Department of Statistics

The London School of Economics and Political Science

Detecting Semi-plausible Response Patterns

by

Tayfun Terzi

A thesis presented for the degree of

Doctor of Philosophy

in the subject of

Statistics

25 May 2017

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of **60,246** words.

Acknowledgements

First and foremost, I would like to thank my supervisors Professor Chris Skinner CBE and Dr Jouni Kuha for offering invaluable advice, their guidance, their contribution of time, as well as the space and freedom I was granted and entrusted with. It was Chris' inspiring intuition and expertise, his very contagious calmness, and his belief in me that gave me the energy to excel in my work. I feel honoured that I was given the chance to prove myself under the supervision of one of the most eminent and distinguished experts in the field. It was Jouni's intellectual brilliance and knowledge (in fact, I personally will never stop believing that he knows everything, even though he would humbly deny that statement), his patience, and his unprecedented time commitment as my second supervisor that made me tirelessly challenge but not doubt myself.

Second, I would like to extend my sincerest thanks to Dr Jason Huang and Professor John A. Johnson for allowing me to use their questionnaire data. The datasets were of paramount importance in evaluating the findings of this thesis and provided an excellent platform for the formation of ideas.

Third, I gratefully acknowledge the Economics and Social Research Council and the London School of Economics and Political Science for providing the funding which allowed me to undertake this research. Particularly, I would like to thank the administration staff at the Department of Statistics for not only just doing their jobs but far more for their almost parental dedication towards improving their PhD students' experiences at LSE.

Fourth, I would like to express my gratitudes to Professor Irini Moustaki and Professor Joop Hox who are honouring me by acting as my examiners in my viva voce examination. There is no greater reward for a finishing PhD student than to discuss our lifetime achievement with researchers to whom we look up to and whose opinion we value and respect.

Fifth, my time at the LSE was made exceptionally enjoyable in large part due to my fellow PhD students. The alternately productive and unproductive chats and

beautiful moments we have shared in the second-floor office of the Columbia House are unforgettable memories and moments, which I will dearly miss.

Lastly, I would like to thank my family for always believing in me and supporting me in all my pursuits. My deepest gratitude goes to my partner in life for his unconditional love and for being my absolute source of happiness.

Abstract

New challenges concerning bias from measurement error have arisen due to the increasing use of paid participants: semi-plausible response patterns (SpRPs). SpRPs result when participants only superficially process the information of (online) experiments/questionnaires and attempt only to respond in a plausible way. This is due to the fact that participants who are paid are generally motivated by fast cash, and try to efficiently overcome objective plausibility checks and process other items only superficially, if at all. Thus, those participants produce not only useless but detrimental data, because they attempt to conceal their malpractice. The potential consequences are biased estimation and misleading statistical inference.

The statistical nature of specific invalid response strategies and applications are discussed, effectually deriving a meta-theory of response strategy, process, and plausibility. A new test measure to detect SpRPs was developed to accommodate data of survey type, without the need of a priori implemented mechanisms. Under a latent class latent variable framework, the effectiveness of the test measure was empirically and theoretically evaluated. The empirical evaluation is based on an experimental and online questionnaire study. These studies operate under a very well established psychological framework on five stable personality traits. The measure was theoretically evaluated through simulations. It was concluded that the measure is successfully discriminating between valid responders and invalid responders under certain conditions. Indicators for optimal settings of high discriminatory power were identified and limitations discussed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the recent rise of new possibilities in research regarding a new form of online recruiting where people are paid to act as participants, the challenge for eliminating response biases through statistical analysis has become more important. Researchers anticipate that those tools which rely on online recruited and paid participants will soon become an important tool for research in many social disciplines (e.g., Mason and Watts, 2009; Buhrmester, Kwang, and Gosling, 2011). Hence, there is an increased need for the evaluation of implications associated with the usage of paid participant pools.

I argue that one of the worrying implications are *Semi-plausible Response Patterns* (SpRPs), which result when participants superficially process the information of (online) experiments or questionnaires and try only to respond in a plausible way. This is due to the fact that participants who are paid are generally interested in earning fast money and efficiently attempt to overcome objective plausibility checks, and process all other items only superficially, if they process them at all. Thus, those people produce not only useless but detrimental data, because they attempt to conceal their malpractice from the ”employer”, or rather the test administrator. The consequences of this new increase in measurement error are biased estimation and blurred or even covered true effect sizes, which contaminate valid models (cf. Mavridis and Moustaki, 2008, 2009).

Huang, Curran, Keeney, Poposki, and DeShon argue that online as well as traditional paper-and-pencil surveys are particularly susceptible to this kind of 'subtle yet insidious threat to data quality [...]' (2012, p. 100). While Huang et al. (2012) use the label *insufficient effort responding* (IER), others refer to this type of responding as random (Beach, 1989; Berry et al., 1992), *careless* and *inattentive* (e.g., Curran, Kotrba, and Denison, 2010; Meade and Craig, 2012), or inconsistent responding (McGrath, Mitchell, Kim, and Hough, 2010), also terms like *content nonresponsivity* (Nichols, Greene, and Schmolck, 1989), *protocol invalidity* (Johnson, 2005) and *speeders* (Greszki, Meyer, and Schoen, 2014) are often used. The foci lie on different causes (e.g., lack of motivation) and contexts (e.g., online vs. paper-and-pencil survey) leading to different labels of types of responding. In spite of different labels, these situations are also leading to fundamentally similar response patterns. Many of these constructs overlap in the idea that participants respond without (any) regard to item content (see Section 1.2 for a detailed distinction). Identifying those response patterns could help to improve the criterion-related validity of measure (McGrath et al., 2010). Couch and Keniston (1960) already stated that these kinds of undesired response patterns should either be treated as outliers (e.g., by controlling for them or removing them from the sample) or they should be seen as manifestations of participants' characteristics.

SpRPs are particularly characterised by the idea that participants do not respond entirely without regard to item content, but rather try to respond in a way that the researcher will not easily detect. This kind of semi-plausible responding is what renders SpRPs a special and more severe version of invalid protocols.

A researcher is confronted with three questions:

- How to prevent or at least minimise SpRPs?
- How to recognise those SpRPs?
- How to deal with biased estimates of any quantity due to SpRPs?

While the first question might be answered by drawing on psychological, empirically-based research, the second and specifically the last question needs to be dealt with on a statistical methodological level. Unfortunately, commonly used techniques for the identification of similar kinds of invalid protocols typically entail only cursory data screening methods. These include, for instance, univariate outlier analysis. However, these methods are only effective given the assumption is met that careless or inattentive responses are rare or extreme in magnitude. Unfortunately, semi-plausible responses are by definition not easily identifiable as merely rare or extreme responses without taking a broader context into account. Other efforts towards capturing measurement error and increasing the reliability of measurements are often very effective, but rarely applicable for systematic measurement error or measurement error that is not produced by everyone in the sample but rather a small group of people. Thus, in the case of SpRPs, comparing response patterns as a whole with plausible response patterns could help with classifying them as valid or invalid protocols. Starting points are procedures such as *person-fit indices* (Meijer and Sijtsma, 2001) which identify the extent to which a response pattern deviates from the latent model. Measures in areas dealing with non-response bias and missing data (Allison, 2009) can also be drawn upon for the treatment of SpRPs.

