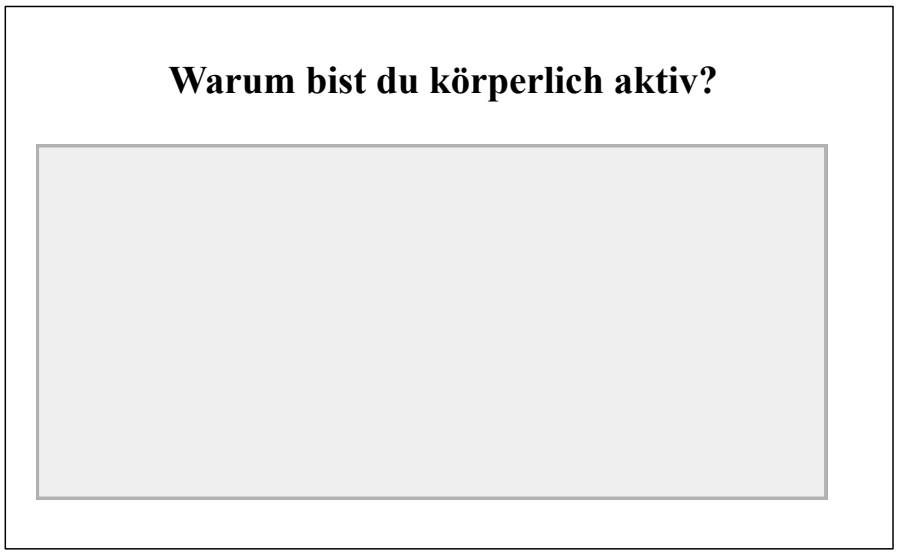
Included in SQT	NO
Description	Related work showed, that the overall amount of given words per open-ended question will be an indicator for the quality of the response.
Example	
Calculation	
Literature reference	• Brüggen, et al., 2010 • Huang, et al., 2012 • Muñoz-Leiva, et al., 2010
Remarks	The given description of this careless response pattern has one big drawback. If a respondent answers an open-ended question only with nonsense and this nonsense had many words then this answer will be ranked better than a short perfect fitting answer. Because of this fact, the author of this thesis dismissed this careless response pattern from SQT. This was the case, because in the opinion of the author, this careless response pattern has not enough power to show negative respondent behaviour.

Table 12. Careless response pattern "Open-ended Questions".The quality of a given answer for an open-ended question is related to the overall amount of used words.

3.3.9 Semantic Synonyms

Included in SQT	YES
Description	If the respondent does not give the same answer to the same question this may be an indicator for the semantic synonyms response pattern. To detect this careless response pattern, the semantic synonyms value (v_{SS}) will be calculated by saving the first answer option and subtracting always the current answer option of the same question. If the same answer option was selected for the same question the delta value will be 0 (careless response pattern not detected) otherwise 1 (careless response pattern detected).
Example	In an online survey the following question showed up "Do you like your job?" and the respondent answered "yes". Later on, the following question showed up "Are you happy with the job" and the respondent answered "no". The respondent gave different answers to almost the same question.
Calculation	$Semantic\ Synonyms\ value: v_{SS} = \sum_{n=0}^{same\ questions} \lvert x_{first\_answer} - x_{current\_answer} \rvert$ $\text{Semantic Synonyms } indicator: ID_{SS}(v_{SS}) = \begin{cases} 0, & v_{SS} = 0 \\ 1, & v_{SS} \neq 0 \end{cases}$ $Semantic\ Synonyms\ index: I_{SS} = \frac{\sum_{n=0}^{questions} ID_{SS}}{questions}$
Literature reference	• DeSimone, et al., 2015 • Meade, et al., 2012 • Francavilla, et al., 2018 • Ward, et al., 2018 • Huang, et al., 2012
Remarks	

Table 13. Careless response pattern "Semantic Synonyms".If the respondent gave dissimilar answers to the same question.

3.3.10 Speeding

Included in SQT	YES
Description	Speeding will be measured with an index as mentioned in (STEINBRECHER, ROßMANN, & BERGMANN, 2013). This index will be calculated as follows: In a first step, the median of the duration of all respondents who completed the whole survey is calculated. To dismiss runaway values the top 5 percent quantile is excluded for calculating the median. For respondents who have a respond time (v_S) between the median and the top 5 percent quantile an indicator value (ID_S) of 0 will be assigned – careless response pattern not detected. For respond times between one second and the median, an indicator value (ID_S) of 1 will be assigned – careless response pattern detected.
Example	
Calculation	$$Speeding\ value: v_S = t_{Respondent}$$ $$Speeding\ indicator: ID_S(v_S) = \begin{cases} 0, & v_S \geq \bar{v}_S \\ 1, & v_S < \bar{v}_S \end{cases}$$ $$Speeding\ Index: I_S = \frac{\sum_{n=0}^{completed\ responses} ID_S}{completed\ responses}$$
Literature reference	• Steinbrecher, Roßmann, & Bergmann, 2013 • Zhang, et al., 2014 • Wimmer, et al.,2018 • Barge, et al., 2012 • Desimone, et al., 2015 • Meade, et al., 2012 • Huang, et al., 2012 • Ward, et al., 2018 • Francavilla, et al., 2018
Remarks	

Table 14. Careless response pattern "Speeding".This pattern will be active, if a respondent answered the questions of the survey faster than the average. Such a behaviour is prone to negative respondent behaviour.

3.3.11 Straightlining

Included in SQT	YES
Description	Selecting a significant number of same answers in series per question matrix can be an indicator for the straightlining response pattern. The calculation of this careless response pattern will be performed by saving each selected answer for the current questions matrix of the respondent in the straightlining value (v_{ST}). If the straightlining value (v_{ST}) contains a significant number of the same answer in a row (v_{ST} greater or equal $\overline{v_{ST}}$) than the straightlining indicator (ID_{ST}) will have the value 1, otherwise 0. $\overline{v_{ST}}$ is defined with 1.
Example	**Du findest hier unterschiedliche Aktivitäten aufgelistet. bitte gib jeweils an, ob du auf die genannte Art zumindest hin und wieder körperlich aktiv bis.** stimmt / stimmt eher / stimmt eher nicht / stimmt gar nicht Ich treibe in meiner Freizeit gezielt Sport um zu trainieren: ○ ◉ ○ ○ Ich bin in meiner Freizeit aktiv, aber meist spontan und ohne festes Trainingsziel: ○ ◉ ○ ○ Ich bin im Rahmen der Schule, meiner Ausbildung oder meinem Beruf körperlich aktiv: ○ ◉ ○ ○ Ich bin bei einem oder mehreren Sportvereinen aktiv: ○ ◉ ○ ○
Calculation	$Straightlining\ value: v_{ST} = \sum_{n=1}^{per\ question} x_{selected}$ $Straightlining\ indicator: ID_{ST}(v_{ST}) = \begin{cases} 0, & v_{ST} < (\overline{v_{ST}} = 1) \\ 1, & v_{ST} \geq (\overline{v_{ST}} = 1) \end{cases}$ $Straightlining\ index: I_{ST} = \frac{\sum_{n=0}^{questions} ID_{ST}}{questions}$
Literature reference	• Francavilla, et al., 2018 • Cibelli, 2017 • McGrath, et al., 2010
Remarks	

Table 15. Careless response pattern "Straightlining".A respondent selected a significant number of the same answer option in series per question matrix.

3.3.12 Trap Questions

Included in SQT	YES
Description	Trap questions are in some sense equal to bogus items, but they give the respondent an instruction to follow and if the respondent does not follow this instruction, this response will be prone to this careless response pattern. The trap value (v_{TR}) stores the answer of the respondent after receiving the instruction of the trap question and the indicator variable (ID_{TR}) will evaluate this answer. If the answer satisfies the instruction the indicator value will be 0 (careless response pattern not detected) otherwise 1 (careless response pattern detected).
Example	These questions instruct the respondent to do something, e.g. "Before starting with the question, answer the first sub question with answering option 2". If the respondent does not comply with the given instruction this can be an indicator for satisficing because the respondent doesn't read the instructions before answering.
Calculation	$$Trap\ value: v_{TR} = x_{answer}$$ $$Trap\ indicator: ID_{TR}(v_{TR}) = \begin{cases} 0, & v_{TR} = v_{followed_{instruction}} \\ 1, & v_{TR} \neq v_{followed_{instruction}} \end{cases}$$ $$Trap\ \ index: I_{TR} = \frac{\sum_{n=0}^{questions} ID_{TR}}{questions}$$
Literature reference	• Miller, et al., 2009 • Desimone, et al., 2015
Remarks	

Table 16. Careless response pattern "Trap Questions".This pattern is related to the bogus items pattern. The difference is, that the respondent has to follow an instruction. If the instruction is not followed this is an indication for negative respondent behaviour.

3.4 Architecture

3.4.1 General Overview

The architecture of SQT (see **Figure 11**) consists of three components – event processing, analytic and statistical output – which will be explained in more detail.

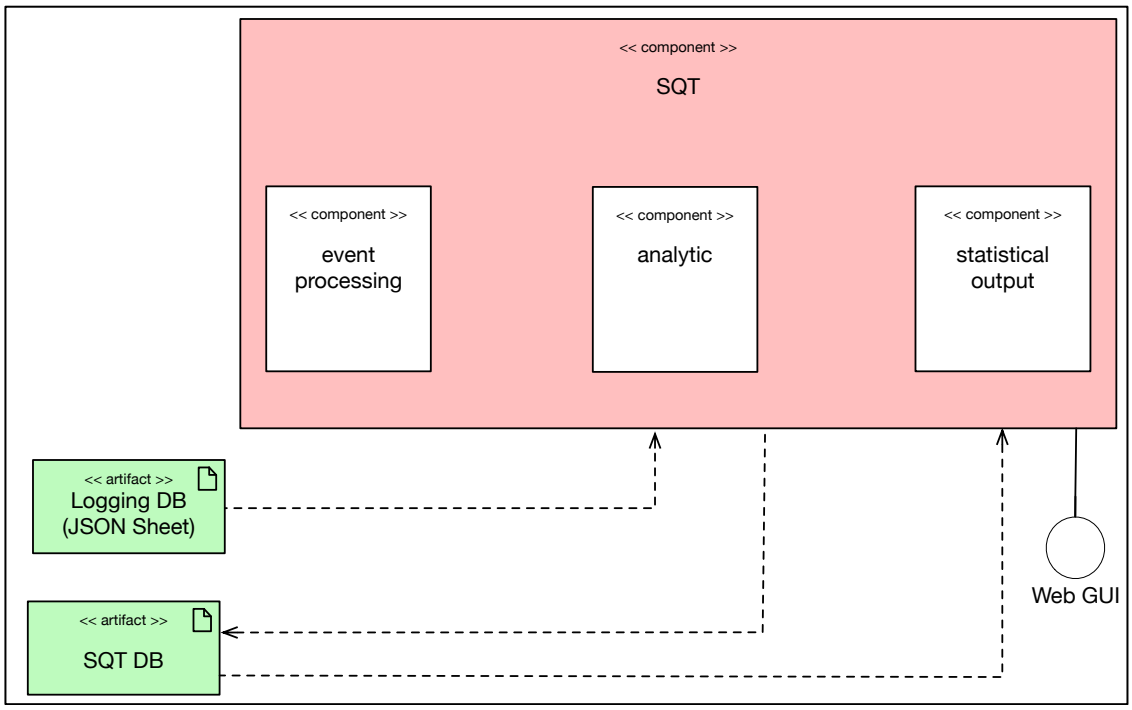


Figure 11. SQT architecture.The event processing component handles the needed input data – response data as JSON file – for SQT. The analytic component calculates the different careless response patterns. The statistical output uses the results from the analytic component and presents them in a Web GUI.

The *event processing component* has the goal to manage the needed input data – survey response data – for the survey quality tool (SQT). The input data can be either automatically logged with the help of an analytical tool like Piwik or Google Analytics or can be created in a manual way. Nevertheless, the input data has to be formatted as a JSON file. The structure of the JSON file is defined via the Extended Backus – Naur form[7]:

actionDetails = {eventCategory, eventAction, timeSpent, eventName, [trapAnswer | compareAnswer | conflictAnswer | dilegenceAnswer | synonymsAnswer]}, eventCategory, eventAction, eventName;

[7] https://en.wikipedia.org/wiki/Extended_Backus–Naur_form (30.08.2020)

eventCategory = “Answering” | “Survey Completion”
eventAction = “Question Text” | “Complete”
timeSpent = “Number”
eventName = “Question Answer ” | “True”
trapAnswer = “Question Answer”
compareAnswer = “Question Answer”
conflictAnswer = “Question Answer”
dilegenceAnswer = “Question Answer”
synonymsAnswer = “Question Answer”

Question Text = ? all available characters to define a question text ?
Question Answer = ? starting from “A1” (first answer option) to “An” (for the last answer option)?
Number = ? integer number ?

Furthermore **Table 17** gives the reader of this thesis more information about the structure of the JSON file and also a practical example will be explained below.

JSON Attribute	Description	Careless response pattern
“actionDetails”	JSON object, which has the logging information of the response.	
“eventCategory”	Either indicates that the following data relates to a question or that no more data for this response is available.	
“eventAction”	Holds either the data for the question text. For example, “Frage 1:Dein Geschlecht?” or holds the information that the response has no more data (keyword “Complete”).	Early Termination (keyword “Complete”)
“timeSpent”	Holds the information about the needed time to give an answer for the current question. For Example, “15”.	Speeding
“eventName”	Holds either the information about the answer for the current question “A1;A2;” or indicates that a response is completed (keyword “True”)	Anchoring Don’t Know Straightlining
“trapAnswer”	Holds the information about the trap answers. For Example, “A1”.	Trap Questions
“compareAnswer”	Holds the information about the compare answers. For Example, “A1”.	Bogus Items
“conflictAnswer”	Holds the information about the conflict answers. For Example, “A1”.	Conflicting Answers
“dilegenceAnswer”	Holds the information about the diligence answers. For Example, “A1”.	Diligence and Interest
“synonymsAnswer”	Holds the information about the synonyms answers. For Example, “A1”.	Semantic Synonyms

Table 17. Component description of the logging JSON sheet.

Detailed description of the different components of the logging JSON sheet and which component is needed to calculate the defined careless response patterns.

Practical example of an entry in the JSON file:

{"actionDetails": *{"eventCategory": "Answering","eventAction": "Frage 1: How old are you?", "timeSpent": "15", "eventName": "A1;" }}*

The *analytic component* of SQT calculates the different predefined careless response patterns based on the input data. In addition, the detection of negative respondent behaviour – careless response pattern value, indicator variable and index – will be saved in a data base (SQT DB).

The *statistical output component* uses the results from the analytic component and presents the detected negative respondent behaviour in an easy and understandable way in a Web GUI.

3.4.2 Used Technologies

SQT uses the following technologies:

- SQL is used for Database queries.
- JSON is used for the logging JSON file.
- Java Script and XHTML is used for the web interface (client part of SQT).
- PHP is used on the server side of SQT.

3.5 The Resulting Survey Quality Tool

SQT can be used either as a web-based tool (hosted as a web application) or as a stand-alone application. In the web-based version SQT offers a user administration interface for survey analysts. In addition to the user administration interface, SQT consists of an overview of all available online surveys, an overview about a single online survey – including the presence of detected careless response patterns – and a more detailed overview of each response based on the detected careless response patterns per question. Another feature of SQT is, the possibility to compare two surveys to see the difference in the careless response patterns at a glance. Based on the statistical output of SQT, the survey analyst has the possibility to dismiss responses with low quality and therefore increase the survey quality. After the evaluation of the different response patterns the survey analyst can export the survey – including response data, careless response pattern detection values – as a CSV file for further statistical evaluation.

If the survey analyst uses SQT as a web-based tool, he had to create a user account for SQT. After the registration process, the survey analyst is able to use SQT with the registered user credentials (see **Figure 12**).

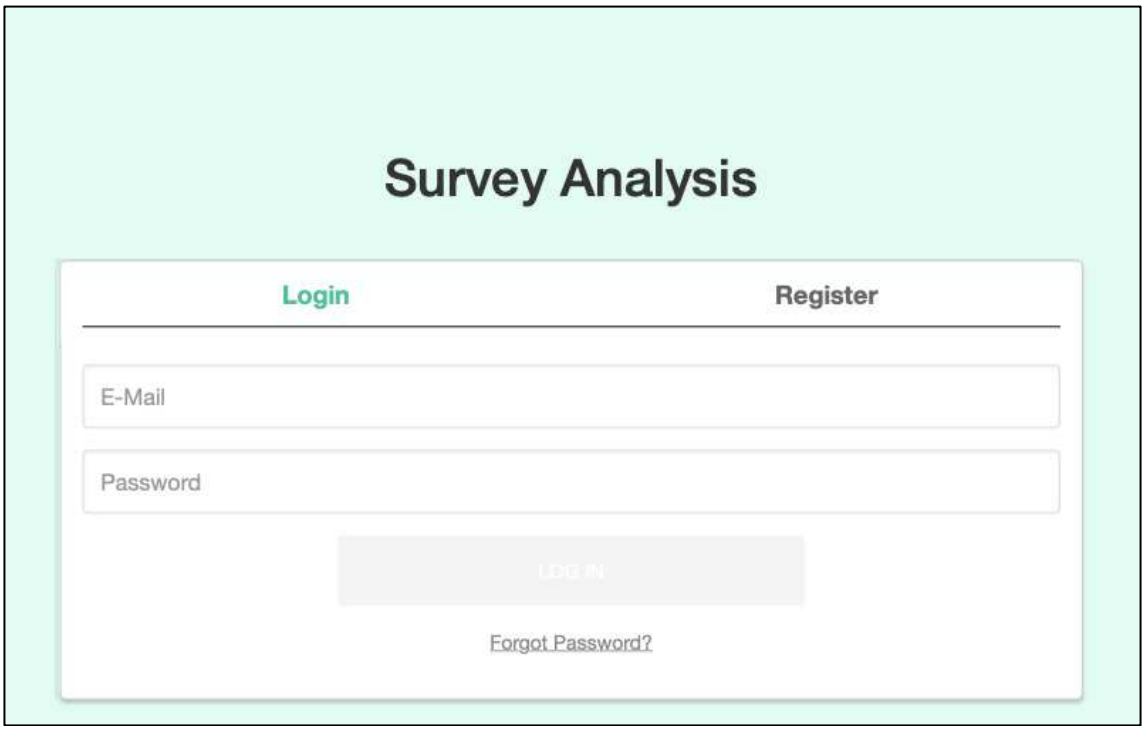


Figure 12. SQT log-in page.Before using the web-based version of SQT the survey analyst has to register. Thereafter, the survey analyst can use his user credentials to start working with SQT.

After the log in, all surveys for the survey analyst will be shown on the *available online survey* page (see **Figure 13**). In this view, the survey analyst has the possibility to interact with the already added surveys. This can be done with different action buttons – information, export, show survey link and delete survey – located next to each survey. If the survey analyst wants to add a new survey, he can press the "Add Survey" button.