In this thesis, I will identify primary sources of SpRPs, discuss their consequences and establish a theoretical framework linked to other already well-established research areas. This will be followed by a literature review of available identification indices developed for different kinds of undesired response patterns. By drawing upon an experimental dataset as well as largely implemented data on an empirically well-investigated framework of the *Big Five* personality factors, I will examine statistical properties of SpRPs. Ultimately, this will help to establish a statistical theory of SpRPs with the focus on latent variable models in order to develop and evaluate an optimal, universally applicable *identification measure*. The thesis will conclude with a brief discussion about attempts to deal with SpRPs, once they are identified.

1.1 Primary Source: Micro-Jobbers

In this section, I will set out the primary source of SpRPs. Although the potential for generalisability of results to other sources will be discussed throughout the thesis, the reader should be aware that *micro-jobbers* introduced in this section serve as the group of focus.

A modern form of data collection for psychology and other social sciences is the use of so-called *micro-jobber platforms* like Amazon's *Mechanical Turk* (MTurk). Usually, those online platforms enable scientists as well as market researchers of every kind to create a task, recruit labour, and financially compensate them for providing data. In 2007, Pontin already reported 100,000 of available micro-jobbers from over

100 countries. This is a diverse potential participants pool available for many kinds of surveys or experiments. However as the term labour already implies, money plays the central role for this work force's motivation to attend as participants. It is essential that the labour commissioner can refuse payment if the work is not done properly, e.g. the participant has not completed the questionnaire in a way a researcher has expected him or her to do. Furthermore, these monetary compensations are typically small. A review of MTurk (Mason and Suri, 2012) but also other reviews of general micro jobber platforms (Buhrmester et al., 2011) report only very small amounts of money such as five to ten cents (USD) for 5 to 10-minute tasks. Paolacci, Chandler, and Ipeirotis (2010) used those platforms to replicate classic studies at a cost of approximately $1.71 per hour per subject.

Another reason why a growing number of researchers make use of micro-jobber platforms besides the low cost is that it is supposed to reduce certain kinds of biases found in traditional samples (Gosling, Vazire, Srivastava, and John, 2004). It is argued that the samples in internet surveys consist of demographically more diverse participants than typical college samples (e.g., in the US, Germany and other countries) broadening the validity beyond undergraduate students (Eriksson and Simpson, 2010). Conducting experiments also appears to be time-saving, allowing for faster cycles with regards to continuously updating of methodology and theory (Mason and Suri, 2012). Concerning the validity of provided responses, Buhrmester et al. (2011) reports satisfying results in terms of psychometric standards based on participants recruited in this manner. However, one cannot deny that a sample consisting of micro-jobbers is a sample of individuals mainly seeking to earn money. It is hard to believe in data provided by Buhrmester et al. (2011) in his brief and very positive review of MTurk which reports participants to be internally motivated (e.g., for enjoyment).

Mason and Suri (2012) collected demographical data in MTurk from nearly 3,000 unique workers and reported 55% being female. Median and average age was reported to be 30 and, respectively, 32 years old with the majority of them earning U.S. $30,000 per annum. About 7% of these participants participated in two studies with only one worker who changed the answer on gender, age, education, and income. Hence, these demographics seem to be based on solid self-consistent measures. An interesting question is why people work as micro-jobbers in spite of the low wages and given their reported high income per annum. The most important driver for MTurk workers is reported to be the monetary outcome. Only 12% of U.S. worker report that MTurk money is irrelevant for them. However, Ipeirotis (2010, cited in Mason and Suri, 2012) states that the vast majority see MTurk also as a fruitful way to spend free time while earning some cash. Nonetheless, nearly 10 % also seem to scrape together a living using MTurk.

Caveats of using online questionnaires are diverse. One major disadvantage is that researchers often are required to deal with duplicates. This means that some individuals might complete an online questionnaire or experiment several times using multiple identities. Although this is partly controlled using browser cookies and tracking IP addresses, experienced users circumvent these and efficiently produce detrimental data. Another problem is the use of software programs or so-called *bots* that complete questionnaires.

Even more concerning, in terms of cursory detecting invalid responders, are individuals who attempt to make as much money as quickly as possible without regard to the instructions or intentions of the study. Mason and Suri (2012) and others refer to those participants as *spammers*. Spammers especially target surveys, since these are easy to complete. This is often done in a random but more predominantly in a semi-plausible manner, since *bogus items* are often implemented in these surveys to identify obviously implausible responders. Semi-plausible/undetectable response patterns are more popular as these less often lead to a refusal of payment, which would, in turn, lead to a bad reputation of this worker on platforms where these kinds of mechanisms are implemented. Furthermore, although the number of these kinds of workers might not be large, the data, and thus the participant entries they produce, are severely detrimental for the subsequent analysis of data.

I primarily focus on paid mass participants because the prevalence of invalid responders is expected to be most severe and invalid response strategies more successfully concealed in these scenarios. However, findings of this thesis are easily generalisable to other settings. Miller (2006) reported in a study of 13 US panels that about 5-10% of participants responded to obvious plausibility checks (also referred to as *red herring* questions) incorrectly, indicating the use of invalid response strategies. R. Smith and Brown (2005) reported 1% of participants in 20 extensive surveys using only the same answer option for all questions (also referred to as *straight-lining* or *long string* response strategy), thus, not even trying to hide their intentions (cited from Greszki et al., 2014). Meade and Craig (2012) used 11 different identification measures and concluded that 5-15% of participants in undergraduate internet surveys lack sufficient attention. Further analysis using factor mixture model analyses also indicated that around 10 to 12% of their undergraduate sample belonged to a *latent class* that can be considered *careless* in their responses, which is nearly identical to results reported by Kurtz and Parrish (2001). Woods (2006) has found that in certain (commonly encountered) scenarios it only requires 1% to 20% of careless responses in the sample for models not to fit the data anymore.

1.2 Semi-plausible Response Patterns

Having set out the primary source and the subject of this thesis, I will continue to discuss causes of SpRPs from a cognitive psychological perspective. In doing so, I seek to define the terminology used throughout the thesis at hand, and depict links to other constructs in the literature. The meta framework, illustrated in Figure 1.1, seeks to capture these links between concepts of response strategy, the cognitive processes involved and on the other hand the actual resulting plausibility of data. Links between response validity, involved cognitive processes, and data plausibility of shown (invalid) response strategies are conceptual examples.