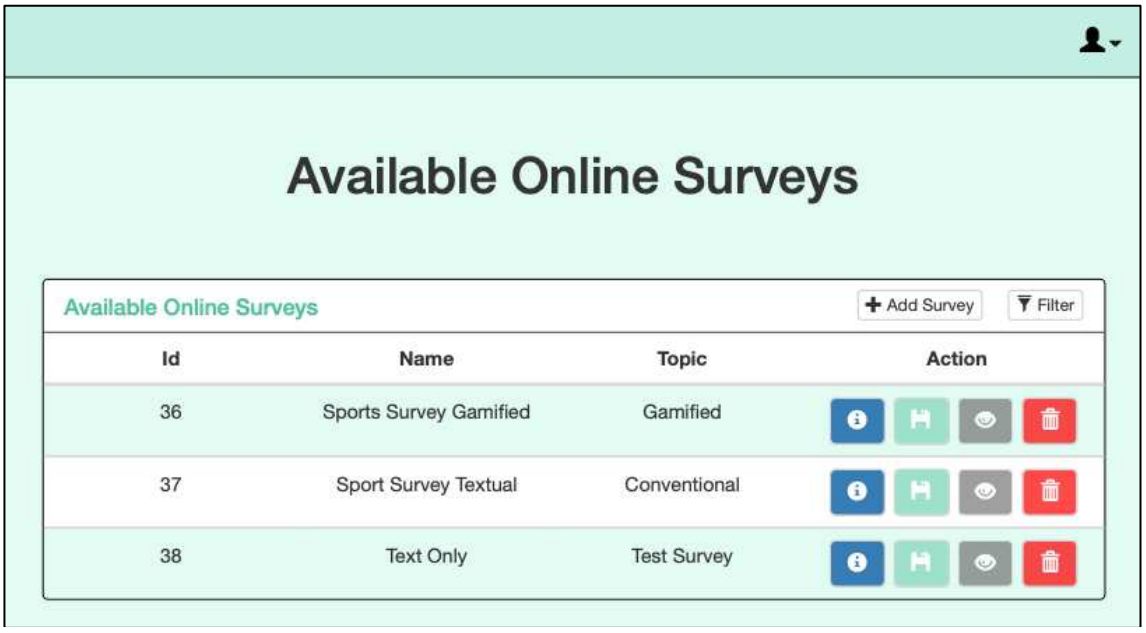


Figure 13. SQT's available online surveys.This view shows all surveys of the currently logged in survey analyst and also offers the possibility to add a new survey. Furthermore, with the help of the different action buttons – information, export, show survey link and delete survey – the survey analyst can interact with each survey.

After that, the *new survey* view will appear (see **Figure 14**). In this view the survey analyst has the possibility to upload response data in SQT. This view contains the general attributes, detection number and measurement section.

First of all, the survey analyst has to select the *general attributes* of the survey. These general attributes contain the name, topic and if available, the survey link to the survey. Furthermore, the coding of the first and the last answer option (e.g. A1) must be inserted by the survey analyst. This information will be used later on by SQT to calculate the different careless response patterns. In the next step the *detection number* for the different careless response pattern must be selected. The detection number defines, how many respondent answers per question matrix must be available for the calculation of the careless response patterns.

The last part of the attributes section is defined as *measurement section*. In this section all available careless response patterns will be listed. Each careless response pattern can be either activated – SQT will calculate this careless response patterns – or deactivated – SQT will ignore this careless response patterns. After setting the measurement attributes the survey analyst has to upload the JSON input file and thereafter press the "Add Survey" button. This will activate the SQT calculation process for the different careless response patterns. During the calculation, SQT saves the careless response values and indicators into the database and redirects the survey analyst to the *available online surveys* page, which presents the analysis results for the new survey.

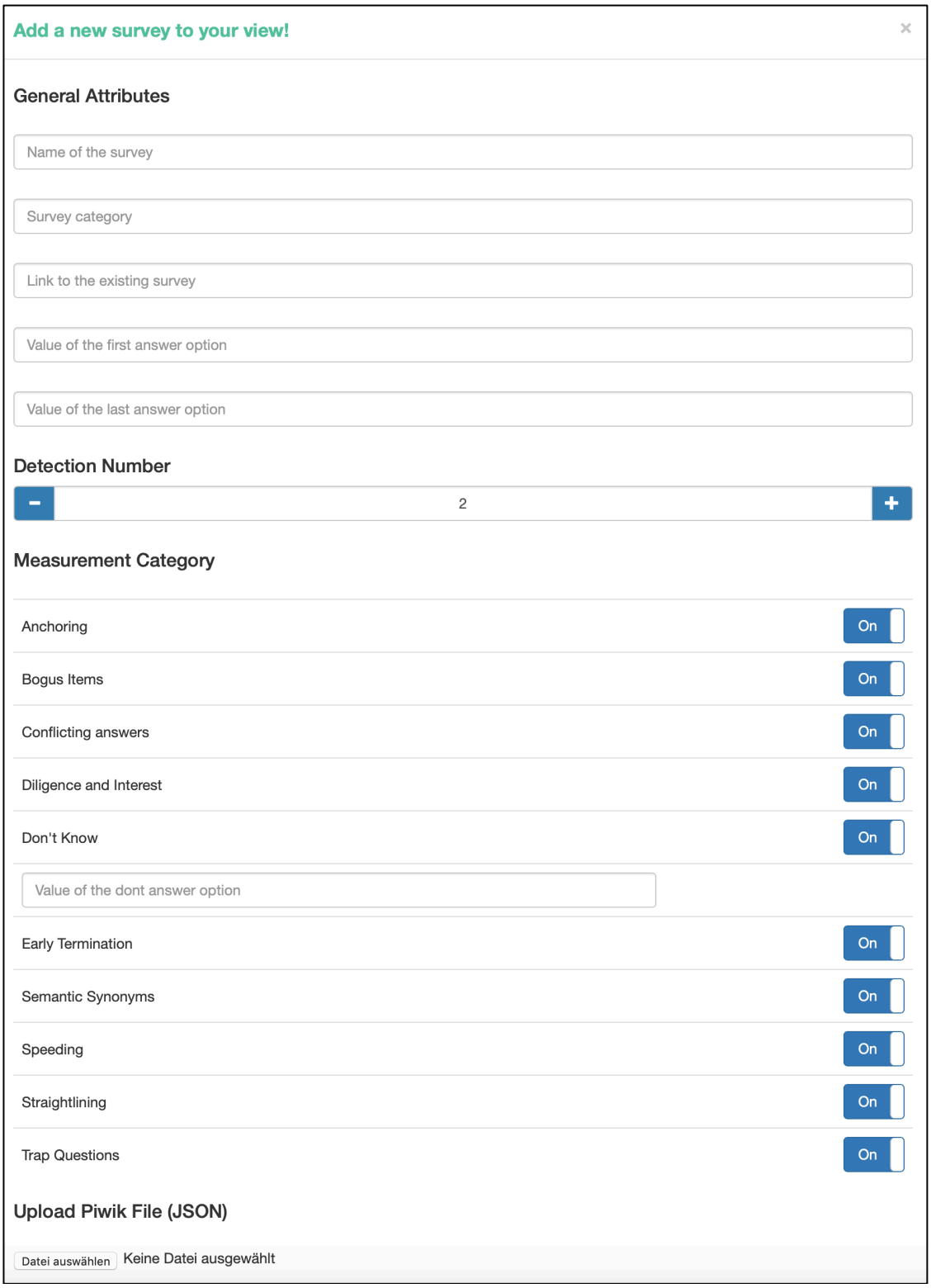


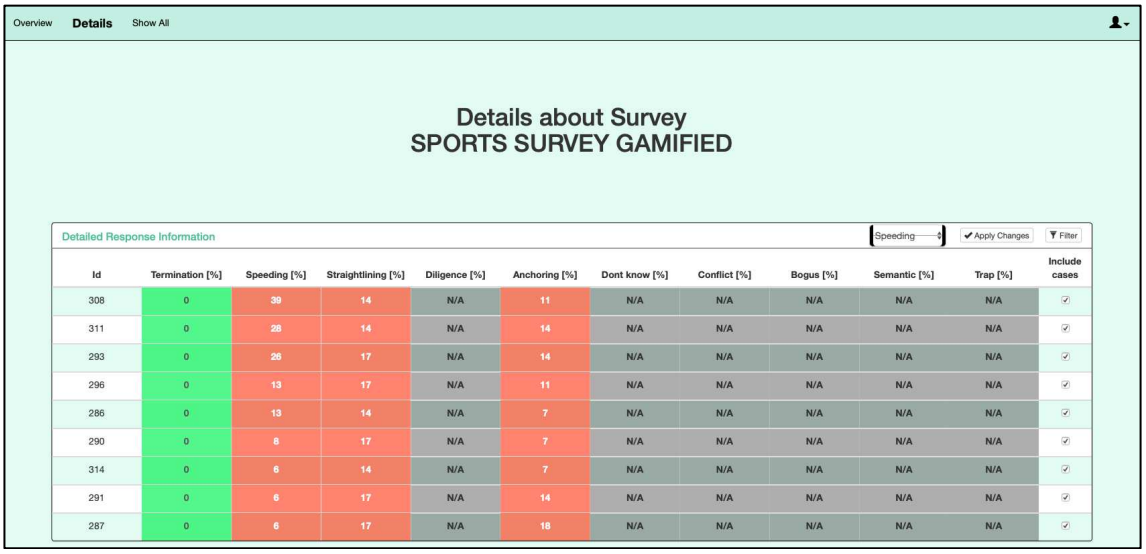
Figure 14. SQT adding a survey.The general attributes section includes the name, topic and if available the link to the current survey. In addition, the first and the last answer option (e.g. A1) must be defined by the survey analyst. The detection number defines, how many questions per question matrix must be available that SQT calculates the careless response patterns. In the measurement section, the different careless response patterns can be activated – SQT will calculate the careless response

pattern – or deactivated – SQT will ignore this careless response pattern. After the survey analyst had selected all attributes the input JSON file must be uploaded to SQT.

To get more information about the careless response patterns for a specific survey the survey analyst has to click the information button in the action column. This action will redirect the survey analyst to the *survey overview* page (see **Figure 15**). The *survey overview* page offers a detailed view of all careless response patterns with additional information. The additional information contains the detection rate– detected, not detected or not available (N/A) – of the careless response pattern for all responses for this specific survey. Besides the additional information, the *survey overview* page shows the amount of responses where the careless response pattern was detected. If a careless response pattern was detected in one response, the font colour of this specific careless response pattern will change to blue. This should attract the attention of the survey analyst and to motivate him to click on the specific careless response pattern. If the survey analyst performs this action, he will be redirected to the *response details* page (see **Figure 16**).

Figure 15. SQT survey overview.The survey overview page presents all careless response patterns to the survey analyst. Furthermore, additional information is available, which contains the detection

rate – detected, not detected or not available (N/A) – of the careless response pattern and how many responses were pollute with careless response patterns.



Id	Termination [%]	Speeding [%]	Straightlining [%]	Diligence [%]	Anchoring [%]	Dont know [%]	Conflict [%]	Bogus [%]	Semantic [%]	Trap [%]	Include cases
308	0	39	14	N/A	11	N/A	N/A	N/A	N/A	N/A	☑
311	0	28	14	N/A	14	N/A	N/A	N/A	N/A	N/A	☑
293	0	26	17	N/A	14	N/A	N/A	N/A	N/A	N/A	☑
296	0	13	17	N/A	11	N/A	N/A	N/A	N/A	N/A	☑
286	0	13	14	N/A	7	N/A	N/A	N/A	N/A	N/A	☑
290	0	8	17	N/A	7	N/A	N/A	N/A	N/A	N/A	☑
314	0	6	14	N/A	7	N/A	N/A	N/A	N/A	N/A	☑
291	0	6	17	N/A	14	N/A	N/A	N/A	N/A	N/A	☑
287	0	6	17	N/A	18	N/A	N/A	N/A	N/A	N/A	☑

Figure 16. SQT response details.Index 0% indicates that the careless response pattern was not detected. Index 100% indicates, that in each question of this response the careless response pattern was detected.

The main goal of the *response details* page is to give the survey analyst an overview about the quality of the different responses, so he can decide, which responses may be excluded from further statistical evaluation because of low-quality data. This can be done via the „Include cases“ checkbox. To fullfil the main goal of the *response details* page, it offers a detailed view about the careless response pattern and their related index values – converted into percentage. Index 0% indicates, that the careless response pattern was not detected. Index 100% indicates, that in each question of this response the careless response pattern was detected. The higher the index value for a given careless response pattern the higher the chance for a low-quality response. To get a better overview about the different index values, the survey analyst has the possibilty to apply a filter to each careless response pattern. If a careless response pattern is detected in a response – the index value is higher than 0% – the survey analyst has the possibilty to click on the index value. This action will open a new view with more details about the detected careless response pattern (see **Figure 17**, **Figure 18**, **Figure 19** and **Figure 20**). The information, which is shown in the new view, depends on the careless response pattern. For example, if the survey analyst clicks on the speeding index, the new window will show the time spent of the respondent per question (see **Figure 17**). With the help of this view, the survey analyst will have a detailed view about the answers given by the respondent and threfore the survey analyst can better decide, whether a single response should be dismissed or not.

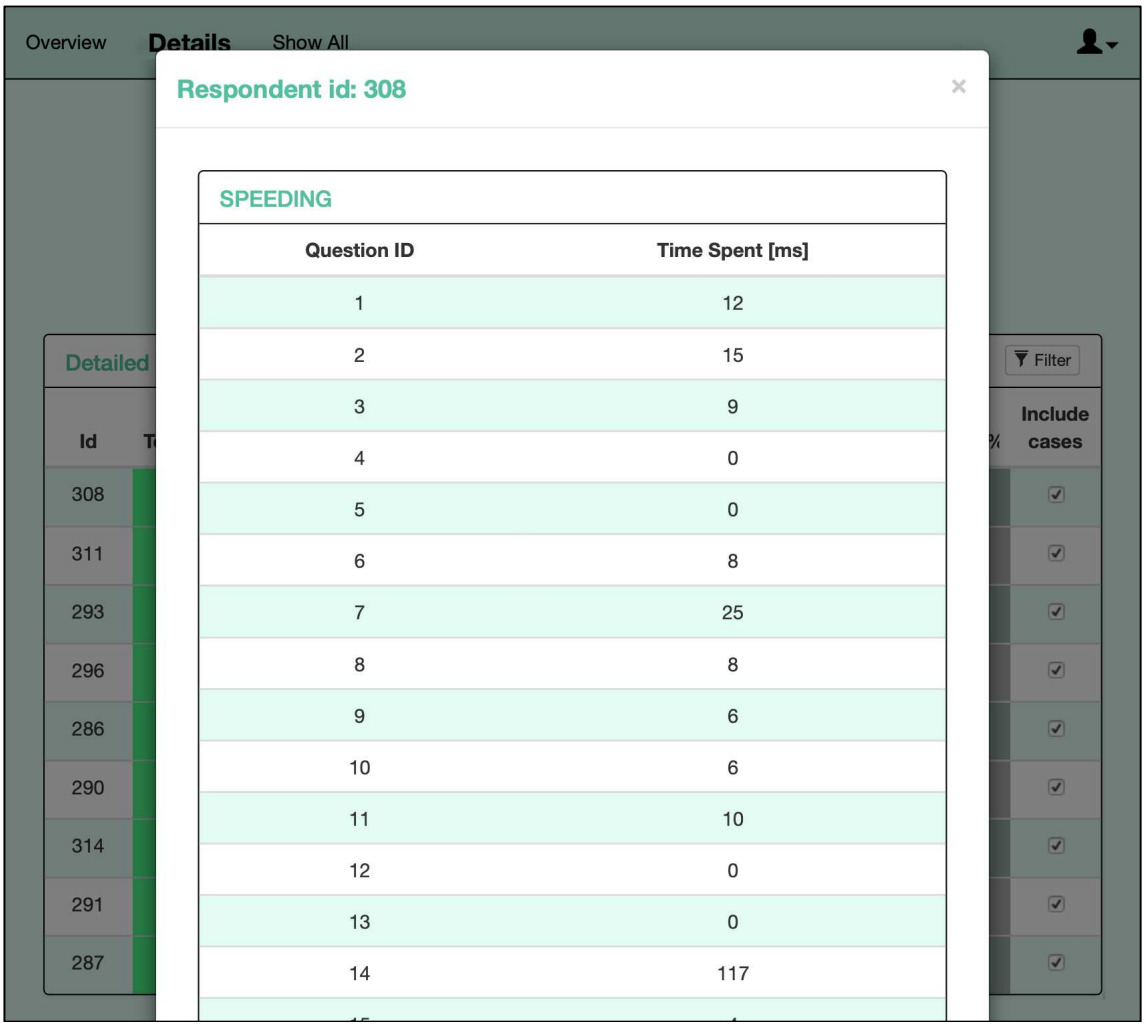


Figure 17. SQT speeding view.The speeding view presents the time spent of the respondent in milli seconds per question.

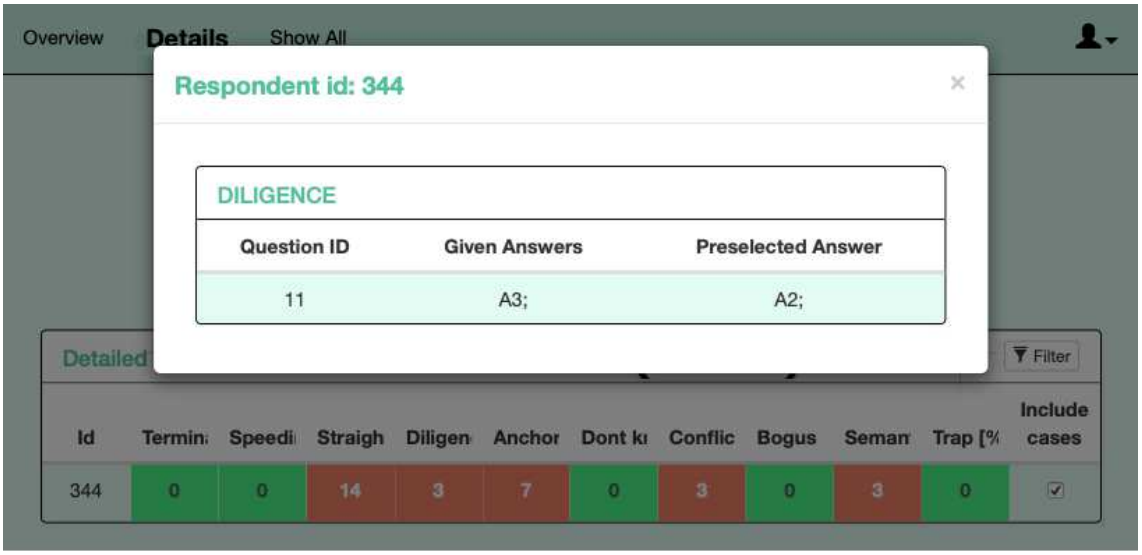


Figure 18. SQT diligence view.The view highlights the question id, if the preselected answer is different to the given answer and therefore the careless response pattern diligence and interests is detected.