In his prominent review of survey research, Krosnick (1999) states that there is wide agreement about the cognitive processes that result in valid response patterns (e.g., Cannell, Miller, and Oksenberg, 1981; Schwarz and Strack, 1985; Tourangeau and Rasinski, 1988). Kahn and Cannell (1957) discuss in detail the so-called *cognitive process model* based on the original work of Tourangeau (e.g., Tourangeau, Couper, and Conrad, 2000; Tourangeau, 1984, 1987). Valid respondents answer questions properly when they, first, read the entire question text to comprehend, interpreting the question and deducing its intent (P1). Secondly, valid response patterns require accessing relevant information in a participant's memory (P2). Thirdly, based on accessible information, a (single) subjective judgement is formed (P3). Lastly, participants then formulate or translate that judgment into a response, e.g. selecting an answer option based on offered alternatives (P4).

Hence, steps P1 to P4 all involve a great deal of cognitive work (e.g., Krosnick and

Fabrigar, 2001). We can assume that if that applies to a single question, it applies even more to a large number of observed variables. We can represent the thoroughness of their execution on individual scales as drawn in Figure 1.1. Psychology of survey participants considers several aspects that lead to this expenditure of cognitive effort (see Warwick and Lininger, 1975): Participants might be motivated by desires for self-expression, intellectual challenge, self-understanding, altruistic feelings or emotional catharsis. Krosnick (1999) further states that motives might include desires for gratification from successful performance to help the survey purpose (e.g., to help an administrator improve working conditions). These motives can be categorised under the psychological concept of intrinsic motivation (for a recent review about intrinsic versus extrinsic motivation, see Ryan and Deci, 2000). As the most mobilising form of motivation, it can easily be considered strong enough to facilitate a valid response. Nevertheless, we cannot always assume that participants are purely intrinsic but rather extrinsically motivated, for instance, driven through automatic compliance processes (e.g., Cialdini, 1993) or as students attempting to collect course credits. Sometimes motivation might even change throughout the course of the questionnaire when participants satisfy their desires to provide valid responses after answering a few questions, and become increasingly fatigued and distracted with each additional assessment. Unfortunately, usual sources of extrinsic motivation in surveys cannot be considered as secure paths for valid response patterns. Krosnick (1991) argues that participants might resolve this dilemma, which is a lack of intrinsic motivation and ineffective extrinsic motivation, by shifting to an invalid response strategy, compromising response standards and expending less energy.

The actual extent to which a response is valid can, apart from a binary classification, be located on a validity continuum as is shown in Figure 1.1. The valid anchor at the positive end of the continuum is often referred to as optimising (e.g., Krosnick, 1999). The actual position on the continuum then indicates the combined degrees of thoroughness executed throughout steps P1 to P4. A response strategy on the continuum not far from valid responses is *weak satisficing* (Simon, 1957). This occurs when responses are the result of the complete but less than fully diligent execution of P1 to P4. Hence, participants settle for a merely satisfactory rather than a thoroughly processed answer. Yet another approach on the continuum might be referred to as *strong satisficing* (borrowing the term from Krosnick, 1999). This invalid response

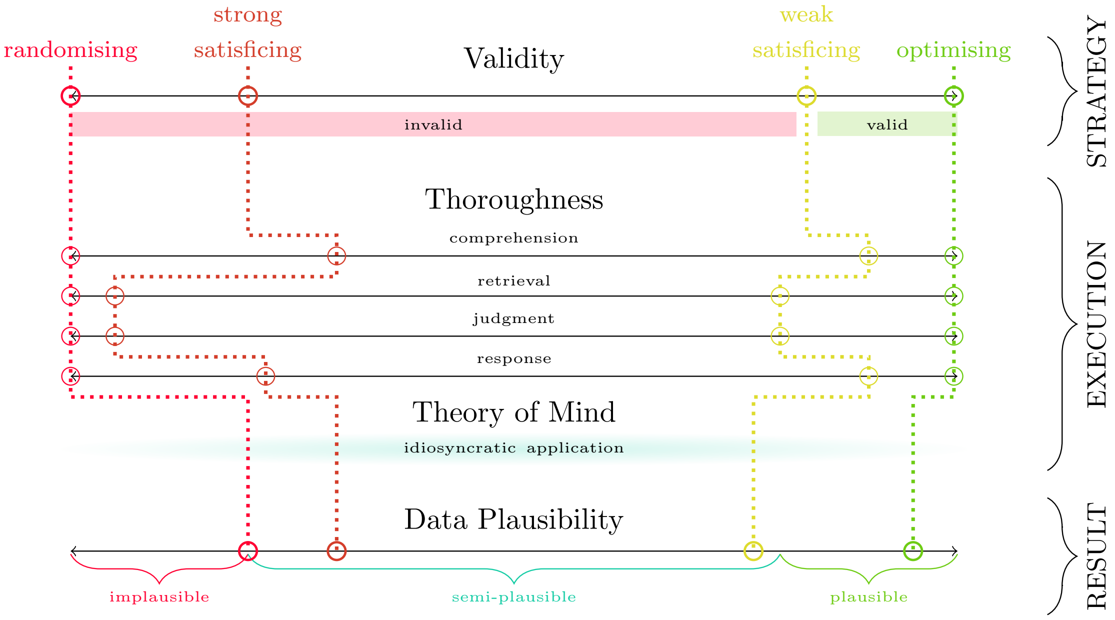


Figure 1.1: A meta-theory of response validity, involved processes, and resulting data plausibility for exemplary response strategies.

strategy leads to answers without going through steps P2 and P3 (retrieval and judgement) altogether. Hence, it means superficially interpreting a question without referring to any internal psychological cues and selecting from given alternatives that are subjectively judged as reasonable answers. As an invalid response strategy without relevance to events of interest, strong satisficing can be considered the most detrimental and most difficult to identify amongst invalid responses, given there is an uncountable number of idiosyncratic heuristics and cues (e.g., question wording) from which participants can choose. If participants further worry about detectability of their invalid response strategy and defensibility of their responses, they will choose safe answer options, such as occasional neutral points of a rating scale avoiding to take more risky stands, and back away from purely random answer selections. Lastly, the negative end of the continuum could be referred to as randomising which would only involve P4. Participants who randomise might try to give (random) answers, always pick the middle category (Schuman and Presser, 1996; Tourangeau, Couper, and Conrad, 2004) or exclusively select first answer option (Malhotra, 2009). As a side note, the reader should be aware that the term randomising should not be taken literally. Following Neuringer (1986, p. 63), humans can learn to behave randomly, but they do not have the natural ability to do so (for reviews Tune, 1964a, 1964b; Wagenaar, 1972). Hence, participants who try to respond in a random manner will produce correlated responses, following some idiosyncratic, systematic way of 'random' responses. This is an important aspect because otherwise statistically random responses (independent responses following a uniform distribution) can to a certain extent be incorporated as random measurement error using latent variable frameworks (Medsker, 1994; Shook, Ketchen, Hult, and Kacmar, 2004). However, even such purely random measurement error would certainly have a bad impact on estimation, if it is completely unrelated to the respondent's true position.

There are numerous methods we might employ with regards to questionnaire design or at data collection stage to minimise the occurrence of SpRPs. For example, we can reduce the perceived cost, such as perceived energy expenditure, by establishing an intrinsically motivating instruction. A balance between monotony and standardisation of question design is very important. A monotone question design can easily lead to boredom. However, if questions do not follow a minimal common standard, cognitive processes require more capacity to adapt to different question formats. In general, a large number of questions should be avoided. Questions should focus on easily accessible memory and allow for additional answer options, such as a 'don't know' answer option. There is a vast and rich range of literature on how to improve data quality, integrity, and response rate, simultaneously. As this thesis primarily seeks to develop and discuss measures to deal with SpRPs after data collection, I would like to refer the reader at this point to standard textbooks on survey methodology and online questionnaires (e.g., Leeuw, Hox, and Dillman, 2008).

To capture further cognitive processes involved in saving energy expenditure in any of the steps P1 to P4, where participants worry about the defensibility of responses, I will borrow from research based on the *theory of mind*. Frith and Frith states that '[t]hrough having a theory of mind we can recognise that another person's knowledge is different from our own' (2005). The theory of mind has found attention especially in psychological areas of developmental psychology trying to assess when human beings start to understand that other people have different cognition, attitudes, emotions and, hence, perception and behaviour separate from ourselves (e.g., Baron-Cohen, 1991). Different processes either developed through social interaction or inferred from introspection enable healthy subjects to develop a theory of mind about the actual intentions of a question. In Figure 1.1, the theory of mind is represented as an additional cognitive variable that reduces validity but, in contrast, might increase data plausibility. Participants can employ an invalid response strategy based upon less cognitively exhausting question-cue-stimulus-response rules and yet produce response patterns that seem plausible from a quantitative data point of view.

Therefore, I introduce the concept of semi-plausible response patterns in order to emphasise that the investigative nature of this thesis focuses on the produced data. Semi-plausible and implausible response patterns are defined through their statistical nature in reference to the valid response model. This is in contrast to the processes and causes of invalid response strategies, which may or may not result in semi-plausible response patterns. For instance in Figure 1.1, we can see that weak and strong satisficing or even randomising can lead to semi-plausible response patterns. As another example, a long string response strategy can depending on features of the valid response model, create an easily detectable implausible response pattern. However, straight-lining might as well result in a semi-plausible response pattern where it is unclear whether the data is based on valid responses. Consequently, the notion of *semi-plausible* response patterns also indicates the difficulty in detecting invalid responses.

The goal is to identify invalid response patterns as a whole rather than matching participants to their chosen response strategies and the cognitive processes with that strategy involved. The concepts of response strategies and cognitive processes are primarily important for understanding the causes of invalid responses in order to research methods for prevention or case-customised detection mechanisms. Hence, in distinction to research about the psychological causes and mechanism involved that lead to invalid response patterns, this thesis will focus on methods for the assessment of, predominantly quantitative, plausibility of the resulting response patterns.

1.3 Thesis Outline

In the previous sections, I motivated and introduced SpRPs as a new problem and, hence, not exhaustively researched topic. I further defined SpRPs as the construct of interest within a framework of existing literature.

In this thesis, I seek to address problems arising from SpRPs and research on solutions to deal with them. In general, I propose two methods to deal with SpRPs, namely, accommodating semi-plausible response strategies into the statistical model and/or using identification measures for the detection of invalid responses to exclude them from further analyses. Methods developed within this thesis will be evaluated on two empirical datasets; one being an experimental study and the other an online questionnaire study. Furthermore, a large-scale simulation study will be conducted such that we can identify relevant information to the prediction of success scenarios in separating valid from invalid responders.

In Chapter 2, detection methods from existing literature will be introduced and reviewed. I will discuss identification measure from other relevant fields such as cheating and fraud detection. This knowledge will serve as an example for the development of an appropriate test measure for the detection of SpRPs. Furthermore, I will draw on *latent class analysis* as a statistical tool for the accommodation of invalid response strategies. Previous implementations of latent class models in related studies will serve as examples for the definition of an appropriate framework to reduce measurement error from SpRPs.

In Chapter 3, I will introduce the empirical studies and analyse them using the traditional latent variable analysis approach without accounting for SpRPs in the sample. Both studies employ the same personality assessment instrument. Hence, under the assumption of measurement equivalence between valid responders of both samples, parameter estimates will be compared between datasets and experimentally induced groups. Here, the online questionnaire study sample is assumed to produce reasonable estimates of the valid response model. I will compare experimentally induced conditions of plausible versus semi/-implausible responding behaviour based on the traditional analysis model. These contrasts have the purpose of investigating the magnitude of estimation bias caused by SpRPs. Furthermore, I seek to assess the statistical nature of SpRPs. The resulting latent variable structure is of particular interest.

In Chapter 4, the traditional latent variable analysis model will be extended to accommodate a latent class for invalid responders. I will propose and analyse an example latent class model that accounts for one type of possible invalid response strategies. This method will be evaluated in three ways. First, model fit test statistics and indices will be compared between the latent class analysis and traditional latent variable analysis results. Secondly, individual parameter estimates are going to be assessed based on whether accommodating an invalid response strategy helps to reduce measurement error and estimation bias. Lastly, based on the posterior distribution of the latent class variable, response patterns will be assigned to either class to evaluate the success of correctly assigning plausible response patterns to the valid response class and semi/-implausible response patterns to the invalid response class.

In Chapter 5, I will motivate, derive, and discuss a new test statistic for the identification of SpRPs. The new measure is a modified version of an existing identification measure and will be interpreted within the framework of latent variable models followed by a discussion of possible application methods. The modification will be further motivated by comparing its performance in detecting semi/-implausible responders in the experimental study sample to the performance of its original version. Concluding the effectiveness of the new test measure, I will derive its theoretical distribution in order to estimate appropriate cut-off values for the separation of valid from invalid response patterns.

Insights gained from the latent class analysis results will be used in Chapter 6 to evaluate the new measure. Several methods introduced in the previous chapter will be applied in a numerical example. First, measurement models for the valid and invalid response classes will serve as known population models to gain a deeper understanding of the new measure as detection instrument. Empirical results for the generated data shall act as validation for the theoretically derived statistical properties of the new measure. Special focus lies on the estimation of cut-off values. Secondly, I will evaluate whether a combined approach towards detection can improve the discriminatory power regarding the experimental study sub-samples. Here, I adapt the more accurate valid response model parameter estimates derived via latent class analysis as the information source for the new identification measure.

In Chapter 7, I undertake a simulation study designed to identify variables that define situations of high and low success in detecting SpRPs. For this purpose, I define a general set of valid response behaviours, and two types of invalid response strategies in the latent variable framework. One type of response strategy is inspired by the empirical latent class analysis measurement model results. Throughout the simulation study, I simulate and alternate numerous attributes of typical empirical study settings, such as sample size, the number of observed/latent variables, and inter-dependence of latent variables. Results of each condition are based on 100 replications following a *Monte Carlo* simulation design. I identify relevant variables which define the success of detecting SpRPs. Ultimately, information collected through the simulation study serve as arguments for the development of guidelines and detailed recommendations for the application of the new identification measure.

The thesis concludes with a summary of outcomes and a global discussion on different approaches towards reducing measurement error and bias from SpRPs. Furthermore, I discuss suggestions for further research on and implications through the use of the new test measure in detecting SpRPs.