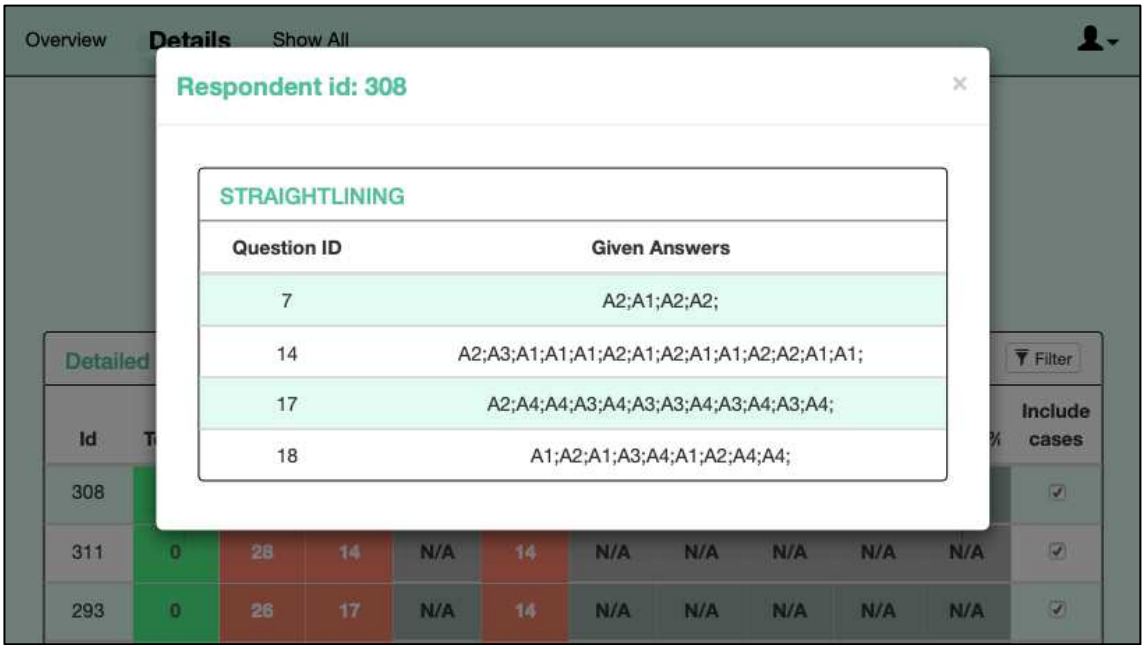


Figure 19. SQT straightlining view.The view highlights the questions, where a significant number of same answers in series per question matrix was given by the respondent.

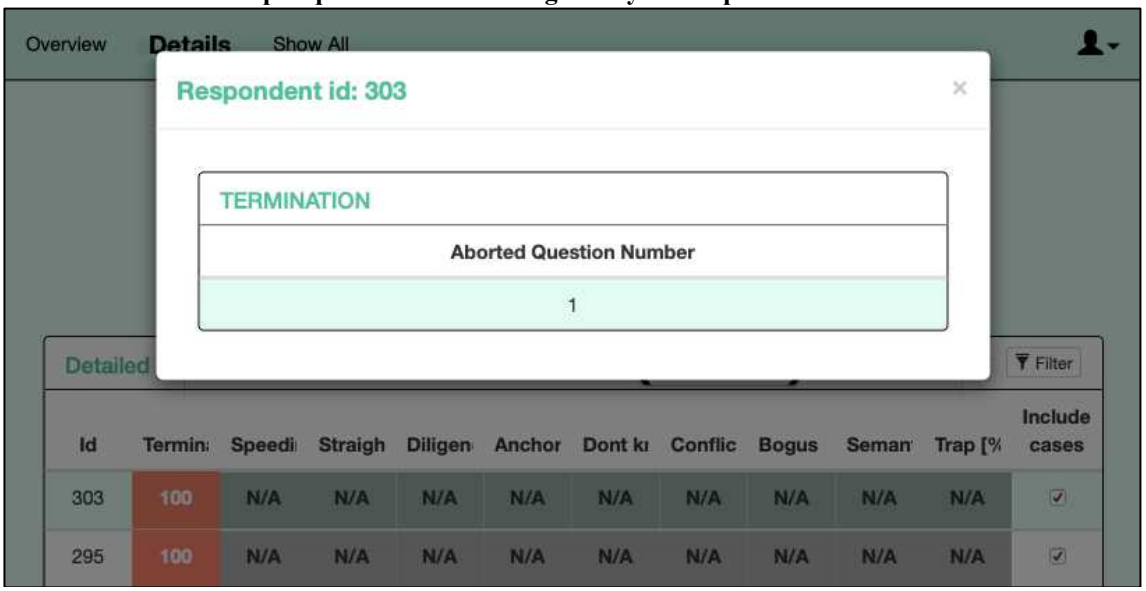


Figure 20. SQT termination view.The view highlights the question, where the respondent had terminated the survey.

Another feature of SQT is the *compare surveys* view (see **Figure 21**). The survey analyst can compare two surveys from the database by means of the different careless response patterns. Therefore, the survey analyst can see at a glance, which careless response pattern was detected in which survey. Provided with this information the survey designer has more insight on different surveys.

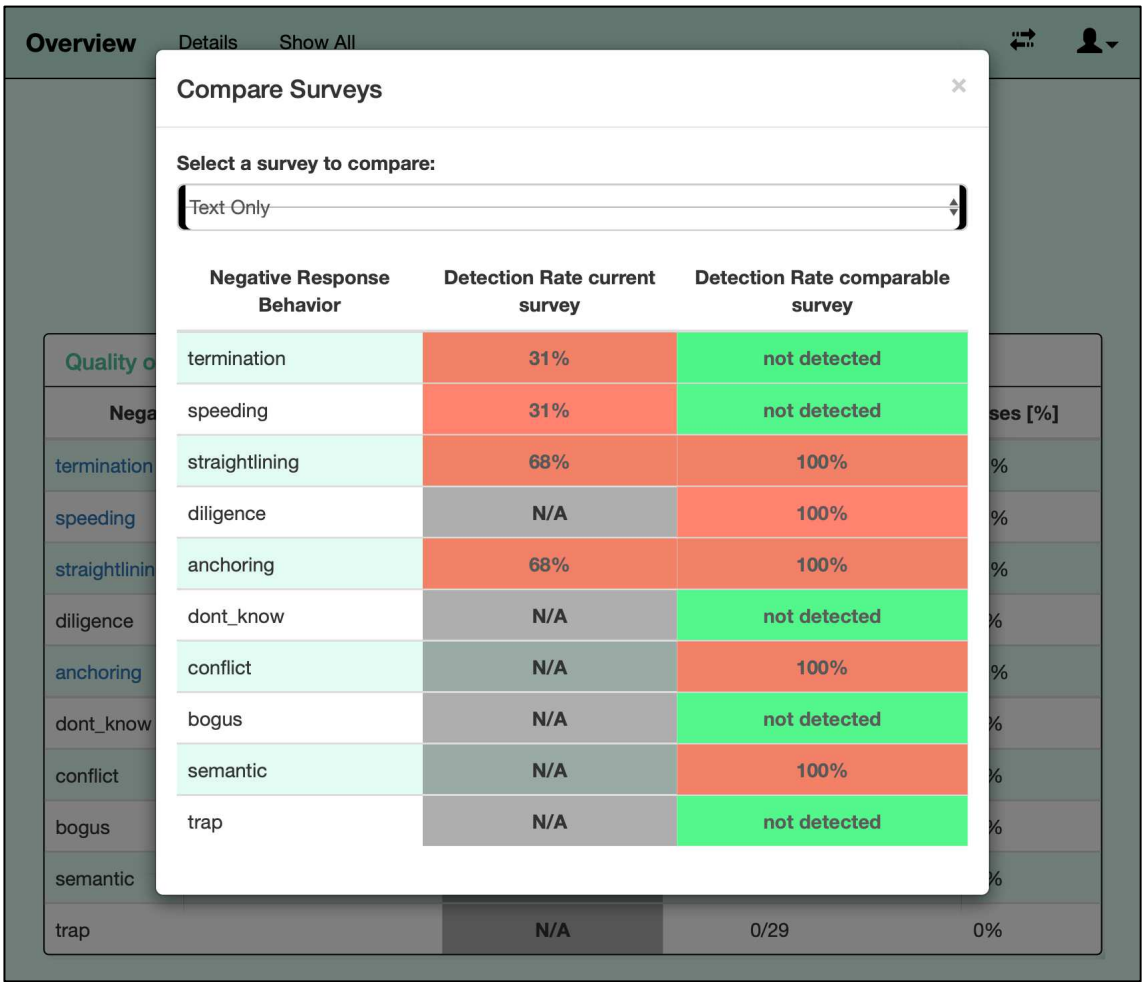


Figure 21. SQT compare surveys.The survey designer has at a glance a view of all careless response patterns for two different surveys.

After the completion of the response data analysis the survey analyst can export the survey data. Therefore, the survey analyst needs to navigate to the *available online survey* page (see **Figure 13**) and click on the green export icon located in the action buttons. Thereafter, a new window will open, where the survey analyst can enter the file name of the exported survey (see **Figure 22**). The survey will be exported as a CSV file and the survey analyst has the possibility to process this file for further statitical analysis with other statistical tools like R or SPSS.

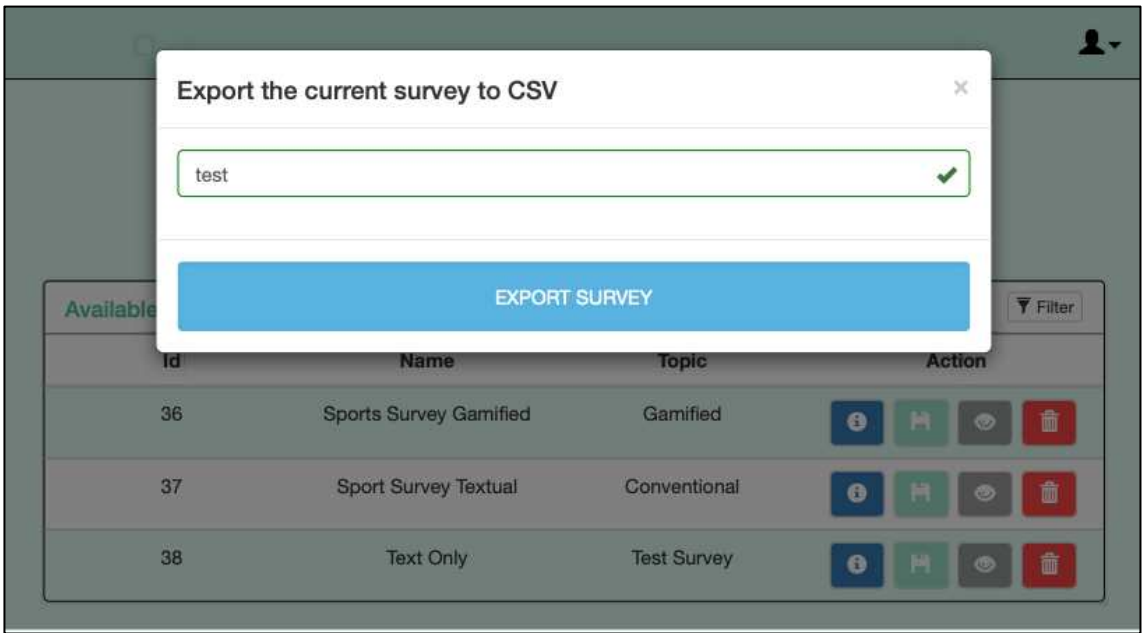


Figure 22. SQT survey export.The survey analyst can export the current survey as a CSV file for further statistical evaluation.

4 Empirical Study 1: Validation of SQT

4.1 Introduction

The goal of empirical study one was to deal with research question one: "Does SQT provide valid output?". The validity of SQT was shown with a two-step validation process (see **Figure 23**). The first step validated the implemented calculation methods of the different careless response patterns with unit tests. In the second step the output of SQT – detection of negative respondent behaviour per response – was validated by manual judgement. The input data for validating SQT was utilized from an online survey used in previous work (HARMS, ET AL., 2015). The title of the survey is "sport and health among teenagers" and a total of 29 responses were collected. The survey was conducted during a period of three months.

4.2 Methodical Approach

Methodically, the author of this thesis, used a two-step sequential validation process (see **Figure 23**.) to answer research question one. Each step was required to show the validity of SQT. If one step failed, this will indicate that SQT is not a valid tool.

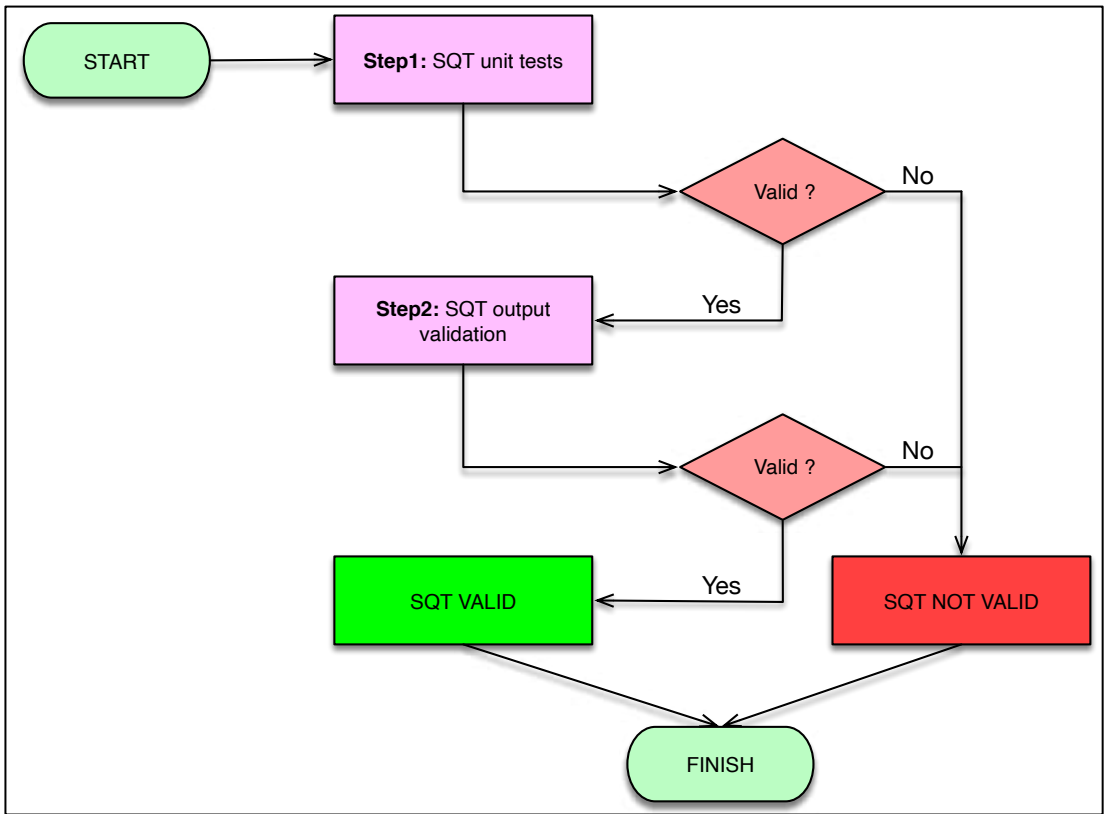


Figure 23. Two-step validation process. The validation process consists of two steps. Step one validated the implementation of the careless response patterns in SQT with unit tests. Step two verified the output of SQT with the help of manual judgment. The result of this process is either SQT is valid or not.

4.3 Data Collection

This part of the empirical study describes, which data was collected to show the validity of SQT and how the validity of SQT was shown.

4.3.1 Unit Tests

As a precondition for step one, of the two-step validation process, the definitions and calculation methods for the careless response patterns were assumed to be valid. This was the case, because the careless response patterns were derived from scientific work found during the literature review and therefore, needed no further evaluation. During step one, of the two-step validation process, the implemented careless response patterns from SQT (compare **Chapter 3.3**) were validated with unit tests (see **Figure 24**).

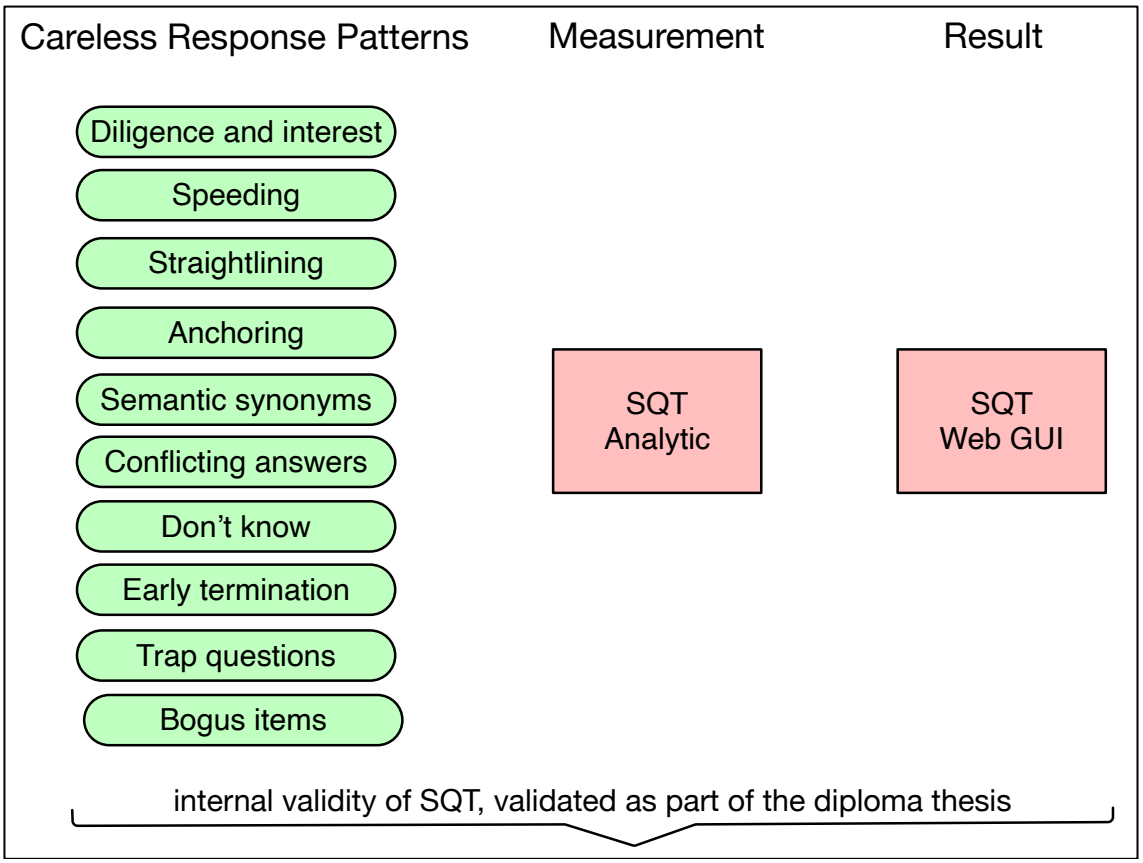


Figure 24. Validation of careless response patterns of SQT.Validated the implemented careless response patterns of SQT with unit tests against the definitions from the literature review.

4.3.2 Output Validation

During step two, of the two-step validation process, the output of SQT – detection of negative respondent behaviour per response – was validated against a manual judgment by the author of this thesis. To perform the output validation of SQT, the author of this thesis evaluated all response data of a survey, which was used in a previous work (HARMS, ET AL., 2015). In this previous work, the survey about sports and health related behaviour among teenagers was chosen, because the survey's questions are easy to understand and an answer without requiring a domain specific expert knowledge can be given. Two designs exist for this survey, one conventional and one gamified. The response data of the gamified version was used to validate SQT's output.

4.4 Analysis

This part of the empirical study describes, how the collected data was evaluated to show that SQT is valid.