Chapter 2

Review of Methods for Detection

In Chapter 1, I described problems that arise through SpRPs and their emergence in online studies because these are increasingly relying on paid micro-jobbers as participants. Main concepts of and causes for SpRPs were discussed, where the emphasis was given to the importance of the plausibility aspect of response patterns for this thesis and, hence, their statistical quantitative nature rather than their qualitative examination. I outlined the thesis structure and pointed out the two primary methods aimed to deal with resulting measurement error and estimation bias: Accommodating invalid response strategies into models for statistical analyses through mixture designs and identifying invalid response patterns such that these can be excluded from further analyses.

In this chapter, I will lay the foundation for my work on the proposed research topic. This requires the introduction of statistical frameworks such as latent variable and latent class models as well as a thorough literature review in this and related fields. Fortunately, research on identifying specific kinds of response patterns received great attention from diverse subject communities, i.e. social and behavioural sciences. Previous literature on identifying other response behaviour such as cheating or malingering in clinical diagnostics is often as unique as the corresponding problem scenarios. However, findings in related areas can serve as very beneficial information sources, especially for the development of test statistics aimed towards a more general definition of undesired response pattern; as is the case for SpRPs.

Following the review on identification measures, I will draw on previous research that focuses on accommodating undesired response patterns into the statistical model. In doing so, we can separate measurement error based on specific kinds of response behaviour from the valid response model. The objective is to ensure that parameters defining the valid response model are not affected by estimation bias once invalid responses are accounted for by the model. We rarely observe or have any indication on whether a sample point is valid or invalid. Consequently, we need to rely on so-called latent class models where group/class membership is latent but not observed. Unfortunately, research in this field does not catch as much attention as work on identification measures due to its nature: Latent class analyses need to be adapted uniquely to each individual study setting and require sophisticated statistical as well as computational knowledge. Mixture designs such as latent class models are very error prone because these are not, except for some limited cases, supported by established software implementations. Furthermore, based on the complexity of latent class models at hand, the statistical implementation requires diligent perusal of numerous problem scenarios, such as how to deal with local maxima in the estimation process. Knowledge in this field remains mostly in the form of journal articles, aimed for a technical rather than applied audience. Hence, in the last section of this chapter, I will focus on research that is few in number but outstanding in quality as an introduction to methods used in this thesis.

2.1 Identification Measures

In the following, I will review the most important and (more or less) established methods for the identification of generally undesired response patterns.

Identification measures discussed in this review are chosen as potential tools to identify undesired participants or to be more specific undesired response patterns. Many of those measures have proven useful for identifying other kinds of response pattern (e.g., random responders, *social desirable* responders). The goal is to evaluate the use of existing identification measures to identify SpRPs and to extract knowledge in order to develop new identification measures specifically tailored to the detection of SpRPs.

Meade and Craig (2012) differentiate between two types of identification measures. The first of them are implemented a priori, i.e. before collecting survey data.

These draw on items or scales which are designed for the very purpose of detecting respondents with a specific response pattern. Among those are scales that are assessing social desirability (e.g., Paulhus, 2002), self-reported response effort (e.g., Student Opinion Scale, SOS, Sundre, 1999; Wolf and Smith, 1995; cited in Wise and Kong, 2005) and *lie scales* (e.g., MMPI-2 Lie scale). Other possibilities include nonsensical or so-called *bogus*/red herring items (e.g., Beach, 1989; Berinsky, Margolis, and Sances, 2014; Miller, 2006; Miller, Officer, and Baker-Prewitt, 2009), special scales designed to assess consistent responding (e.g., the MMPI-2 VRIN and TRIN scales), and questions which explicitly instruct the participant how to respond to an item (e.g., 'To monitor quality, please respond with a three for this item'). Those are integrated in the survey prior to administration as well as any self-report measures of response quality usually placed at the end of a survey (for discussion on self-report measures, see Wise and Kong, 2005).

Although such identification measures could be very useful for the very purpose of identifying semi-plausible responders, such measures are not available and implemented in the majority of existing surveys. Therefore, this thesis is focusing on the second kind of identification measures, namely, *post hoc* identification measures. Nevertheless, I will draw on surveys which also include identification measures that are implemented prior to the assessment in order to validate proposed and existing post hoc identification measures.

Post hoc methods can be applied to a broad range of surveys which did not a priori integrate specialised items. By drawing on several indices that are computed post hoc, the data can be screened for specific response patterns. Post hoc measures are either based on actual responses (response-driven measures) or on data which is acquired simultaneously to the survey process itself (e.g., response time per item, para-data measures). Response-driven and *para-data* measures that could potentially be useful for the identification of semi-plausible responders shall be discussed in following sections.

2.1.1 Measures for (theory-driven) Outlier Detection

There are a lot of generic ways of detecting outliers (Hodge and Austin, 2004). Outliers can be the result of several unexpected aspects of collecting data, for instance valid but extreme manifestations of the construct of interest or answers caused by poor survey design leading to misinterpretations. However, outlier measures may also provide means for the detection of response patterns that are the result of invalid response strategies. Here, we often have the choice of classifying certain sample members by purely data-driven procedures or feed further theory into the decision making process (e.g., distributional assumptions). In the following, some procedures shall briefly be discussed.

Individual Consistency

Individual Consistency measures are based on the assumption that a set of observed variables should be internally consistent by design. That is, we assume perfect measurement of a single construct of interest (latent variable) via observed variables. Hence, we can simply compute composite values of sub-scales (e.g., the sum of observed variables). In other words, a specific set or subset of observed variables that seek to measure the same general construct (latent variable) is supposed to produce similar scores.

The simplest of individual consistency measures is the *Longest String* measure. For each of the available answer categories, we compute the longest successive occurrence of that category. The reason for not only computing a single long string score for only the middle category is that response time, which is another major indicator of response validity, is usually negatively correlated with other additional answer categories, e.g. 'no answer' or 'don't know' (Greszki et al., 2014). Furthermore, some long string answer strategies often involve an idiosyncratic tendency to favour one or more answer options over others. Hence, it is sensible to assess the consecutive use of all available answer options. A scree test of sudden drops (Cattell, 1966) based on the frequency distribution of the values for one answer category can be used to determine cut-off values. Too long strings are considered as an inattentive use of the same response category. Studies on satisficing response strategy recommend a cut-off value of five or more consecutive choices of the middle category (Kaminska, McCutcheon, and Billiet, 2011; Krosnick, Narayan, and Smith, 1996). One disadvantage of the long string methods is that its effectiveness remains unstudied (Huang et al., 2012). Furthermore, long strings of identical responses might represent participants' substantive preferences (Kaminska et al., 2011; Krosnick et al., 1996).

Other identification methods basically separate information of a single response pattern in a certain manner in order to enable the computation of a within-person correlation. This is to provide a measure of consistent responding. Johnson (2005) proposes his individual reliability score which is derived from numbering the sequence of observed variables as they appear in a survey and dividing them into odd-numbered and even-numbered subsets. Each even and odd subset are used to accumulate scores of the respective observed variables. Finally, the correlation of the two half-scale scores is supposed to indicate a respondent's response pattern consistency. Furthermore, Johnson's individual reliability score can be corrected for decreased number of observed variables by the Spearman-Brown Formula. A high positive value indicates that the person is responding to inter-related items in a consistent way. Negative or small values indicate inconsistent response patterns. One major disadvantage of Jackson's even-odd score is that there needs be a reasonable number of observed variables.

Similar to Johnson's individual reliability score, the psychometric antonym/synonym measures produce pairs of observed variables to enable the computation of a correlation between responses of an individual. Depending on the actual procedure, we would expect a participant to respond in opposite directions between items that are, for instance, highly negatively correlated. The psychometric antonym procedure is derived from the so-called *semantic consistency indices* initially used by Goldberg and Kilkowski (1985). In general, first, antagonistic item pairs (e.g., highly negatively correlated items) are a priori or post hoc identified. Secondly, the within-person correlation between a participant's responses of these antagonistic item pairs is then computed for each participant. The aim is to examine the difference in two items that are highly similar in content (Meade and Craig, 2012). One way to define psychometric antonyms and synonyms is a priori drawing on dictionaries. However, this procedure is susceptible to subjective judgements and not feasible for universal use. Another way of assessing consistency via psychometric antonyms suggested by Goldberg (2000, cited in Johnson, 2005) is by identifying some number of unique pairs of observed variables with the highest negative correlations. Hence, this is a purely data-driven procedure. Plausible response patterns are supposed to consist of observed variable pairs where responders answer in opposite directions. Correlations across antonyms are consequently negative within each response pattern, and higher negative correlations indicate larger consistency. For psychometric antonym indices, high negative values would indicate that a participant has a consistent response pattern, whereas for psychometric synonym indices a highly positive correlation between responses of item pairs, which are similar to each other, would indicate a consistent response pattern. Widely used and scientifically validated tests (e.g., personality inventories) sometimes have semantic consistency scales customised for the test itself to assess the validity of response patterns (Kurtz and Parrish, 2001). A cut-off score might be obtained by drawing on the first percentile of the frequency distribution. Alternatively, Monte Carlo simulations based on random response patterns determined by actual survey properties can provide the frequency distribution for identifying a cut-off value for valid response patterns.

Response Time

Further useful tools for the detection of invalid response strategies are response time measures. Studies show clear associations between very quick response time and low data quality (Callegaro, Yang, Bhola, Dillman, and Chin, 2009; Malhotra, 2009; Rossmann, 2010). The two main types of response time are variable specific response time and total study completion response time assessment. Since the assumption that reading questions and processing information requires a certain amount of time seems obvious (Tourangeau et al., 2000), cut-off values for unreasonable times could help to identify semi-plausible response patterns. There are diverse procedures for the actual computation and assessment of response time (e.g., Fraley, 2004; Heerwegh, 2003; Kaczmirek, 2009). Lower-bound cut-off scores can be obtained by simulating a fastest possible responder. This could be simulated by the survey designer himself or a third party who is instructed to complete the questionnaire as fast as possible. The person simulating a fastest possible response should be allowed some time to practice before the actual response time assessment. This is to allow adjustment for factors other than effort that could affect response time such as cognitive ability. Cut-off scores can also be based on a variable predicting response time, for instance, the text length of a question (Bergstrom, Gershon, and Lunz, 1994; Halkitis, 1996) or if question draws upon further reading material or figures and illustrations (Bergstrom et al., 1994). The exclusion criterion may also be based on a posteriori analysis. There is supposed to be a notable characteristic within response time frequency distributions common to speeded high-stakes tests, namely, short time spikes. These are especially associated with observed variables that appear at the end of surveys. Short time spikes in item response time frequency distributions are supposed to be located at very low response time values and are often used as thresholds. Wise and Kong (2005) use response time measure for each observed variable to assess *response time effort* (RTE). Drawing on low-stakes tests where participant have no time limit, Wise and Kong (2005) hypothesise rapid-guessing behaviour for unmotivated examinees who will try to respond quickly versus solution behaviour for motivated examinees. An advantage of using the response time to identify outliers is that this measure is characterised by an unobtrusive and non-reactive assessment of which participants are usually unaware. It would further allow an observed-variable specific assessment of a valid versus an invalid response. Caveats of this type of outlier measure are that it has only been found to have modest correlations with other evaluations of valid response patterns (e.g., self-reported effort, Wise and Kong, 2005). Furthermore, we would usually focus on lower-bound cut-off values leaving out those who respond semi-plausibly but slowly. Furthermore, using only lower bounds does not seem to alter substantive findings in terms of marginal distributions and multivariate models (Greszki et al., 2014). Lastly, raw response time should not automatically be seen as an indicator of response quality since it can be assumed to be affected by traits like, for instance, cognitive ability or prior training.

Multivariate Outlier

In order to introduce the terminology used in this section we need to define some notation. Suppose that there are p continuous observed variables and the vector $\boldsymbol{x}^T = (x_1, \ldots, x_j, \ldots, x_p)$ denotes these variables. Let $\boldsymbol{x_i}^T = (x_{i,1}, \ldots, x_{i,j}, \ldots, x_{i,p})$ denote the observed response pattern of the i'th participant with $i = 1, \ldots, n$ and sample size n. Furthermore, let $\bar{\boldsymbol{x}}^T = (\bar{x}_1, \ldots, \bar{x}_j, \ldots, \bar{x}_p)$ be the vector of means for observed variables $\boldsymbol{x}$, where $\bar{x}_j = n^{-1} \sum_{i=1}^{n} x_{i,j}$ (for a symbol directory of notation see Table in A.5).

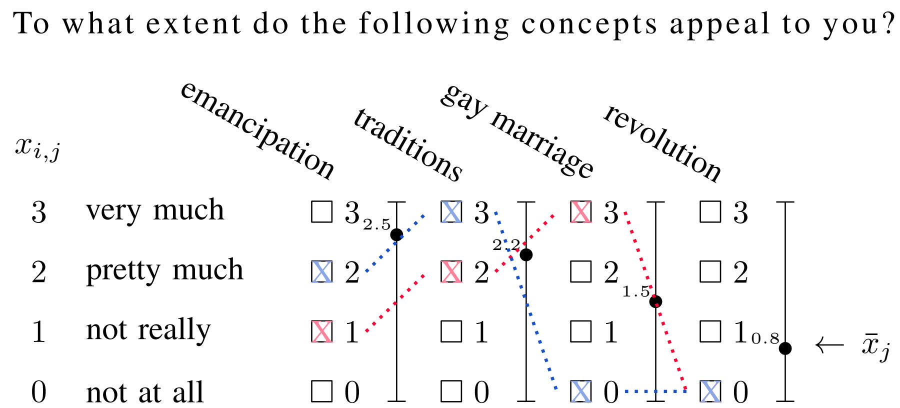
The *Mahalanobis distance*

$$D_i^2 = (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T S^{-1} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \tag{2.1}$$

measures the distance of a response pattern ($\boldsymbol{x_i}$) from the vector of means ($\bar{\boldsymbol{x}}$) of the sample, taking account of the associations between observed variables in the sample covariance matrix S (Mahalanobis, 1936).