4.4.1 Unit Tests

To evaluate, whether the implemented careless response patterns from SQT were valid or not, the author of this thesis developed unit tests for each careless response pattern. The unit tests were created in Java Script and in such a way, that the result of the test cases was automatically matched against the definition of the careless response patterns derived from the literature review. Therefore, it can be assumed, when all unit tests executed successfully the implementation of the careless response patterns into SQT is valid. To give the reader an insight about the definition of the different unit tests one example of a unit test, for the careless response pattern straightlining, will be explained in more detail below:

```
TestStraight_Return_5(test_data_3);

// IN: test_data 3: A2;A2;A2;A3;A3;A3;A3;A3;
// OUT: MaxStraightLine = 5
function TestStraight_Return_5(data)
{
	result = Straightlining(splitAnswers(data, ";"), 1);

	if (result = = 5)
	{
		addTestCaseToTable(4, "Straightlining", data, 5, result, "YES");
	}
	else
	{
		addTestCaseToTable(4, "Straightlining", data, 5, result, "NO");
	}
}
```

The method “TestStraight_Return_5” had one parameter (test_data_3), which will be handed to the test method. The comment section of test case “TestStraight_Return_5” defined the precondition (IN) and the postcondition (OUT) of this test case. After the test method had received the test data and converted the format, this data was transferred to the method “Straightlining” (calculation of the careless response pattern straightlining). In the last step, the returned result of the “Straightlining” function will be compared to the expected result. If the expected result equals the actual result the value “YES” in all other cases a “NO” will be entered in the test table.

4.4.2 Output Validation

Before the author of this thesis started comparing the output of SQT with manual judgment, all response data from the previous work (HARMS, ET AL., 2015) were imported into MAXQDA[8] (see **Figure 25**). The advantage of using MAXQDA was, the response data was well-aranged and it was easier for the author of this thesis to compare the ouput of SQT with the manual judgment.

	1: Id	2: Question Text	3: Response Answer	4: Check Ques...	5: Answering ...
549	287_18	Frage 18: Hier findest du nochmals ein paar Aussagen, bitte gib jeweils an, ob du ihnen sehr zustimm	A2;A2;A2;A3;A2;A3;A2;A2;A1;	null	86
548	287_17	Frage 17: Hier noch ein paar Gruende, koerperlich nicht aktiv zu sein. Bitte gib jeweils an, ob dies	A4;A4;A4;A3;A3;A3;A3;A2;A4;A4;A4;A4;	null	112
547	287_16	Frage 16: Warum bist du koerperlich nicht aktiv?	wenn keine Zeit vorhanden oder wenn es wehwechen gibt;	null	0
546	287_15	Frage 15: Wie sieht es in deinem Freundeskreis aus? Sind die Leute in deinem Freundeskreis alles in	A3;	null	8
545	287_14	Frage 14: Hier noch ein paar Gruende, koerperlich aktiv zu sein, bitte gib jeweils an, ob diese auf	A1;A3;A1;A1;A1;A1;A1;A2;A1;A1;A1;A4;A2;A2;	null	189
544	287_13	Frage 13: Warum bist du koerperlich aktiv?	für die ladies, gesundheit, gewicht zu halten;	null	0
543	287_12	Frage 12: Es gibt ja eine Vielzahl von koerperlichen Aktivitaeten, angefangen vom Fussballspielen ue	laufen, wandern, radfahren, inline skaten, spazieren;	null	0
542	287_11	Frage 11: Und wenn du koerperlich aktiv bist, bist du dann lieber gemeinsam mit anderen Menschen akt	A3;	null	11
541	287_10	Frage 10: Wenn du koerperlich aktiv bist, wo ist das dann?	A3;	null	18
532	287_1	Frage 1:Dein Geschlecht?	A1;	null	1
560	287_0	null	null	null	null
569	286_9	Frage 9:Und wie wuerdest du deine sportlichen und koerperlichen Aktivitaeten einschaetzen? Tust du	A3;	null	10

Figure 25. Imported responses for manual judgment into MAXQDA.Raw data of the responses from previous work (HARMS, ET AL., 2015) imported into MAXQDA.

In the next step the author of this thesis created for each careless response pattern a colour coding (see **Figure 26**).

[8] https://www.maxqda.com (18.03.2020)

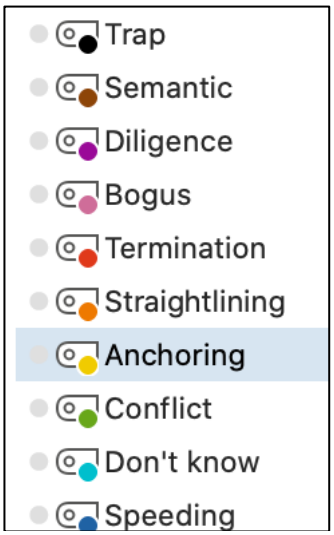


Figure 26. MAXQDA colour coding.Colour coding for the different careless response patterns.

With the help of MAXQDA and the colour coding, the author of this thesis started to rate the different responses according to the definitions of the careless response patterns. One example for the straightlining careless response pattern will be explained below. Response id 287 had for question 17 the following response: "A4;A4;A4;A3;A3;A3;A3;A2;A4;A4;A4;A4;". According to the definition of the straightlining careless response pattern, used in SQT, a sequence of two or more consecutive equal answers indicate straightlining. For question 17, of the current response, this definition was fulfilled for the answer sequence A3 and A4. Therefore, this question was affected by the straightlining careless response pattern and therefore color coded (see **Figure 27**).

548	287_17	Frage 17: Hier noch ein paar Gruende, koerperlich nicht aktiv zu sein. Bitte gib jeweils an, ob dies	A4;A4;A4;A3;A3;A3;A3;A2;A4;A4;A4;A4;	null	112

Figure 27. MAXQDA straightlining example.One example of a response data, which was affected by the straightlining careless response pattern and colour coded with MAXQDA.

In the last step the author of this thesis coded the different careless response patterns for each response (see **Appendix III**) and compared them with the output of SQT. During this step the true positive, true negative, false positive and false negative samples were assessed based on the manual judgment, which was defined as gold standard (see **Table 18**). Later on, the true values from **Table 18** were used to calculate the sensitivity (see **Formula 3**) and the specificity (see **Formula 4**). The sensitivity is defined as the possibilty that SQT detected a negative respondent behaviour correctly for a given response. The specificity is defined as possibility that SQT rejected a response correctly if a response is free of negative respondent behaviour. If the sensitivity and the specificity values were greater than 99% the author of this work assumed that SQT's output is valid.

True Condition	SQT	Manual Judgment
True positive	Careless response pattern detected	Careless response pattern detected
False positive	Careless response pattern detected	Careless response pattern not detected
True negative	Careless response pattern not detected	Careless response pattern not detected
False negative	Careless response pattern not detected	Careless response pattern detected

Table 18. SQT output validation.Validates the output of SQT based on the different true conditions[9].

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negative}$$

Formula 3. Sensitivity of SQT[10].Possibility that SQT detected a negative respondent behaviour correctly for a given response.

$$specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives}$$

Formula 4. Specificity of SQT[11].Possibility that SQT rejected a response correctly if no negative respondent behaviour is available.

[9] https://en.wikipedia.org/wiki/Sensitivity_and_specificity (19.04.2020)

[10] https://en.wikipedia.org/wiki/Sensitivity_and_specificity (08.05.2020)

[11] https://en.wikipedia.org/wiki/Sensitivity_and_specificity (08.05.2020)

5 Empirical Study 2: Usefulness and Practical Applicability of SQT

5.1 Introduction

The second empirical study was designed to deal with research question two: "What is the usefulness, practicability and usability of SQT?". Therefore, SQT was provided to Maximilian Störchle who conducted an online survey about gamification and used SQT to detect negative respondent behaviour (early termination, speeding and straightlining) in his scientific work (Störchle, 2020). During the online survey 117 responses were collected. 17 responses showed the early termination careless response pattern, 61 showed the speeding careless response pattern and 98 responses showed the straightlining careless response pattern. After Maximilian Störchle had finished his work, the author of this thesis conducted an interview with him. Furthermore, the author of this work performed a qualitative analysis based on the interview and a quantitative analysis based on the SUS questionnaire. Both analyses had the goal to detect, whether SQT was useful and practical for Maximilian Störchle's scientific work or not.

5.2 Methodical Approach

Before the author of this work conducted the interview, he created an interview protocol with different questions. This protocol was divided in two parts, first general questions and second usability questions.

The general questions covered the name and the skill of the interviewed person and questions about the usage of SQT.

The second part highlighted the usability of SQT. Therefore, the author of this thesis used questions from the SUS[12] questionnaire to measure the usability of SQT. The advantage using the SUS score are, SUS is a valid tool, it is easy to scale on the participants and it can be used on a small sample size.

[12] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html (13.02.2020)

5.3 Data Collection

After the interview protocol was created the interview was performed. Therefore, an appointment of about 15 minutes was scheduled with Maximilian Störchle. The interview was performed via a Skype audio call. In the first minutes, the author of this thesis asked the first two questions of the general part – his name and his technical skills. Thereafter, Maximilian Störchle was asked to talk about SQT in a general way – without asking the questions of the interview protocol. The author of this thesis noted the different remarks about SQT and assigned the remarks to the related questions of the interview protocol. After Maximilian Störchle had finished his statements about SQT, the author of this thesis asked him specific questions, which were not properly answered. In addition, Maximilian had rated the usability of SQT with the SUS questionnaire provided in the interview. After 15 minutes all questions of the interview protocol were answered, and the interview was finished (see **Appendix IV**).

5.4 Analysis

Before the analysis of the interview started, the author of this thesis created a transcript version of the interview. The results of the interview were analysed and presented with the help of a qualitative analysis method from (BURNARD, GILL, STEWART, TREASURE, & CHADWICK, 2008). In the first step, the transcript of the interview was screened, to create themes and categories, which correlated to the text. Therefore, short phases from the interview were summed up and grouped into categories – this method is called open coding (BURNARD, ET AL., 2008). Generating codes for the open coding method can be done by two methods (Stuckey, 2015): „a priori“ or emergent.

“A priori” means, that some codes were derived from questions of the survey or the interview e.g. related questions to the topic success will be coded to the category success.

In contrast to “a priori” the emergent codes were codes, which evolved from the interview data (e.g. actions, concepts or meanings). For this interview the coding was executed with a combination of “a priori” and emergent codes. The result of the first step was a summary for each element of the transcript (see **Table 19**).

Interview transcript	Initial coding framework
Author of this thesis: “In which context did you used SQT?”	
Maximilian Störchle: “I used SQT in the context of quality control for my own survey for the master thesis. SQT was	• Quality control

used to detect speeding, straightlining and termination. I used SQT as a stand-alone application with the help of a virtual container (Docker). Therefore, I didn't used the user administration of SQT".	• Detect negative respondent behaviour • Stand-alone application
Author of this thesis: "Which positive aspects did you noticed during your work with SQT?"	
Maximilian Störchle: "The import of respondent data into SQT was very easy and simple. This was the case, because SQT uses JSON as a data format for the import of respondent data. Therefore, I could use any survey analysis tool to track the different user actions and therefore import it via JSON into SQT. Another positive feature of SQT was the representation of the different careless response patterns in the Web GUI."	• Easy import of respondent data • Use any survey analysis tool • Good representation of careless response patterns in Web GUI
Author of this thesis: "Which negative aspects did you noticed during your work with SQT?"	
Maximilian Störchle: "I used SQT as a stand-alone application, therefore the installation of SQT was not so easy. To overcome the installation issue, I used a virtual container application (docker). In addition to the table view of the careless response patterns a graphical view would be nice."	• Not easy to install • Graphical view about careless response patterns would be nice
Author of this thesis: "Was SQT useful for your study and why?"	
Maximilian Störchle: "SQT was very useful for detecting careless response patterns (speeding, straightlining, termination) I needed for my master thesis. I used the output of SQT (detection of negative respondent behaviour for one response) to dismiss low quality responses."	• Useful for detecting careless response patterns • Was useful for master thesis • Deselected low-quality response data
Author of this thesis: "Did SQT fulfil your expectations?"	
Maximilian Störchle: "Yes, it fulfilled all the expectations."	• Fulfilled expectations
Author of this thesis: "Is SQT a useful tool for real-world online surveys?"	
Maximilian Störchle: "SQT is a useful tool for real-world surveys and not only for research topics. I can imagine that	• Useful tool in real-world

SQT is also useful for big commercial companies which use online survey for product feedback.”	• Maybe useful also for big companies

Table 19. Initial coding of the interview transcript.Coding of the interview transcript into an initial coding framework.

In the second step, the author of this thesis grouped and removed the duplicated categories to get a consolidated list of categories - a final coding for the interview transcript (see **Table 20**). In the next step, the final coding from **Table 20** was assigned to different colours and with the help of the colour coding the interview transcript was highlighted (BURNARD, ET AL., 2008). For this task the tool MAXQDA was used.

Final coding framework	Initial coding framework

Practical Application of SQT	• Positive Comments
	  • Quality control
	  • Detect negative respondent behaviour
	  • Stand-alone application
	  • Easy import of respondent data
	  • Use any survey analysis tool
	  • Representation of careless response patterns in Web GUI
	• Negative Comments
	  • Not easy to install
	  • Graphical view about careless response patterns would be nice

Usefulness of SQT	• Positive Comments
	  • Useful for detecting careless response patterns
	  • Useful for master thesis
	  • Deselected low-quality response data
	  • Useful tool in real-world
	  • Useful also for big companies
	  • Fulfilled expectations
	• Negative Comments

Table 20. Final coding of the interview transcript.The initial coding will be grouped and merged into umbrella categories to receive the final coding framework.

In addition to the qualitative analysis of the interview transcript, also the SUS questionnaire, which was rated during the interview, was evaluated in terms of usability. The SUS questionnaire consists of ten questions with five answering options ranging from strongly disagree to strongly agree. For each answering option points will be assigned – e.g. the answering option "strongly disagree" gets 0 points, the next answering option gets 1 point and so on. At the end, all points of the ten questions will be summed up resulting in a maximum of 40 points. The resulting points will be multiplied with the factor of 2.5 to get a score between 0 and 100 points. Based on SUS research[13] a score of 68 points will be considered as an average usability for a software program.

The qualitative and quantitative analysis of the interview provided an answer for research question two, whether SQT is useful and practical in a real-world scenario.

[13] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html (13.02.2020)

6 Results

The results of this thesis contain the survey quality tool (SQT) – which was explained in detail in **Chapter 3.5** – and the outcome of empirical study one – covers research question one: "Does SQT provide valid output?" – and empirical study two – covers research question two: "What is the usefulness, practicability and usability of SQT?". In terms of validity of SQT the results showed that all of the 39 unit tests for the different careless response patterns were successfully executed. Furthermore, the sensitivity and specificity values showed 100%. This indicated, that the output of SQT – whether a response shows negative respondent behaviour or not – and the manual judgment of the response data were congruent. Results related to the usefulness, practicability and usability indicated that Maximilian Störchle – who used SQT for his scientific work (STÖRCHLE, 2020) – found SQT useful for detecting negative respondent behaviour. Furthermore, SQT was rated by Maximilian Störchle with a SUS score of 80 out of 100 points.

6.1 Validity

The validity of SQT was shown by the two-step validation process.

Step one showed the validity of implemented careless response patterns from SQT. These careless response patterns were tested with 39 successfully executed unit tests.

In the second step, of the two-step validation process, the output of SQT – whether a response shows negative respondent behaviour or not – was validated against manual judgements. Therefore, the author of this work evaluated all responses of a survey, which was used in a previous scientific work (HARMS, ET AL., 2015). The title of the survey is "sport and health among teenagers" and a total of 29 responses were collected. These response data were rated by the author of this thesis according to the definition of the careless response patterns. **Figure 28** shows the ranking of the detected careless response pattern by manual judgment. 9 responses had terminated before the online survey ended and therefore fulfilled the early termination careless response pattern. The other 20 responses were affected by at least one careless response pattern. The output of SQT can be seen in **Figure 29**. For these two results the sensitivity (possibility that SQT detected a negative respondent behaviour correctly for a given response) and the specificity (possibility that SQT rejected a response correctly if no negative respondent behaviour is available) was calculated. The sensitivity and specificity values for SQT resulted both in 100%. This indicated that the manual judgment of the author of this work and SQT's output were congruent.

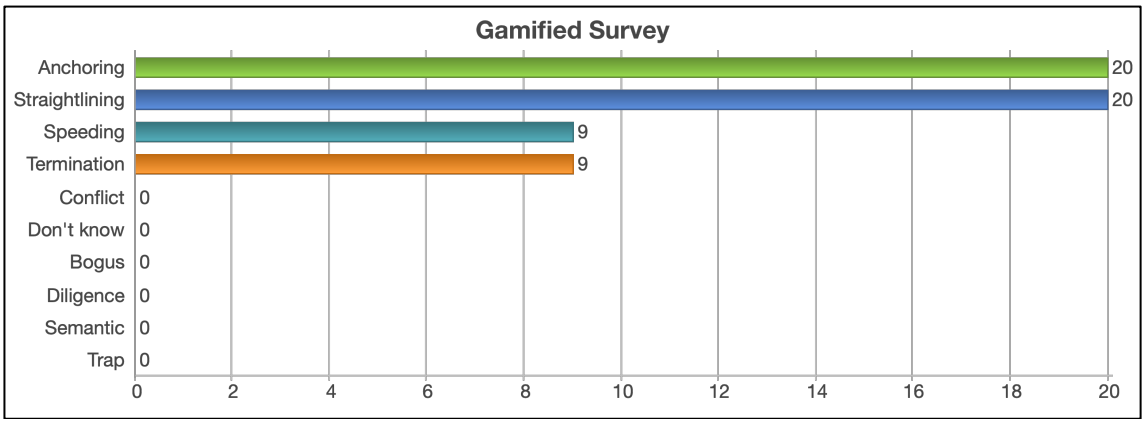
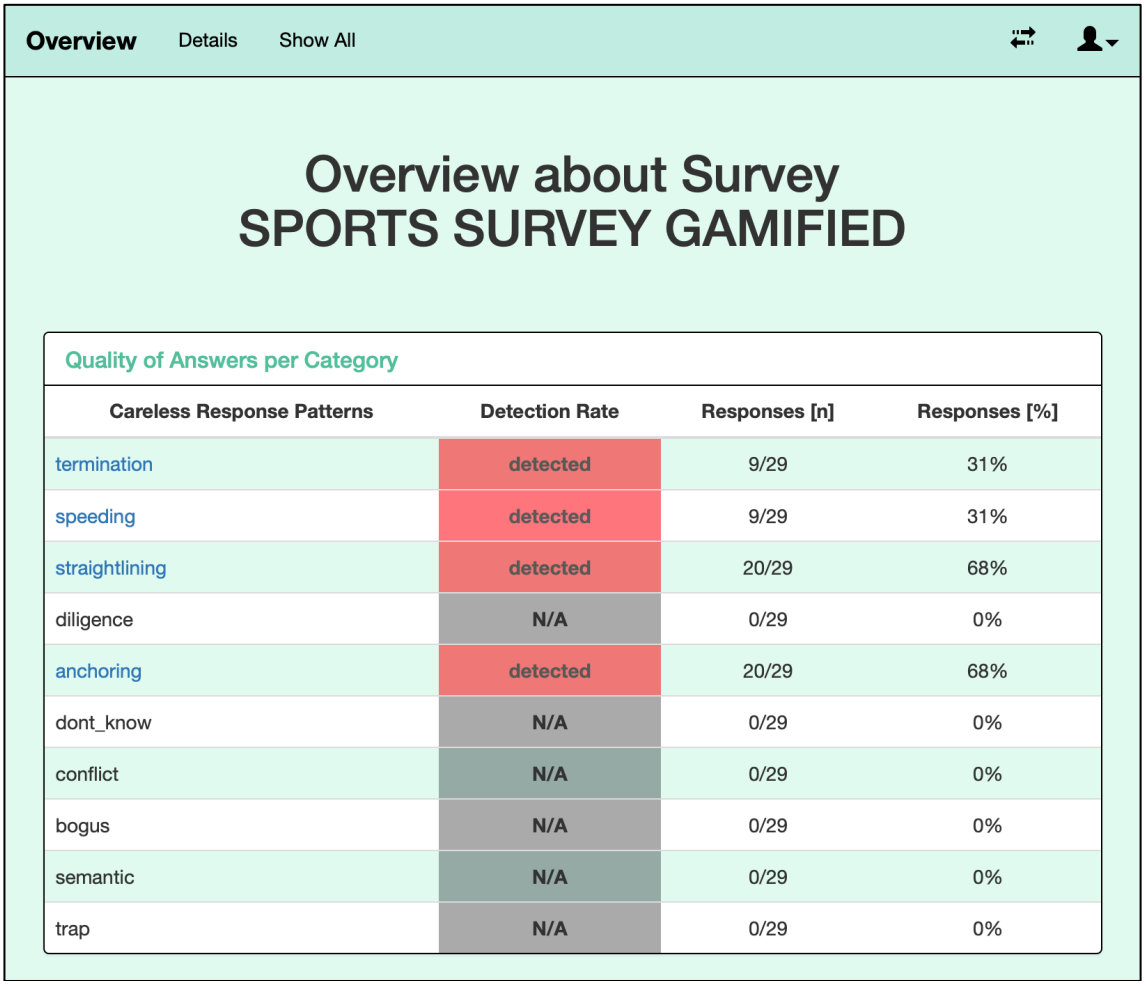


Figure 28. Manual judgment of the responses from previous work.(HARMS, ET AL., 2015). The occurrence of the different careless response patterns was rated by the author of this thesis.



Careless Response Patterns	Detection Rate	Responses [n]	Responses [%]
termination	detected	9/29	31%
speeding	detected	9/29	31%
straightlining	detected	20/29	68%
diligence	N/A	0/29	0%
anchoring	detected	20/29	68%
dont_know	N/A	0/29	0%
conflict	N/A	0/29	0%
bogus	N/A	0/29	0%
semantic	N/A	0/29	0%
trap	N/A	0/29	0%

Figure 29. SQT rating of the responses from previous work.(HARMS, ET AL., 2015). The occurrence of the different careless response patterns was rated by SQT.

6.2 Usability, Usefulness and Practicability

To show the usability, usefulness and practicability of SQT, the author of this thesis distributed SQT to Maximilian Störchle, who used SQT during his scientific work (STÖRCHLE, 2020). To get the needed information about the usability, usefulness and practicability of SQT an interview with Maximilian Störchle was performed after he had finished his scientific work. The results for the usability, usefulness and practicability of SQT indicated, that SQT fulfilled Maximilian Störchle's expectations.

6.2.1 Usefulness

The overall result for the usefulness of using SQT is presented in **Figure 30**. 100% of the answers indicated that SQT was very useful for his scientific work and maybe also useful for other companies using online surveys to get a feedback for their products. The positive comments were related to the usefulness of SQT – maybe useful also for big companies, useful tool in real-world for his scientific work and useful for the master thesis – and to the successful detection of the different careless response patterns – deselection of low-quality responses and useful for detecting careless response patterns. The overall statement given by Maximilian Störchle was, that SQT fulfilled his expectations.

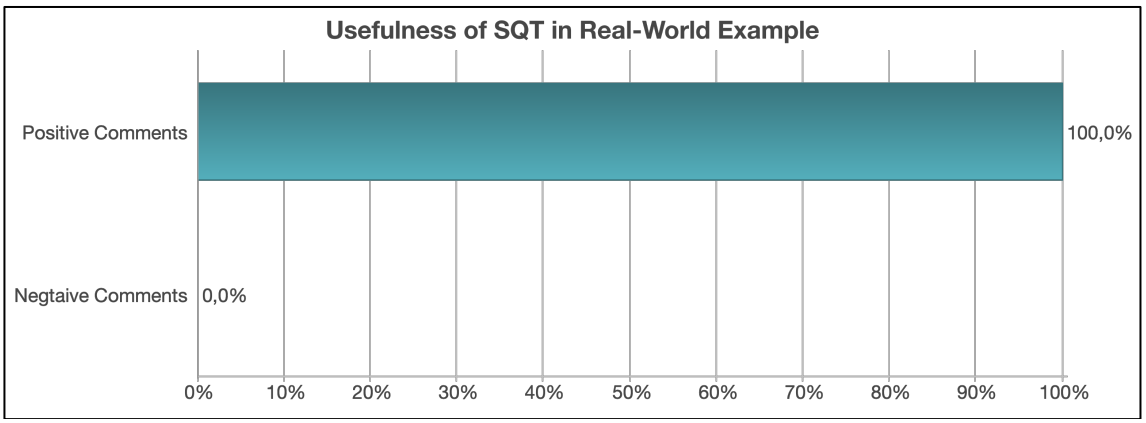


Figure 30. Usefulness of SQT.Facing the positive against the negative comments from Maximilian Störchle given in the interview.

6.2.2 Practicability

The overall result for the practicability using SQT in a real-world example can be seen in **Figure 31**. 75% of the given answers indicated, that SQT was a practicable tool for his scientific work. 25% of the given answers indicated, that there were some problems using SQT. One problem was the installation process of SQT, when it was used as a local version, instead of a hosted version. Another issue, which was mentioned during the interview, was the missing graphical view about the careless response patterns. This feature would give the survey analyst a better understanding

and overview about his survey. In contrast to the negative comments, the positive comments highlighted the clear representation of careless response patterns, the possibility to use any survey analysis tool and the easy import function of response data in SQT. Furthermore, the possibility to use SQT as a stand-alone version and many options for detecting different careless response patterns was also positively mentioned during the interview.

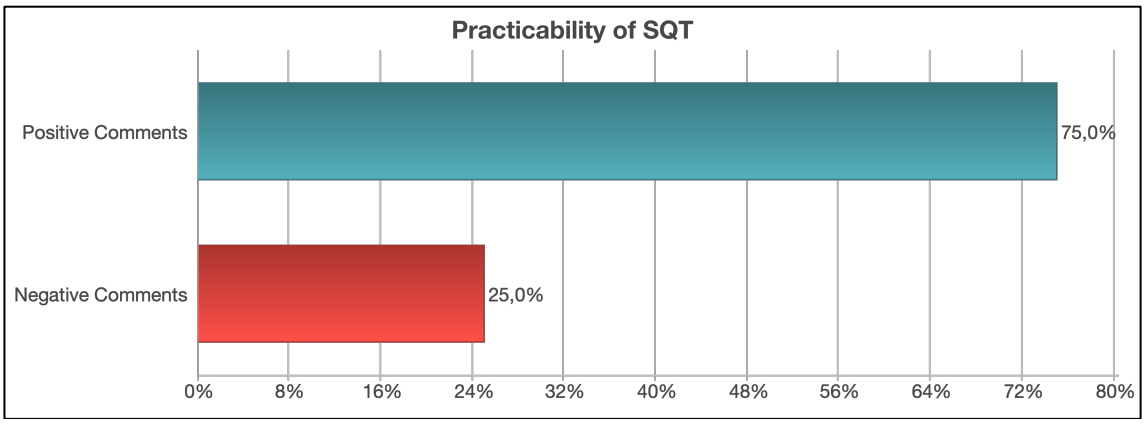


Figure 31. Practicability of SQT.Facing the positive against the negative comments from Maximilian Störchle given in the interview.

6.2.3 Usability

The usability rating the interview can be seen in **Table 21**. A total of ten questions were provided to Maximilian Störchle, who had rated each question with 0 to 4 points. To receive the score for the negative asked questions, the maximum points per question subtracted by the answer given resulted in the points for this particular question. The SUS questionnaire for SQT resulted in a total of 32 points. To norm the resulting value in a 0 (minimum points) to 100 (maximum points) point scale the 32 points were multiplied with the factor of 2.5 resulted in 80 points for SQT. According to the SUS research[14] a total of 68 points can be seen as an average value for the usability of a software system.

[14] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html (13.02.2020)

Nr.	Question	SUS Score
1	The respondent thinks that he would like to use SQT frequently.	4
2	The respondent found the system unnecessarily complex.	(4-1) = 3
3	The respondent thought the system was easy to use.	3
4	The respondent thinks that he would need the support of a technical person to be able to use SQT.	(4-1) = 3
5	The respondent found the various functions in SQT were will integrated.	4
6	The respondent thought there was too much inconsistency in SQT.	(4-0) = 4
7	The respondent would imagine that most people would learn to use SQT very quickly.	2
8	The respondent found the system very cumbersome to use.	(4 – 1) = 3
9	The respondent felt very confident using the system.	4
10	The respondent needed to learn a lot of things before he could get going with SQT.	2
	Overall Score	**32**

Table 21. SUS score for SQT.The different points per questions were noted in the table. For the negative asked questions, the maximum points per questions will be subtracted by the answer of Maximilian Störchle to receive the resulting points.

7 Discussion

This work set out to develop and validate a tool for the automatic detection of negative respondent behaviour in online surveys.

Methodically, the development of SQT started with a literature research about negative respondent behaviour. At the beginning, the author of this thesis used an unstructured approach, resulted in too many and not suitable scientific work. Therefore, the author of this thesis decided to perform a literature review based on the PRISMA framework resulted in 31 high-quality scientific work. Thereafter, SQT was developed and later on evaluated in two empirical studies, which answered the two research questions.

Research question one: "Does SQT provide valid output?" can be positively answered. Results from empirical study one showed successfully the internal validation of SQT.

Research question two: "What is the usefulness, practicability and usability of SQT?" was answered during the interview of Maximilian Störchle, who used SQT during his scientific work (STÖRCHLE, 2020). Maximilan Störchle rated SQT as a useful and practical tool, which fulfils his expectations. Furthermore, he rated the usability of SQT with a SUS score of 80 out of 100 points.

The remainder of this section discussed methodical details, the validity and limitations of the results of this work, as well as directions for future work.

7.1 Methodical Approach

Methodically, the development of SQT started with a literature search to cover the main research goal of this thesis. This goal was to develop, implement and empirically evaluate a survey quality tool (SQT), which automatically detects negative respondent behaviour with the help of careless response patterns.

At the beginning, the author of this thesis started an unstructured literature search with the help of keywords – negative respondent behaviour, online surveys, etc. – related to the main research goal. The application of these keywords in different literature databases resulted in a huge amount of scientific work. This resulted that it was impossible to select relevant literature for answering the main research goal.

To overcome this problem, the author of this thesis decided to terminate the unstructured literature search and performed a structured literature review based on the PRISMA framework. The literature review started with the scoping based on the main research goal of this thesis. This goal was very general and ,in the opinion of the author of this thesis, not suitable for the scoping of the literature review. Therefore, the author of this thesis introduced two more precise questions about negative respondent behaviour in online surveys. The first question was: “What is negative respondent behaviour in online surveys?“ and the second one: “How can negative respondent behaviour be measured and detected in online surveys?“.

The literature review resulted in a total of 70 scientific work, but not all of them were relevant for answering the scoping of the literature review. To dismiss non relevant scientific work, defined filters were used – exclusion and inclusion criteria. Applying such exclusion and inclusion criteria on the result of the literature review may introduce a bias. This can be the case if, a scientific work may be deselected from the result set because of a single exclusion criterion, which is not strictly relevant in the present context – and the decision would be better to include instead of excluding the scientific work in the result set. Furthermore, the judgement if an exclusion/inclusion criterion should be applied or not may be biased or tainted. In the worst case this bias can generate an unexhausted list of careless response patterns by applying a “wrong” exclusion criterion and therefore excluding a suitable careless response pattern. To keep such a bias as small as possible, the author of this work used two countermeasures:

- A pre-defined inclusion, exclusion list from (LIBERATI, ET AL., 2009) will be used. This will help to avoid a bias through post-hoc modifications of inclusion and exclusion criteria.
- The author of this thesis has experience in the current topic, and this experience will be beneficial for his judgements.

Despite the fact of introducing a bias by using inclusion and exclusion criteria, the author of this thesis had decided to apply this method. The decision was based on the following facts. First of all, the author of this work used the above defined countermeasures to keep the bias as small as possible. In addition, the resulted list of 70 scientific work from the literature review significantly varies in terms of context and document quality (ranging from simple PowerPoint presentations to journal papers). Therefore, the author assumed, that overall the positive effect of using inclusion and exclusion criteria on the result quality of the literature review had a bigger impact than the possible negative effect of the bias. After the application of the inclusion and exclusion criteria the scientific work reduced from 70 to 31. Nevertheless, it must be critically mentioned, that applying inclusion and exclusion criteria to the result of the literature review, could result in an unexhausted list of careless response patterns. This had the effect, that some careless response

patterns were not included in SQT. This issue can be highlighted in future work. One possibility to solve this issue would be the implementation of a machine learning approach into SQT. This would solve the issue by learning new negative respondent patterns which were not covered in the literature.

In the next step the different requirements were defined. After the definition of the requirements a mockup prototype was created to get an overview about the Web interface and the usability of the SQT. Thereafter the functional prototype of SQT was developed.

In the last step of the methodical approach SQT was evaluated in two empirical studies.

Empirical study one had the goal to show the validity of SQT. The validity of SQT was split up in three different parts.

Part 1: The validity of careless response patterns was performed during the literature review (see **Figure 32a**).

Part 2: The internal validity was shown by a two-step validation process (see **Figure 32b**).

Part 3: The external validity – psychological effect of the respondent – was postponed in future work, because it was not in the scope of this thesis.

During empirical study two, SQT was provided to Maximilian Störchle who conducted an online survey about gamification and used SQT to detect negative respondent behaviour for his scientific work (STÖRCHLE, 2020). Therefore, the goal of empirical study two was gathering experience and lessons learned regarding the validity of postulated requirements, possible missing requirements, and practical applicability and usefulness of SQT in a real-world scenario. Furthermore, empirical study two validates SQT in terms of usability (see **Figure 32c**). This was done by a SUS questionnaire during the interview with Maximilian Störchle. It must be critically assumed that SQT was distributed only to one survey analyst. The distribution of SQT to a bigger audience – survey analysts – is a goal for future work.

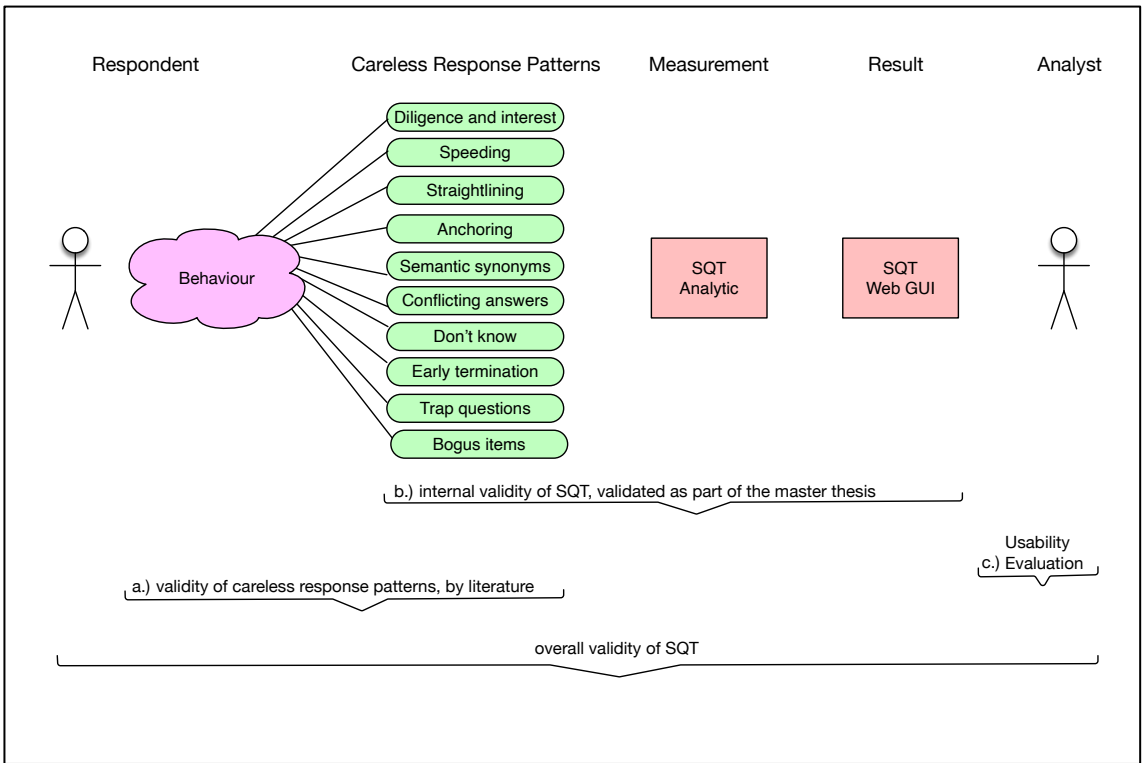


Figure 32. Validation of SQT.The validity of the careless response patterns (a) and the internal validity of SQT (b) – unit tests and output validation – was performed by empirical study one. The evaluation of the usability (c) was performed during empirical study two with the help of the SUS questionnaire.

7.2 Validity of SQT

Results of empirical study one indicated that SQT is a valid tool. Therefore, research question one "Does SQT provide valid output?" can be positively answered because the author of this thesis performed different validity steps (see **Figure 32a-b**), which were described in this chapter.

7.2.1 Validity of Careless Response Patterns

The validation of the careless response patterns (see **Figure 32a**) was performed by the literature review (compare **Chapter 3.1**). This literature review had the goal to find scientific work, which defined different careless response patterns in terms of detection and measurement. Later on, these careless response patterns were used in SQT to detect negative respondent behaviour. Because the careless response patterns for SQT were derived from scientific work resulted from the literature review, the author of this thesis assumed, that the careless response patterns were valid.

7.2.2 Internal Validity

The internal validity of SQT (see **Figure 32b**) consists of two parts.

The first part had the goal to validate the implementation of the defined careless response patterns in SQT. This validation was performed by creating unit tests for each careless response pattern. Each careless response pattern was tested according to their defined behaviour – derived from the literature review. This method resulted in 39 different test cases, which were successfully executed. Therefore, the author of this thesis assumed that the implementation of the careless response patterns in SQT is valid. Nevertheless, it must be critically mentioned, that 39 test cases won't cover all possible use cases of each careless response pattern. But the author of this thesis was convinced, that 39 test cases for ten careless response patterns is a good trade-off between time effort and practicability.

The second part had the goal to validate SQT's output – whether a response shows negative respondent behaviour or not. This validation was performed by using a manual judgment (referenced as a gold standard). Results of empirical study two indicated a sensitivity and specificity value of 100% for the manual judgment of the author of this thesis and SQT's output. This implies, that SQT is able to detect different careless response patterns in the same fashion as a survey analyst.