Figure 2.1 shows an example of two different response pattern $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ in a multivariate context with given inter-correlations between observed variables in the correlation matrix S. Comparing D_a^2 with D_b^2 here, we can see the response pattern $\boldsymbol{x}_b$ is further from the sample mean pattern. The example has been chosen such that the result is intuitively interpretable: The participant with response pattern $\boldsymbol{x}_b$, answers to the question 'To what extent do the following concepts appeal to you?' with a 'not really' when asked about emancipation but strongly endorses the concept of 'gay marriage'. These two variables are supposedly highly correlated with a correlation coefficient of .6. However, the resulting penalty manifested as a large value in D_b^2 is also a function of participant b's extreme responses on univariate level, i.e. $x_{\text{emancipation}} = 1$ where $\bar{x}_{\text{emancipation}} = 2.5$. Hence, a common caveat remains: extreme values in D_i^2 can in some cases be merely the result of extreme but valid responses.

Nonetheless, this measure has very useful properties as a purely data-driven procedure and accounts for all covariances between observed variables. Meade and Craig (2012) argue that D_i^2 is a powerful indicator of careless response. However, they also point out limitations to purely data-driven approaches such as D_i^2. The efficacy of outlier analysis depends upon the distribution of responses in the sample and, as such, also depends on undesired responses in the data. When careless responses followed a uniform random distribution, D_i^2 performed well in separating valid from invalid response patterns. However, the more observed variables were found to follow a normal distribution regarding careless respondents' data, the less well or even poorly D_i^2 performed in differentiating between valid and invalid responders.



$$S = \begin{pmatrix} 1.0 & & & \\ -0.4 & 1.0 & & \\ 0.6 & -0.8 & 1.0 & \\ 0.4 & -0.8 & 0.4 & 1.0 \end{pmatrix}, \ \bar{\boldsymbol{x}} = \begin{pmatrix} 2.5 \\ 2.2 \\ 1.5 \\ 0.8 \end{pmatrix}, \ \boldsymbol{x}_a = \begin{pmatrix} 2 \\ 3 \\ 0 \\ 0 \end{pmatrix}, \ \boldsymbol{x}_b = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 0 \end{pmatrix}, \ \begin{matrix} \text{hence,} \\ D_a^2 = 13.45, \\ \text{and} \\ D_b^2 = 29.12. \end{matrix}$$

Figure 2.1: Illustrative example for the Mahalanobis distance, for two response patterns labelled a and b.

2.1.2 Person-Fit for categorical Variables

The main goal of using person-fit indices is to identify any kind of aberrant response patterns. A person-fit statistic is best described as an indicator of the degree of reasonableness of a response pattern $\boldsymbol{x_i}$ for a given respondent i. The reasonableness of a participant's response pattern is also judged based on the information provided by all the other response patterns. Person-fit indices are roughly classifiable as parametric or non-parametric person-fit statistics. Where non-parametric person-fit statistics are not based on modelled and estimated parameters, parametric person-fit statistics measure the distance between the actual observed data and the predicted responses under a statistical model. In *item-response theoretically* (IRT) constructed models, the combination of item difficulties and person trait levels help to reveal if persons' response patterns fit the applied model (e.g., via multilevel logistic regression, Conijn, Emons, van Assen, and Sijtsma, 2011; Reise, 2000). The person-fit indices presented in this section are only feasible for binary or ordered categorical observed response variables (e.g., $x_j \in \{0, 1, 2, 3\} \ \forall \ j$).

In his review Karabatsos (2003) found that the first work in person-fit measures is traceable to the early part of the 20th-century (e.g., Cronbach, 1946; Fowler, 1954; Glaser, 1949, 1950, 1951, 1952; Guttman, 1944, 1950; Mosier, 1940; Sherif and Cantril, 1945, 1946; Spearman, 1910; Thurstone, 1927) while research intensified during the late 70s. This increase in research is partly due to the establishment of item response theory models in mainstream psychological assessment (Lord and Novick, 1968; Mokken, 1971; Rasch, 1960). Many researchers have already attempted to compare the quality of over forty currently existing statistics (Birenbaum, 1985, 1986; Drasgow, Levine, and McLaughlin, 1987; Harnisch and Linn, 1981; Harnisch and Tatsuoka, 1983; Kogut, 1986; Li and Olejnik, 1997; Meijer, 1998; Meijer and Sijtsma, 1995; Meijer, 1994; Meijer and Sijtsma, 2001; Meijer, Muijtjens, and van der Vlueten, 1996; Nering and Meijer, 1998; Noonan, Boss, and Gessaroli, 1992; Rogers and Hattie, 1987; Rudner, 1983; Karabatsos, 2003). Comparisons are usually carried out by drawing on either simulated or real empirical data. The following section tries to give only a brief overview of most commonly used person-fit indices. Selected indices have been chosen after personal review (for a list of reviewed indices see Table in A.1) and are to provide an essential understanding of the general concepts and mechanisms of person fit. For more detailed information, please refer to the reviews mentioned previously.

Binary-descriptive Models

The simplest type of person-fit indices are based on a purely descriptive *Guttman model*. The Guttman model does not require any statistical inference due to its strong set of assumptions. The basic assumption is that observed binary response variables $\{x_1, \ldots, x_j, \ldots, x_p\}$ (items) can be ordered such that $\gamma_1 > \gamma_j > \gamma_p$, where γ_j (in IRT terminology, item difficulty) indicates the probability of $x_j = 1$ (correct answer) versus $1 - \gamma_j$ for $x_j = 0$ (incorrect answer). Under the Guttman model, we simply calculate

$$\hat{\gamma}_j = \frac{\sum_{i=1}^{n} x_{i,j}}{n} = \bar{x}_j \tag{2.2}$$

as estimate for γ_j. Another important quantity is y_i (person trait) which indicates an individual's average probability for $x_{i,j} = 1$. Under the Guttman model this is simply estimated by

$$\hat{y}_i = \frac{\sum_{j=1}^{p} x_{i,j}}{p}. \tag{2.3}$$

The most important property of a Guttman model is the perfect pattern. A Guttman perfect pattern is given if $x_{i,j} = 1$ for all $j \leq p\hat{y}_i$ and $x_{i,j} = 0$ for all $j > p\hat{y}_i$, where py_i is the number of items with the response $x_{i,j} = 1$ for respondent i. Any deviation from this is not Guttman conforming.