7.2.3 External Validity

The external validity – the psychological effect of the respondent – was out of the scope of this thesis and will be postponed to future work.

7.2.4 Overall Validity

All parts of the overall validity – careless response patterns, internal validity and external validity (out of scope of this thesis) – showed that SQT is a valid tool for the detection of negative respondent behaviour. Because of these facts, research question one, "Does SQT provide valid output?" can be positively answered. The validity of SQT implies, that a survey analyst who will use SQT, will be able to remove or ignore a low-quality response from the survey result set. This post-hoc adjustments, if carefully conducted, will have a positive effect on the overall response quality as mentioned in WARD, ET AL., (2018) and FRANCAVILLA, ET AL., (2018) and on the statistical power of the survey data (MANIACI, ET AL., 2014).

7.3 Usability, Usefulness and Practicability of SQT

Results from empirical study two indicated, that SQT is useful and pleasant to use. This statement can be derived from the positive comments of Maximilan Störchle, who used SQT during his scientific work (Störchle, 2020).
Maximilian Störchle described the usefulness of SQT with the following terms, SQT fulfilled his expectations, SQT is useful in real-world applications and SQT helped him to detect negative respondent behaviour using the different careless response patterns. Particular the last statement of Maximilian Störchle demonstrates the usefulness of SQT in a real-world environment. The main goal for the author of this thesis was to create SQT for detecting careless response patterns and help the survey analyst to decide whether to dismiss a response from further statistical evaluation or not. According to Maximilian Störchle this goal was completely achieved.

The practicability of SQT was defined by Maximilian Störchle mainly with the following terms, an easy import function for response data, the possibility to use any survey analytic tool to track respondent behaviour and the clear representation of the careless response patterns and their detection rates.

The overall good impression about SQT can also be seen in the scores of the SUS questionnaire, which was used to validate SQT in terms of usability (see **Figure 32c**). SQT received a total score of 80 out of 100 points. Based on the defined average usability score of 68 for a software program[15] the 80 points indicates that SQT's usability is above the average. The main positive statements derived from the SUS questionnaire were, the easy use of SQT, well integrated functions and the confidence of using SQT for detecting negative respondent behaviour.

Despite these positive outcomes the author of this thesis was surprised that the installation process of SQT as a local version and the missing graphical appearance of the careless response patterns were highlighted in a negative way. SQT was intended to use as a web-based version – there is one host and the survey analysts register to the SQT web server. For the scientific work of Maximilian Ströchle, he needed a local version of SQT and therefore had problems to install it. This issue must be highlighted in future work by offering a simple installation process for a local use of SQT. In addition the SUS questionnaire highlighted two questions, which indicate a not optimal usability, question 8 "The respondent would imagine that most people would learn to use SQT very quickly." and question 10 "The respondent needed to learn a lot of things before he could get going with SQT.". Both questions were related to the topic of learning the features and the

[15] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html (13.02.2020)

usage of SQT. To address this issue, the author this paper, will provide in a further work an updated manual with more background knowledge about the different careless response patterns and how they will be detected by SQT. Furthermore, SQT was only distributed to one survey analyst and therefore the expressiveness of the usability, usefulness and practicability of SQT is restricted. Therefore, the author will distribute in a next step SQT to more survey analyst to get a more multifaceted feedback.

7.4 Contributions

Detecting negative respondent behaviour with the help of careless response patterns is a promising way for improving the response quality, reducing the total survey error and therefore improving the statistical power of the response data of an online survey. This work extends prior research by making the following two contributions.

Firstly, this work documents the successful development of a valid quality survey tool (SQT), as displayed in **Figure 33**, which is able to automatically detect negative respondent behaviour based on careless response patterns. Because SQT is a valid tool survey analyst can expect valid results.

As a second contribution, SQT was evaluated in a real-world example by Maximilian Störchle during his scientific work (Störchle, 2020). He used SQT to detect negative respondent behaviour in his online survey to dismiss responses with low-data quality. The application of SQT in the real-world example indicates, that SQT is useful and practical.

Figure 33. SQT as the result of this thesis.

7.5 Outlook

Future work can use this tool and investigate new methods for measuring negative respondent behaviour. The need arises because the used methods, which were applied in SQT for defining careless response patterns, are based on psychological measurements and therefore are very static and can't be adopted dynamically. The examination of future work should focus on ways how a more dynamic detection of negative respondent behaviour can be implemented. One solution to fulfil this need could be the implementation of a machine learning approach, so SQT uses machine learning to classify respondent behaviour and to potentially identify new careless response patterns that are not yet covered in related work.

Furthermore, the need for an easy to use installation for a local version of SQT and a detailed user manual are goals for future work.

Nevertheless, the actual version of SQT is a valid, useful and practical tool for detecting ten different careless response patterns and therefore enables the survey analyst to dismiss low-quality responses from statistical evaluation.

8 Conclusion

Detecting negative respondent behaviour with the help of careless response patterns is a promising way for improving the response quality, reducing total survey error and therefore, improving the statistical power of the response data of an online survey. This work extends prior research by making the following two contributions.

Firstly, this work documents the successful development of a survey quality tool (SQT), which is able to automatically detect negative respondent behaviour based on careless response patterns. Furthermore, the different careless response patterns were derived by a literature review and mathematically defined and integrated into SQT. Thereafter SQT was applied in a case study to validate it with a three-step validation process resulting that SQT is a valid tool. This indicates that survey analyst who use this tool can expect valid results.

As a second contribution, SQT was evaluated in a real-world example by Maximilian Störchle. He used SQT to detect negative respondent behaviour in his online survey to dismiss responses with low-data quality. The application of SQT in the real-world example indicated, that SQT is useful and practical. This was achieved by different features of SQT, like the easy import function of response data, the detection of negative respondent behaviour using different careless response patterns, the good representation of the results of the different careless response patterns etc. These positive results are, however, accompanied by critical issues including problems during the installation of a local version of SQT and the need of a deep background knowledge to understand how SQT work and the meaning of each careless response patterns. These issues warrant further improvement of the tool and explanation.

Future work may investigate the need of a more dynamic detection of negative respondent behaviour. The reason for this need is that the used psychological measurements of negative respondent behaviour are very static and with the help of machine learning SQT can learn new negative respondent patterns which were not covered in the literature. Therefore, an improvement for the future of the tool is to add artificial learning for the detection of negative respondent behaviour.

References

Akbulut, Y. (2015). Predictors of inconsistent responding in web surveys. *Emerald Insight, 25 (1)*, pp. 131-147.

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53(2)*, pp. 182-200.

Beach, D. A. (1989). Identifying the random responder. *The Journal of psychology, 123(1)*, pp. 101-103.

Biegler, S. (2015). Gamification of Online-Surveys.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly, 74(5)*, pp. 817-848.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111(2)*, p. 218.

Brüggen, E., & Dholakia, U. M. (2010). Determinants of participation and response effort in web panel surveys. *Journal of Interactive Marketing, 24(3)*, pp. 239-250.

Bradburn, N. (1978). Respondent burden. *In Proceedings of the Survey Research Methods Section of the American Statistical Association (Vol. 35)*, p. 40.

Burnard, P., Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Analysing and presenting qualitative data. British dental journal, 204(8). p. 429.

Cechanowicz, J., Gutwin, C., Brownell, B., & Goodfellow, L. (2013). Effects of gamification on participation and data quality in a real-world market research domain. *In Proceedings of the first international conference on gameful design, research, and applications*, (S. pp. 58-65).

Cibelli, K. L. (2017). The Effects of Respondent Commitment and Feedback on Response Quality in Online Surveys. *Doctoral dissertation*.

Décieux, J. P., Mergener, A., Neufang, K. M., & Sischka, P. (2015). Implementation of the forced answering option within online surveys: Do higher item response rates come at the expense of participation and answer quality?. *Psihologija, 48(4)*, pp. 311-326.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36(2)*, pp. 171-181.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining" gamification". *In Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, (S. pp. 9-15).

Deutskens, E., Ruyter, K. D., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing letters, 15(1)*, pp. 21-36.

Dolnicar, S., Grün, B., & Yanamandram, V. (2013). Dynamic, interactive survey questions can increase survey data quality. *Journal of Travel & Tourism Marketing, 30(7)*, pp. 690-699.

Egan, W. J., & Morgan, S. L. (1998). Outlier detection in multivariate analytical chemical data. In *Analytical chemistry, 70(11),* (S. pp. 2372-2379).

Francavilla, N. M., Meade, A. W., & Young, A. L. (2018). Social Interaction and Internet-Based Surveys: Examining the Effects of Virtual and In-Person Proctors on Careless Response. *Applied Psychology, 68 (2)*, pp. 223-249.

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. *In Survey research methods (Vol. 2, No. 1)*, pp. 21-32.

Gehlbach, H., & Barge, S. (2012). Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology, 34(5)*, pp. 417-433.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). Survey Methodology. *(Vol. 561)*.

Guin, T. D.-L., Baker, R., Mechling, J., & Ruyle, E. (2012). Myths and realities of respondent engagement in online surveys. *International Journal of Market Research, 54(5)*, pp. 613-633.

Halbherr, V. R. (2017). Data quality in cross-national surveys: a longitudinal and cross-cultural analysis of the quality indicators response rate, fieldwork efforts, and nonresponse bias. *Doctoral dissertation*.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does Gamification Work? — A Literature Review of Empirical Studies on Gamification. *47th Hawaii international conference on system sciences*, (S. pp. 3025-3034).

Harms, J., Biegler, S., Wimmer, C., Kappel, K., & Grechenig, T. (2015). Gamification of online surveys: Design process, case study, and evaluation. *In IFIP Conference on Human-Computer Interaction* (S. pp. 219-236). Springer, Cham.

Harms, J., Seitz, D., Wimmer, C., Kappel, K., & Grechenig, T. (2015). Low-cost gamification of online surveys: Improving the user experience through achievement badges. *In Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, (S. pp. 109-113).

Harms, J., Wimmer, C., Kappel, K., & Grechenig, T. (2014). Gamification of online surveys: conceptual foundations and a design process based on the MDA framework. *In Proceedings of the 8th Nordic conference on human-computer interaction: Fun, fast, foundational*, (S. pp. 565-568).

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27(1)*, pp. 99-114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100(3)*, p. 828.

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. *In Proceedings of the AAAI Workshop on Challenges in Game AI (Vol. 4, No. 1)*, (S. p. 1722).

Huotari, K., & Hamari, J. (2012). Defining gamification: a service marketing perspective. *In Proceeding of the 16th international academic MindTrek conference*, (S. pp. 17-22).

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

Kühne, S., & Kroh, M. (2018). Kühne, S., & Kroh, M. (2018). Personalized Feedback in Web Surveys: Does It Affect Respondents' Motivation and Data Quality? *Social Science Computer Review, 36(6)*, pp. 744-755.

Kaminska, O., McCutcheon, A. L., & Billiet, J. (2010). Kaminska, O., McCutcheon, A. L., & Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly, 74(5)*, (S. pp. 956-984).

Keusch, F., & Zhang, C. (2017). "A review of issues in gamified surveys.". *Social Science Computer Review, 35(2)*, pp. 147-166.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied cognitive psychol- ogy, vol. 5, no. 3*, pp. 213-236.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New directions for evaluation, (70)*, pp. 29-44.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., . . . Moher, D. (2009). "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration.". *Annals of internal medicine 151.4*, pp. W-65.

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public opinion quarterly, 72(5)*, pp. 914-934.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48,*, pp. 61-83.

McGrath, R. E., Mitchell, K. M., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological bulletin, 136(3)*, p. 450.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods, 17(3)*, p. 437.

Miller, J., & Baker-Prewitt, J. (2009). Beyond 'trapping'the undesirable panelist: The use of red herrings to reduce satisficing. *In CASRO Panel Quality Conference (Vol. 2).*

Muñoz-Leiva, F., Sánchez-Fernández, J., Montoro-Ríos, F., & Ibáñez-Zapata, J. Á. (2010). Improving the response rate and quality in Web-based surveys through the personalization and frequency of reminder mailings. *Quality & Quantity, 44(5)*, pp. 1037-1052.

Puleston, J. (2011). Online research–game on!: A look at how gaming techniques can transform your online research. *Proceedings of the 6th ASC (Association for Survey Computing) International Conference*, (S. pp. 20-50).

Schacht, S., Keusch, F., Bergmann, N., & Morana, S. (2017). Web survey gamification–increasing data quality in web surveys by using game design elements. *ECIS 2017 Proceedings.*

Siddaway, A. (2014). What is a systematic literture review and how do I do one? pp. 1-13.

Störchle, M. (2020). Macro vs. Micro Gamification von Online Umfragen Ein empirischer Vergleich vierer Umfragedesigns. *Unpublished manuscript, Technical University of Vienna.*

Steinbrecher, M., Roßmann, J., & Bergmann, M. (2013). The Short-Term Campaign Panel of the German Longitudinal Election Study 2009. *GESIS-Technical Reports 2013 (20).*

Stuckey, H. L. (2015). The second step in data analysis: Coding qualitative research data. *Journal of Social Health and Diabetes, 3(01)*, pp. 007- 010.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology, 67(2)*, pp. 231-263.

Wimmer, C., Biegler, S., Harms, J., Kappel, K., & Grechenig, T. (2018). SQT: A tool for the automated measurement of respondent behaviour and response quality in health-related gamified online surveys. *In 2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)*, (S. pp. 1-8).

Zhang, C., & Conrad, F. G. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *In Survey Research Methods (Vol. 8, No. 2)*, pp. 127-135.

Appendix I

Appendix I presents the scientific work, which was found during the literature review. The search term indicates the query, which was used to find the scientific work. Each scientific work was evaluated with the different inclusion and exclusion criteria and a remark about the decision of the author of this thesis, why this particular scientific work was excluded or included.

# 1	
Search Term	allintitle: (Respondent OR responding OR responder OR response OR result OR feedback OR answer OR reply) AND (effort OR engagement OR random OR quality OR careless OR destructive OR negative) AND surveys AND -telephone AND -mail -population
Number of scientific work found	49

ID	Title	Author	Status	Remark
1.1	Response rate and response quality of internet-based surveys: An experimental study.	(Deutskens, Ruyter, Wetzels, & Oosterveld, 2004)	Included	Explaining the "don't know" careless response pattern.
1.2	Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.	(Kane, & Staiger, 2012)	Excluded	No essential information about negative response behaviour.
1.3	Non-response in student surveys: The role of demographics, engagement and personality.	(Porter, & Whitcomb, 2005)	Excluded	Explaining the non-response error based on a small population (college student) with no essential

				background information.
1.4	Detecting and deterring insufficient effort responding to surveys.	(Huang, Curran, Keeney, Poposki, & DeShon, 2012)	Included	Describing the variety of careless response patterns including the insufficient effort responding.
1.5	Informed consent: Consequences for response rate and response quality in social surveys.	(Singer, 1978)	Excluded	No essential information about negative response behaviour.
1.6	Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point.	(Borgers, Sikkel, & Hox, 2004).	Excluded	No essential information about negative response behaviour.
1.7	Myths and realities of respondent engagement in online surveys.	(Guin, Baker, Mechling, & Ruyle, 2012)	Included	Definition of the respondent burden.
1.8	Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality?	(Smyth, Dillman, Christian, & McBride, 2009).	Excluded	Describes GUI improvements for open-ended questions.
1.9	Improving retention rate and response quality in Web-based surveys.	(Sánchez-Fernández, Muñoz-Leiva, & Montoro-Ríos, 2012)	Excluded	Detailed description about improving of retention rate with the help of reminder messages and prize draws.
1.10	The effect of personalization on response rates and data quality in web surveys.	(Heerwegh, Vanhove, Matthijs, & Loosveldt, 2005)	Excluded	No essential information about

				negative response behaviour.
1.11	Improving the response rate and quality in Web-based surveys through the personalization and frequency of reminder mailings.	(Muñoz-Leiva, Sánchez-Fernández, Montoro-Ríos, & Ibáñez-Zapata, 2010)	Included	Definition of response quality.
1.12	Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.	(Kane, & Staiger, 2012)		Already excluded above, see 1.2.
1.13	Drop downs and scroll mice: The effect of response option format and input mechanism employed on data quality in web surveys.	(Healey, 2007).	Excluded	No essential information about negative response behaviour.
1.14	Item sampling in service quality assessment surveys to improve response rates and reduce respondent burden: The "LibQUAL+® Lite" example.	(Thompson, Kyrillidou, & Cook, 2009).	Excluded	Case study about "LibQUAL+® Lite" no other information was provided.
1.15	The impact of incentives and interview methods on response quantity and quality in diary-and booklet-based surveys.	(Bonke & Fallesen, 2010).	Excluded	Comparison between telephone and web surveys. No new information provided.
1.16	Effects of researcher presence and appeal on response quality in hand-delivered, self-administered surveys.	(Webster, 1997).	Excluded	No essential information about negative response behaviour.
1.17	Superficial survey choice: An experimental test of a potential method for increasing response rates and response quality in correctional surveys.	(Pickett, Metcalfe, Baker, Gertz, & Bedard, 2014).	Excluded	No essential information about negative response behaviour.

1.18	Determinants of participation and response effort in web panel surveys.	(Brüggen & Dholakia, 2010)	Included	Describes a measurement technique for open-ended questions, titled as survey response effort.
1.19	Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys.	(Ward & Pond III, 2015).	Excluded	Type of publication is a master thesis (but used references checked for useful information).
1.20	Item sampling in service quality assessment surveys to improve response rates and reduce respondent burden: The "LibQUAL+® Lite" randomized control trial (RCT).	(Kyrillidou, 2010)	Excluded	Case study about "LibQUAL+® Lite" no other information was provided.
1.21	Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys.	(Jones, House, & Gao, 2015)	Included	Definition of different forms of careless response patterns (speeding, straightlining, etc.)
1.22	Reducing respondent burden, information processing effort, and incomprehensibility in stated-preference surveys: Principles and properties of the pairwise design strategy.	(Wang & Timmermans, 2001)	Excluded	No essential information about negative response behaviour.
1.23	Cognitive probes in web surveys: on the effect of different text box size and probing exposure on response quality.	(Behr, Bandilla, Kaczmirek, & Braun, 2014)	Excluded	No essential information about negative response behaviour.
1.24	Implementation of the forced answering option within online surveys: Do higher item response rates	(Décieux, Mergener, Neufang, & Sischka, 2015)	Included	Definition of random answering.

	come at the expense of participation and answer quality?			
1.25	Investigating cognitive effort and response quality of question formats in web surveys using paradata.	(Höhne, Schlosser, & Krebs, 2017)	Excluded	No essential information about negative response behaviour.
1.26	Editors and researchers beware calculating response rates in random digit dial health surveys.	(Martsolf, Schofield, Johnson, & Scanlon, 2013)	Excluded	No essential information about negative response behaviour.
1.27	Improving the quality of community health surveys and community health promotion campaigns by feedback from the community: experience from the Wallsend Community and Health Project.	(Clarke, Allen, McBay, & Heaney, 1990)	Excluded	No essential information about negative response behaviour.
1.28	Increasing response rates & data quality of Web surveys: Pre-notification and questionnaire paging format.	(Lusinchi, 2015)	Excluded	No essential information about negative response behaviour.
1.29	Using optical mark read surveys: an analysis of the response rate and quality.	(Klose & Ball, 1995)	Excluded	No essential information about negative response behaviour.
1.30	Mobile and dirty: Does using mobile devices affect the data quality and the response process of online surveys?	(Schlosser & Mays, 2018)	Excluded	Only focused on mobile devices and the case study.
1.31	Optimizing response rates and data quality in web surveys: the immediacy effect and prize values.	(Tuten, Galešić, & Bošnjak, 2008)	Excluded	Lottery tickets for response quality, no information about

				negative response behaviour.
1.32	'Designing Questions for Web Surveys: Effects of Check-List, Check-All, and Stand-Alone Response Formats on Survey Reports and Data Quality.	(Dykema, Schaeffer, Beach, Lein, & Day, 2011)	Excluded	Type of publication is a power point presentation.
1.33	Improving data quality, accuracy, and response in on-board surveys: Application of innovative technologies.	(Oliveira & Casas, 2010)	Excluded	No essential information about negative response behaviour.
1.34	Response processes and response quality in business surveys.	(Haraldsen, 2018)	Excluded	No essential information about negative response behaviour.
1.35	Respondent behaviour and data quality aspects in panel surveys: four empirical contributions.	(Serfling, 2006)	Excluded	No useable information for web surveys, only other survey types like mail or telephone surveys were highlighted.
1.36	Response to surveys of high-profile topics: the effects of media coverage and public engagement on response to the National 2009 H1N1 Flu Survey.	(Davis, Singleton, & Balluz, 2011)	Excluded	No essential information about negative response behaviour.
1.37	Limitations of the random response technique and a call to implement the ballot box method for estimating recreational angler compliance using surveys.	(Bova, Aswani, Farthing, & Potts, 2018)	Excluded	No essential information about negative response behaviour.

1.38	Using Personalized Feedback to Increase Data Quality and Respondents' Motivation in Web Surveys?	(Kühne & Kroh, 2016)	Excluded	Publication is not more available.
1.39	Assessing the Effects of Survey Instructions and Physical Attractiveness on Careless Responding in Online Surveys.	(Rauti, 2017)	Excluded	Type of publication is a master thesis (but used references checked for useful information).
1.40	SQT: A tool for the automated measurement of respondent behaviour and response quality in health-related gamified online surveys.	(Wimmer, Biegler, Harms, Kappel, & Grechenig, 2018)	Included	Definition of different careless response patterns.
1.41	Data Quality in Cross-National Surveys. A Longitudinal and Cross-Cultural Analysis of the Quality Indicators Response Rate, Fieldwork Efforts, and Nonresponse Bias.	(Halbherr, 2017)	Included	Detailed description about nonresponse.
1.42	Forced answering in online surveys: Do higher item response rates come at the expense of participation and answer quality?	(Sischka, Decieux, Mergener, & Neufang, 2016)	Excluded	Type of publication is a power point presentation.
1.43	Focus On Family-Centered Outcomes: Understanding Response Rates To Surveys About Symptoms Of Psychological Distress Among Family Members Of Critically Ill Patients.	(Long, Downey, Engelberg, Nielsen, Ciechanowski, & Curtis, 2017)	Excluded	No essential information about negative response behaviour.
1.44	Detecting and Deterring Insufficient Effort Responding to Surveys.	(Huang, Curran, Keeney, Poposki, & DeShon, 2012)		Already included in 1.4.
1.45	The Effects of Respondent Commitment and Feedback on Response Quality in Online Surveys.	(Cibelli, 2017)	Included	Details about careless response patterns (straightlining, etc.)

1.46	Personalized Feedback in Web Surveys: Does It Affect Respondents' Motivation and Data Quality?	(Kühne & Kroh, 2018)	Included	Definition of survey satisfaction.
1.47	Read It From My Fingertips–Can Typing Behaviour Help Us to Predict Motivation and Answer Quality in Online Surveys?	(Hörmann & Bannert, 2016)	Excluded	Type of publication is a poster version.
1.48	The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys.	(McKay, Garcia, Clapper, & Shultz, 2018).	Excluded	Type of publication is a manuscript version.
1.49	Social Interaction and Internet-Based Surveys: Examining the Effects of Virtual and In-Person Proctors on Careless Response.	(Francavilla, Meade, & Young, 2018)	Included	Description of nine careless response indicators.

# 2	
Search Term	allintitle: (Respondent OR responding OR responder OR response OR result OR feedback OR answer OR reply) AND (effort OR engagement OR random OR quality OR careless OR destructive OR negative) AND surveys AND (measure OR scale OR grade OR evaluate OR score)
Number of scientific work found	1

ID	Title	Author	Status	Remark
2.1	Detecting and Deterring Insufficient Effort Responding to Surveys.	(Huang, Curran, Keeney, Poposki, & DeShon, 2012)		Already included in 1.4.

# 3	
Search Term	allintitle: (Respondent OR responding OR response OR result OR feedback OR answer OR reply) AND (effort OR engagement OR random OR quality OR careless OR destructive OR negative) AND surveys AND (detect OR discover OR find OR observe) AND -mail -patients
Number of scientific work found	1

ID	Title	Author	Status	Remark
3.1	Detecting and Deterring Insufficient Effort Responding to Surveys.	(Huang, Curran, Keeney, Poposki, & DeShon, 2012)		Already included in 1.4.

Literature: Detecting and deterring insufficient effort responding to surveys. Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012).

Title	Author	Status	Remark
Identifying the random responder.	(Beach, 1989)	Included	Information about random responses.

Literature: Myths and realities of respondent engagement in online surveys. Guin, T. D. L., Baker, R., Mechling, J., & Ruyle, E. (2012).

Title	Author	Status	Remark
Respondent burden.	(Bradburn, 1978)	Included	Additional information about the respondent burden.
Completion time and response order effects in web surveys.	(Malhotra, 2008)	Included	Information about satisficing.

Literature: Implementation of the forced answering option within online surveys: Do higher item response rates come at the expense of participation and answer quality? Decieux, J. P. P., Mergener, A., Sischka, P., & Neufang, K. (2015).

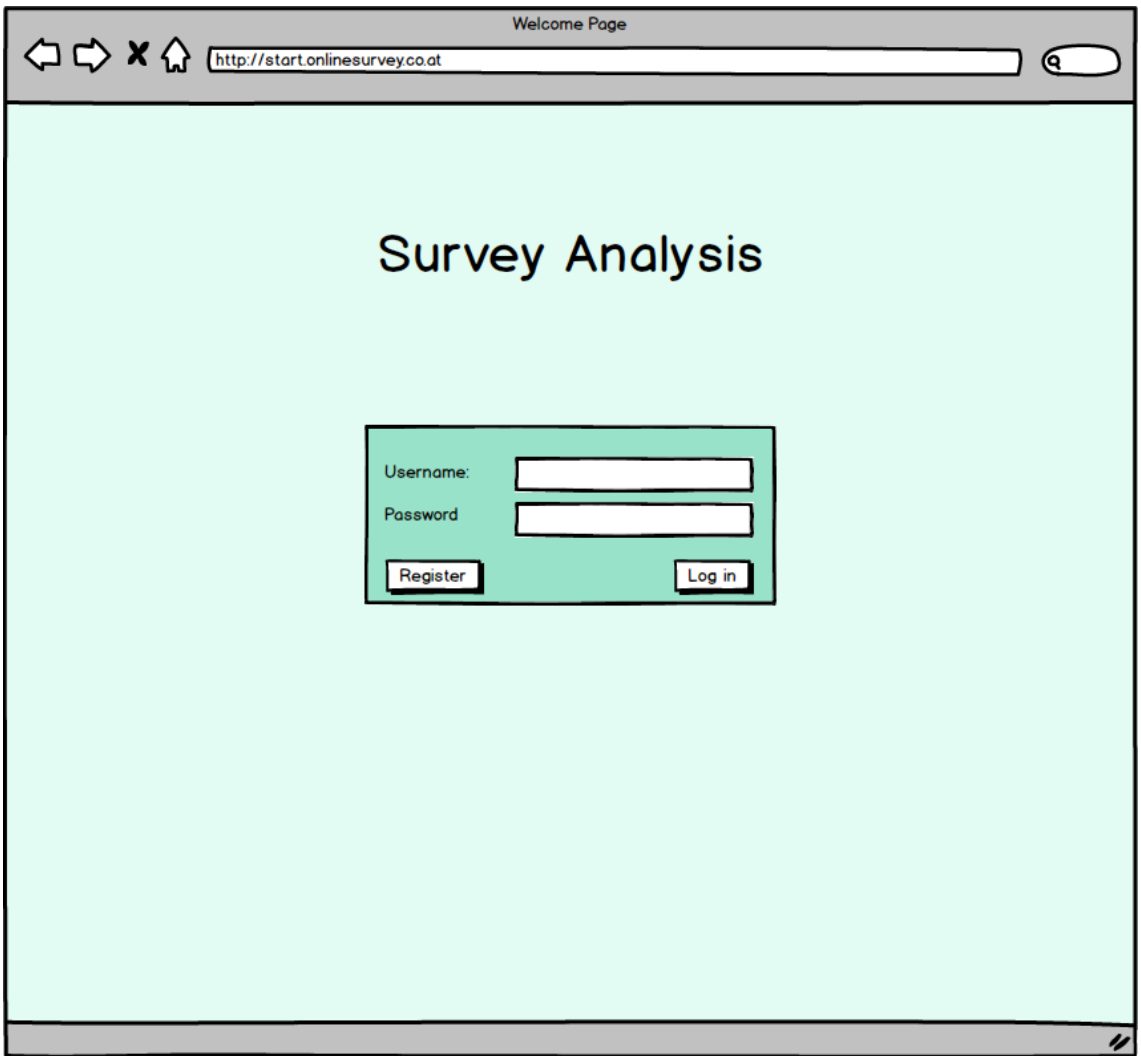
Title	Author	Status	Remark
Satisficing in surveys: Initial evidence.	(Krosnick, Narayan, & Smith, 1996)	Included	Additional information about satisficing.

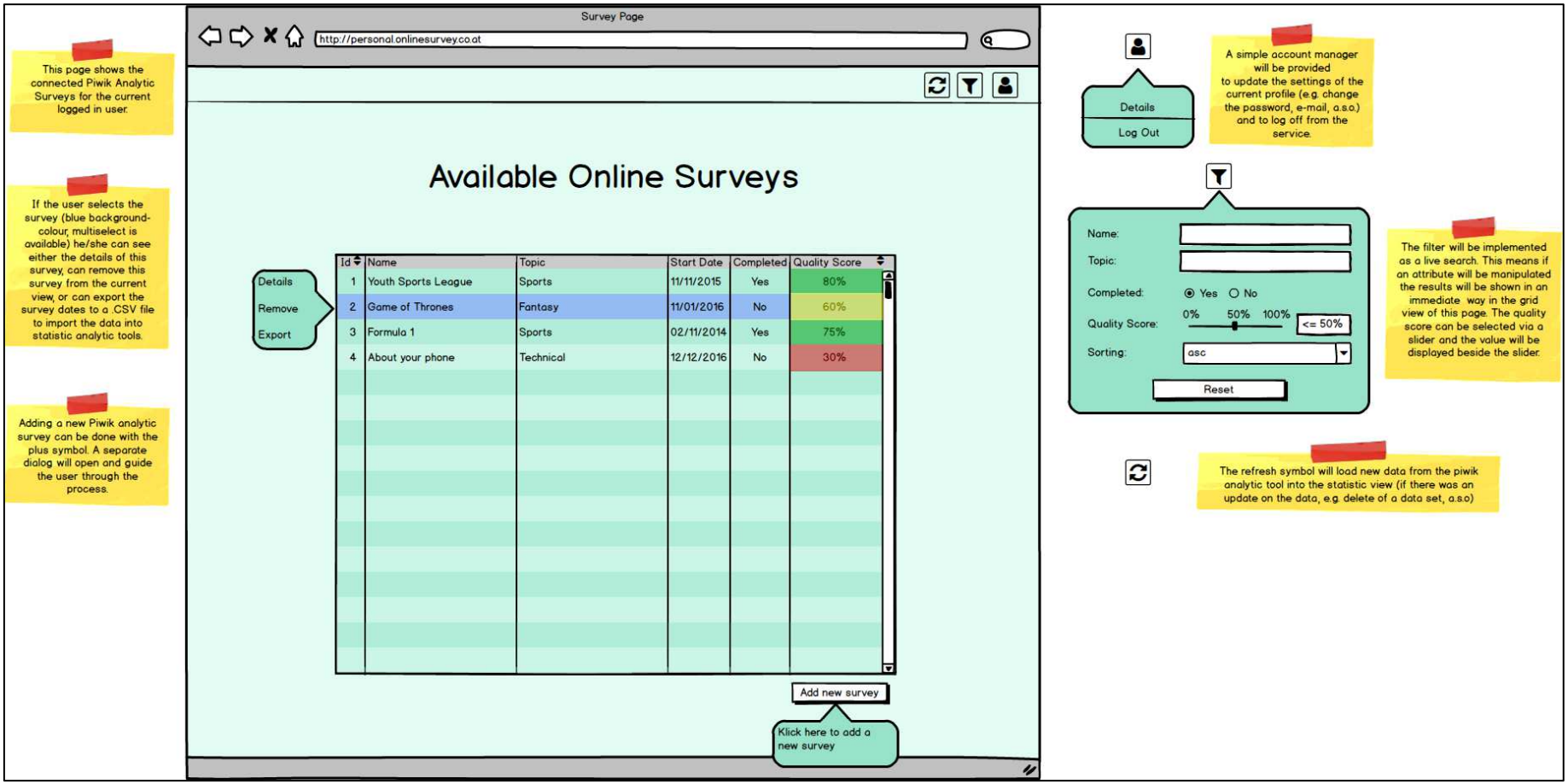
Literature: SQT: A tool for the automated measurement of respondent behaviour and response quality in health-related gamified online surveys. Wimmer, C., Biegler, S., Harms, J., Kappel, K., & Grechenig, T. (2018, May).			
Title	**Author**	**Status**	**Remark**
The short-term campaign panel of the german longitudinal election study 2009: Design, implementation, data preparation, and archiving; version 5.0. 0.	(Steinbrecher, Roßmann, & Bergmann, 2013)	Included	Computation of the careless response pattern speeding.
Insufficient effort responding: Examining an insidious confound in survey data.	(Huang, Liu, & Bowling, 2015)	Included	Defining careless responses and respondent burden.
Speeding in web surveys: The tendency to answer very fast and its association with straightlining.	(Zhang & Conrad, 2014)	Included	Definition of speeding and straightlining.
Literature: Social Interaction and Internet-Based Surveys: Examining the Effects of Virtual and In-Person Proctors on Careless Response. Francavilla, N. M., Meade, A. W., & Young, A. L. (2018).			
Title	**Author**	**Status**	**Remark**
Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality.	(Bowling, et al., 2016)	Included	Groundwork about careless responding.
Applying social psychology to prevent careless responding during online surveys.	(Ward & Meade, 2018)	Included	Definition of different markers for careless responses.
Identifying careless responses in survey data.	(Meade & Craig, 2012)	Included	Definition of careless responses.
Literature: Assessing the Effects of Survey Instructions and Physical Attractiveness on Careless Responding in Online Surveys. Rauti, C. M. (2017).			
Title	**Author**	**Status**	**Remark**
Predictors of inconsistent responding in web surveys.	(Akbulut, 2015)	Included	Definition of satisficing.
Using the theory of satisficing to evaluate the quality of survey data.	(Barge & Gehlbach, 2012)	Included	More information about satisficing and some examples of negative respondent behaviours.

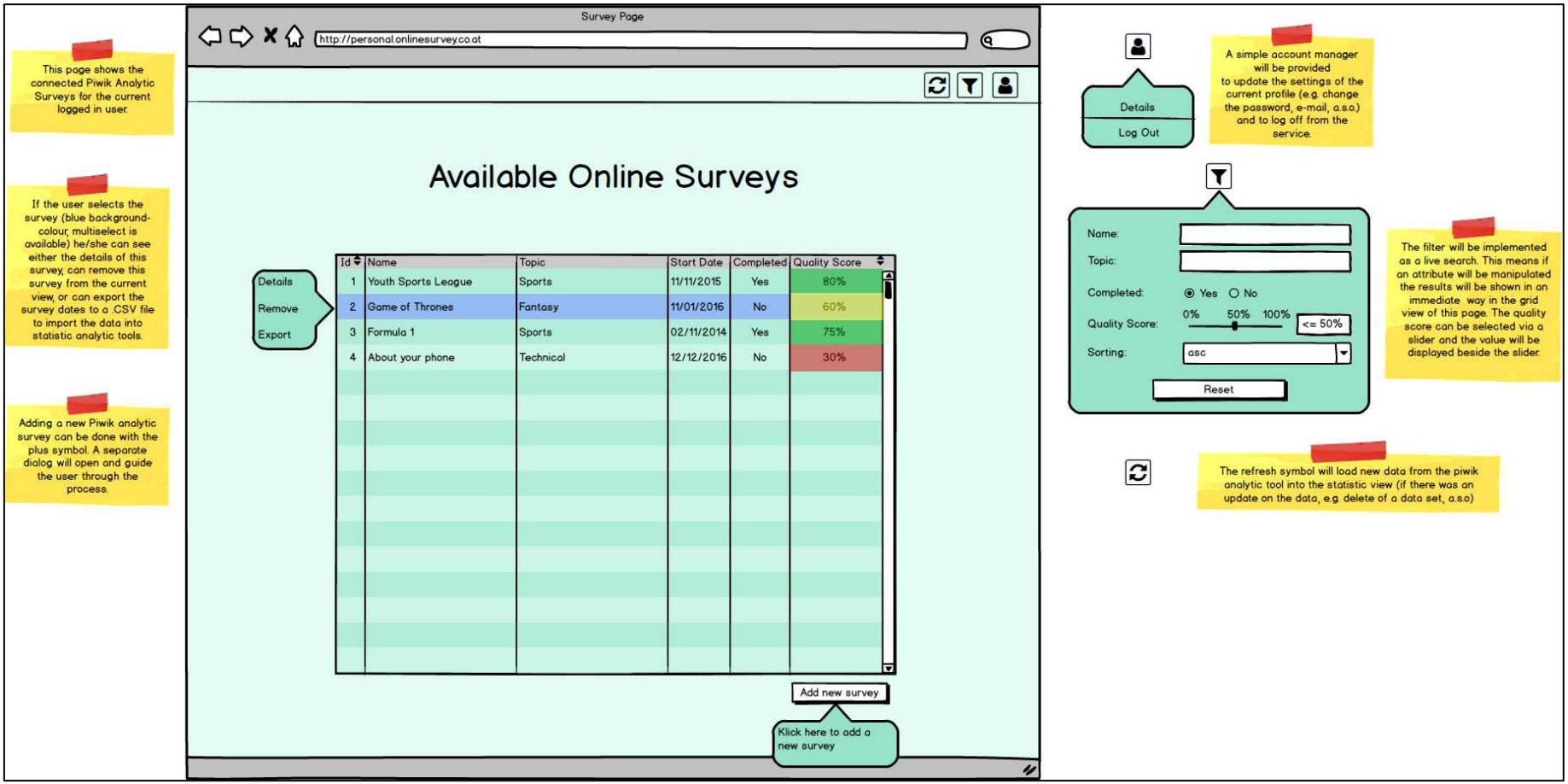
Best practice recommendations for data screening.	(DeSimone, Harms, & DeSimone, 2015)	Included	Data screening methods to detect careless responses.
Anchoring and Adjusting in Questionnaire Responses.	(Gehlbach & Barge, 2012)	Included	Definition of anchoring and measurement strategy.
Response strategies for coping with cognitive demands of attitude measures in surveys.	(Krosnick, 1991)	Included	Background information about satisficing.
Caring about carelessness: Participant inattention and its effects on research.	(Maniaci & Rogge, 2014)	Included	Definition of data quality and careless responses.
Evidence for response bias as source of error variance in applied assessment.	(McGrath, Mitchell, & Hough, 2010)	Included	Definition of the different types of responder and why there exists a respond bias.
Beyond 'trapping' the undesirable panelist: The use of red herrings to reduce satisficing.	(Miller & Baker-Prewitt, 2009)	Included	Explaining satisficing.
The Influence of the Design of Web Survey Questionnaires on the Quality of Responses.	(Ganassali, 2008)	Included	Different definitions about negative response behaviour.

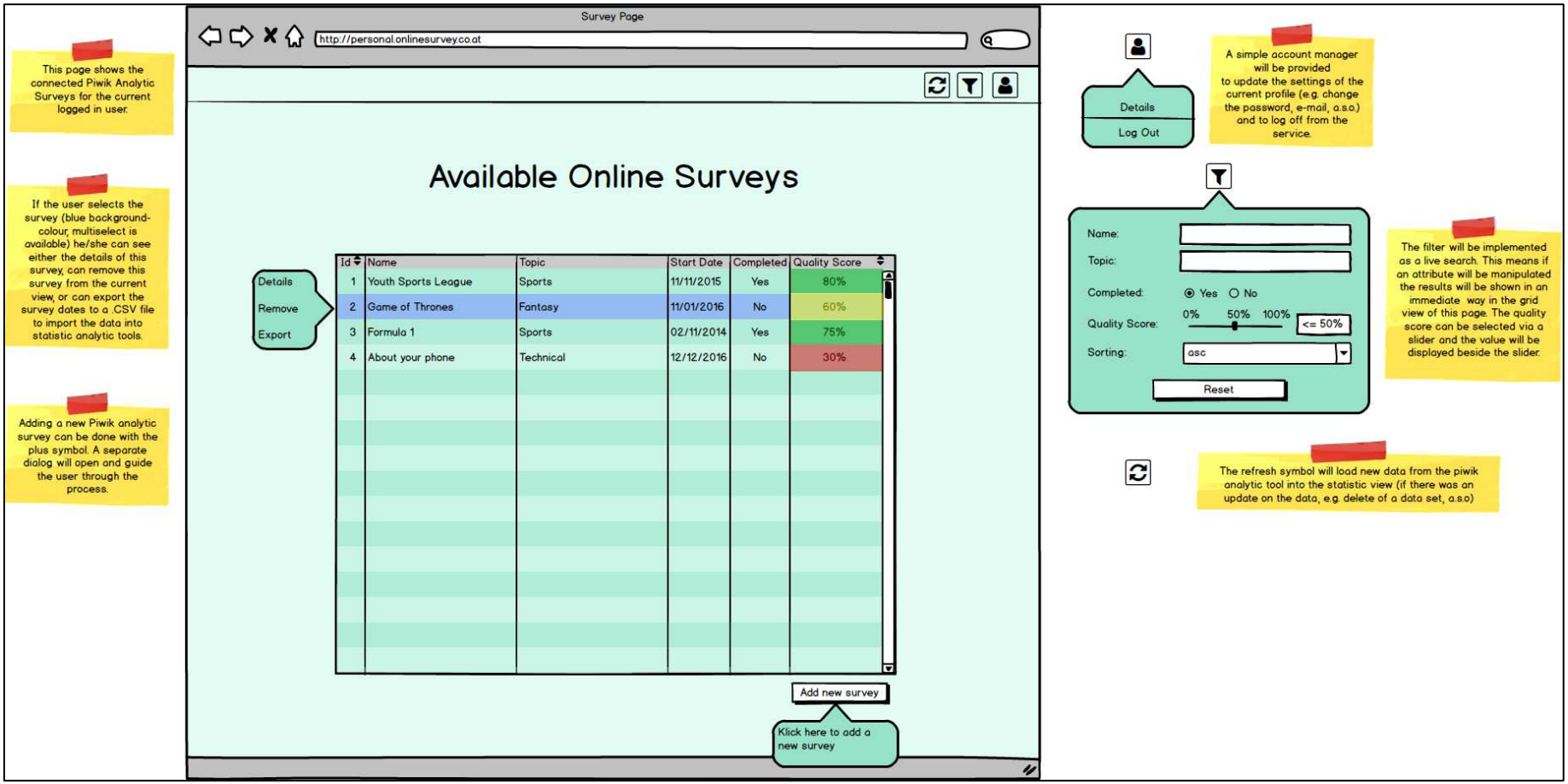
Appendix II

Appendix II presents the mockup GUI prototype.



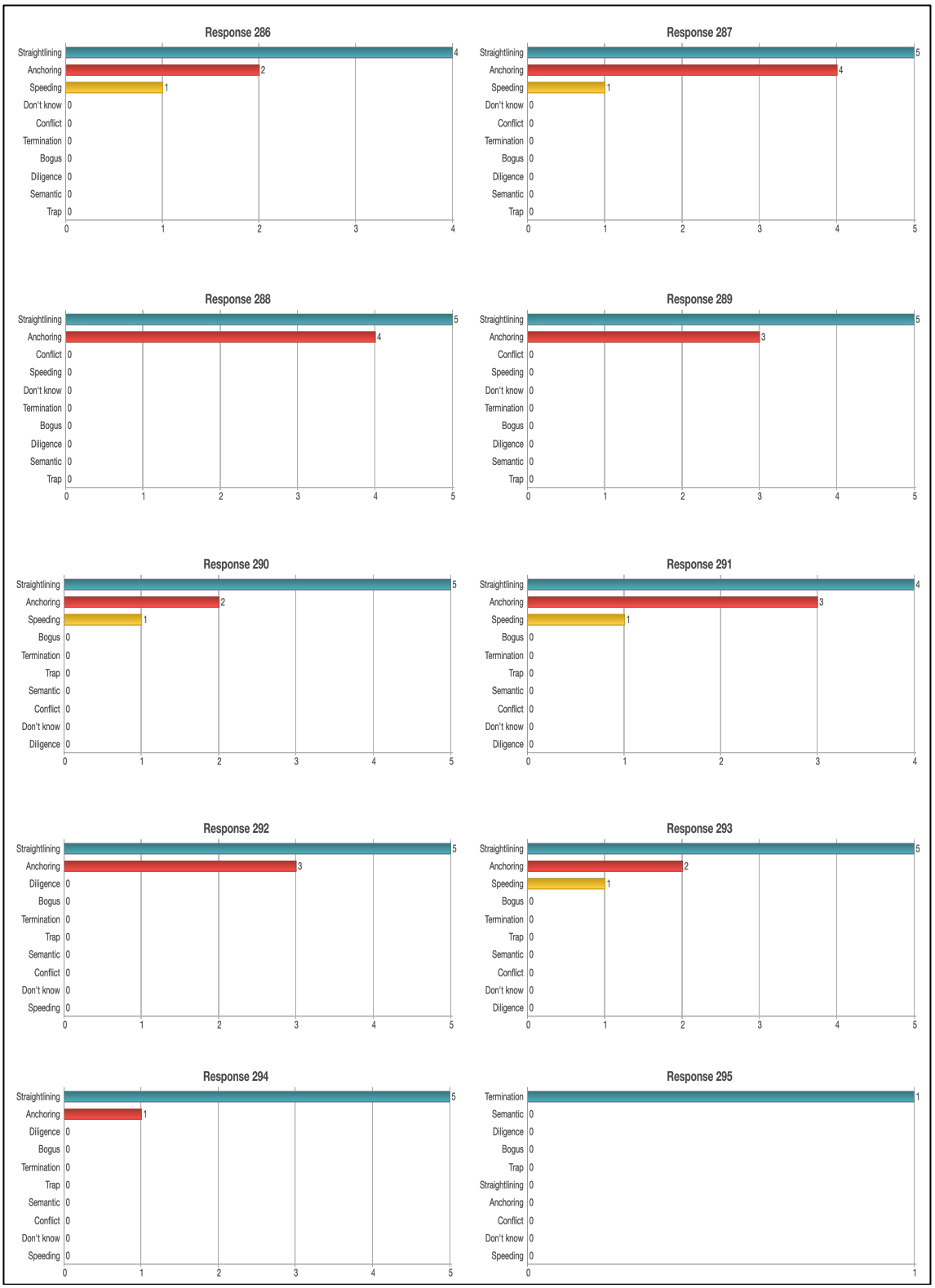


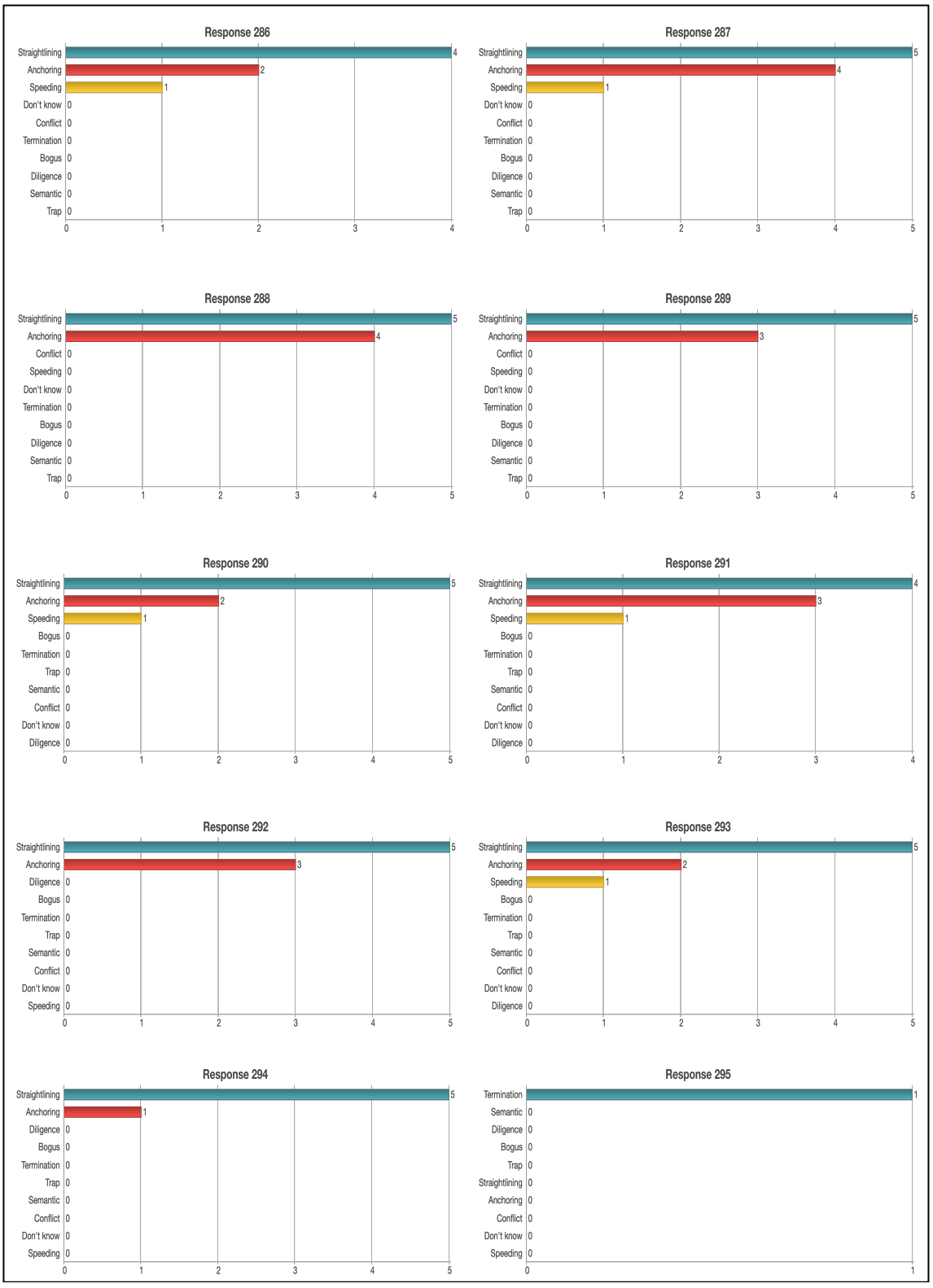


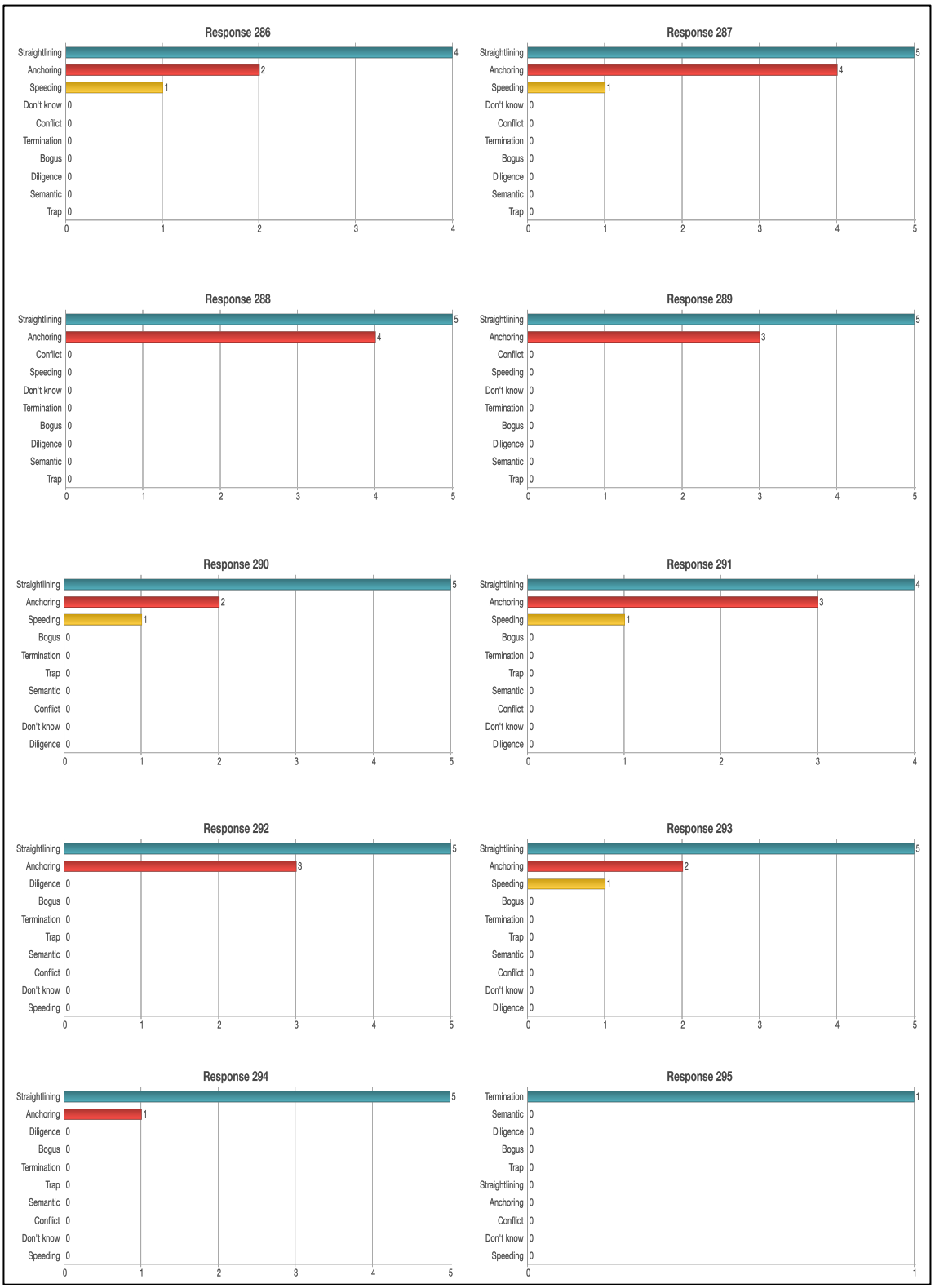


Appendix III

Appendix III presents the manual judgement for each careless response pattern from a survey from a previous scientific work (HARMS, ET AL., 2015). The manual judgment will be compared with SQT's output.







Appendix IV

Appendix IV presents the interview of Maximilian Störchle about the usefulness and practicability of SQT including the SUS questionnaire to rate the usability.

General Questions

Name of the respondent.
Maximilian Störchle.

Skills of the respondent.
Student of TU Wien, in the area of software engineering.

In which context the respondent used SQT?
Maximilian used SQT in the context of quality control for his own survey for the master thesis. SQT was used to detect speeding, straightlining and termination. Normally SQT is used as a host platform, Maximilian used SQT as a stand-alone application with the help of a virtual container (Docker). Therefore, Maximilian didn't used the user administration of SQT.

Positive things about SQT?
Maximilian highlighted that the import of respondent data of his survey into SQT was very easy and simple. This was the case because SQT uses JSON as a data format for the import of respondent data. Therefore, Maximilian could use any survey analysis tool to track the different user actions and therefore import it via JSON into SQT. Another positive point of SQT was the representation of the different careless response data in the Web GUI of SQT.

Negative things about SQT?
Because Maximilian didn't use the host version of SQT it was not so easy to install SQT as a local version. In the end Maximilian used a virtual container application (docker) to solve this situation. In addition to the table view of the careless response patterns a graphical view would be nice.

Was SQT useful for your study and why?

According to Maximilian SQT was very useful for the careless response pattern (speeding, straightlining, termination) he needed for his master thesis. He used the output of SQT (detection of negative respondent behaviour for one response) to deselect low-quality responses.

Did SQT fulfil your expectations?
Yes, it fulfilled all the expectations.

Is SQT a useful tool for real-world online surveys?
According to Maximilian and his survey SQT is a useful tool in real-world surveys and not only for research topics. Maximilian can imagine that SQT is also useful for big commercial companies which use online survey for product feedback.

Usability Questions

The respondent thinks that he would like to use SQT frequently.

Strongly disagree				Strongly agree
				X
0	1	2	3	4

The respondent found the system unnecessarily complex.

Strongly disagree				Strongly agree
	X			
0	1	2	3	4

The respondent thought the system was easy to use.

Strongly disagree				Strongly agree
			X	
0	1	2	3	4

The respondent thinks that he would need the support of a technical person to be able to use SQT.

Strongly disagree				Strongly agree
	X			
0	1	2	3	4

The respondent found the various functions in SQT were well integrated.

Strongly disagree				Strongly agree
				X
0	1	2	3	4

The respondent thought there was too much inconsistency in SQT.

Strongly disagree				Strongly agree
X				
0	1	2	3	4

The respondent would imagine that most people would learn to use SQT very quickly.

Strongly disagree				Strongly agree
		X		
0	1	2	3	4

The respondent found the system very cumbersome to use.

Strongly disagree				Strongly agree
	X			
0	1	2	3	4

The respondent felt very confident using the system.

Strongly disagree				Strongly agree
				X
0	1	2	3	4

The respondent needed to learn a lot of things before he could get going with SQT.

Strongly disagree				Strongly agree
		X		
0	1	2	3	4