Meijer and Sijtsma (2001) introduced a general framework common to person-fit indices that are based on a Guttman or similarly parsimonious models. Let ω_j denote a particular choice of weight for each item j, e.g. $\omega_j = \gamma_j$. ω_j is usually ordered such that $\omega_1 > \omega_j > \omega_p$, as is the case for γ_j. Then, a general person fit index is of the form

$$G_i = \frac{\sum_{j=1}^{py_i} \omega_j - \sum_{j=1}^{p} \omega_j x_{i,j}}{\sum_{j=1}^{py_i} \omega_j - \sum_{j=p(1-y_i)+1}^{p} \omega_j}. \tag{2.4}$$

We can interpret G_i as a contrast of i's response pattern $\boldsymbol{x_i}$ to what would on average be expected given y_i and the information provided by all response patterns in the sample. The first term in both the nominator and the denominator is the sum of the weights w_j assigned to the first (ordered) py_i observed variables. In the nominator, the first term is subtracted by the sum of weights that belong to items that an individual answered correctly. Since the items j are ordered according to their weights ω_j (e.g., item difficulty) the nominator is always positive, or 0 for a perfectly Guttman model conforming response pattern. In the denominator, the first term is subtracted by the sum of py_i smallest weights, e.g. most difficult items. Consequently, the denominator equals the nominator in case an individual has a perfectly Guttman model contradicting response pattern. In this case, we have $G_i = 1$. Furthermore, in case the measurement instrument under the Guttman model provides little to no information for the differentiation between individuals' trait scores, e.g. $\omega_j = \omega$ is constant, all patterns are model conforming. In the extreme case of constant ω_j weights, the denominator becomes 0 where G_i is not defined and set to $G_i = 0$, instead. In other words, the measure will penalise non-conforming response patterns more when the measurement instrument is well-designed.

Table 2.1: Values of G_i as simplified index with $\omega_j = \gamma_j$ for four example response patterns $\boldsymbol{x}_i$ of participants i given $\boldsymbol{\gamma}$

Sample Point i	Response Pattern $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,9})$	Value in G_i
1	$(1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$	0.00
2	$(1, 1, 0, 0, 0, 1, 1, 0, 0, 0)$	0.25
3	$(1, 0, 0, 0, 0, 0, 0, 1, 1, 0)$	0.57
4	$(0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$	1.00

$^{\text{Note}}$ $\gamma_1 = 1, \gamma_2 = 0.9, \ldots, \gamma_5 = 0.6, \ldots, \gamma_9 = 0.1$

In the following example, we define the weights such that $\omega_j = \gamma_j$. In IRT terminology, a perfectly Guttman conform pattern consists of correct answers $x_j = 1$ for the easiest py_i questions and incorrect answers $x_j = 0$ for the remaining more difficult, smaller γ_j, questions. Table 2.1 illustrates an example where G_i scores were estimated for a generic sub-group of individuals with response patterns $\boldsymbol{x}_i$ given $p = 10$ binary observed response variables, which are ordered such that $\boldsymbol{\gamma} = (1, 0.9, \ldots, 0.1)$. For $i = 1$ we see a perfectly Guttman conform pattern, whereas $i = 4$ consists of a perfectly Guttman contradicting response pattern. The remaining response patterns take on values between $G_i = 0$ (Guttman model conforming) and $G_i = 1$ (Guttman model contradicting).

Throughout the section of person-fit statistics for binary and categorical variables, I establish common indices under the assumption that observed variables $\boldsymbol{x}$ (items) are measures of a single latent variable y (trait). Sets of observed variables that fulfil this requirement are often referred to as unidimensional scales. I would like to focus on the introduction of more relevant concepts. Where applicable I will comment on limitations of approaches that require this assumption to be true.

Binary-logistic Models

In a binary-logistic model we extend the above binary-descriptive approach by describing the probability of a response pattern $\boldsymbol{x}$ as a function of a single latent (unobserved) variable y and further observed variable specific parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_p)$. We limit the dimension of observed variables to binary responses, $x_j \in \{0, 1\}$ for all j. Hence, for a binomial distribution we express the joint probability function of $\boldsymbol{x}$ given y as

$$g(\boldsymbol{x}|y; \boldsymbol{\theta}) = \prod_{j=1}^{p} g_j(x_j = 1|y; \theta_j)^{x_j}(1 - g_j(x_j = 1|y; \theta_j))^{1-x_j} \tag{2.5}$$

assuming conditional independence of the x_j given y, where g_j is defined as $\Pr(x_j = 1|y; \theta_j)$. For g_j, we may for example use the logistic model

$$g_j(x_j = 1|y; \gamma_j, \alpha_j) = \frac{\exp[\alpha_j(y - \gamma_j)]}{1 + \exp[\alpha_j(y - \gamma_j)]}, \tag{2.6}$$

where $\theta_j = (\gamma_j, \alpha_j)$ are item-specific parameters. Setting $\alpha_j = 1$ leads to the one-parameter logistic model. In a two-parameter logistic model, α_j is freely estimated. α_j is an effect size measure between y and an observed variable x_j controlling for all other parameters in the model (item discrimination parameter).

Where the previously discussed descriptive person-fit indices evaluate the fit to a simple Guttman model based on sample information, IRT-based person-fit statistics give us the possibility to evaluate the fit of a response pattern to a binary-logistic model. This is, in general, a more realistic and flexible representation of phenomena underlying the data. Furthermore, this model allows for statistical inference of goodness of fit to empirical data, rather than just setting untestable assumptions for the detection of participants that do not respond in a Guttman conforming way.

A disadvantage of using person-fit indices that rely on statistical models with latent variables is that, apart from estimating the model parameters, it requires us to estimate individual (latent) y_i scores, as well. Before these are specified, the person-fit formulas in the following sections are not yet usable. In practice, y_i scores are estimated by treating the model parameters as if these were known (Brown and Croudace, 2015). With fixed model parameters, item response probabilities can be estimated assuming that they only depend on y_i (local independence). This assumption allows us to use a *maximum likelihood* (ML) procedure to estimate latent scores, maximising the joint likelihood function of each response pattern. Furthermore, any replacement of true parameter values by their respective ML estimates generally has an impact on the distribution of person-fit statistics.

Generic Person-Fit Snijders (2001) introduced a general framework common to person-fit indices that are based on a binary-logistic model. Let $\omega_j(y)$ and $\omega_0(y)$ be suitable functions for weighting a response and adapting person-fit scale scores respectively, and define

$$G_i = \sum_{j=1}^{p} x_{i,j}\omega_j(y_i) - \omega_0(y_i). \tag{2.7}$$

We can see that the j specific component $w_j(y_i)$ in the first term is only included if $x_{i,j} = 1$. Furthermore, it is a function of the subject-specific variable y_i. This in turn is adjusted by an overall weight $w_0(y_i)$ for all observed variables, which also is a function of y_i. Therefore, a large $w_0(y_i)$, e.g. based on a large y_i value, can undo (justify) a large $x_{i,j}\omega_j(y_i)$ value. As this is a highly abstract generalisation of many person-fit statistics, I shall give more intuition on a specific person-fit index further below.

By defining

$$\omega_0(y_i) = \sum_{j=1}^{p} g_j(x_j = 1|y_i;\theta_j)\omega_j(y_i) \tag{2.8}$$

we can express the person-fit statistic in the centred version

$$G_i^* = \sum_{j=1}^{p}[x_{i,j} - g_j(x_j = 1|y_i;\theta_j)]\omega_j(y_i). \tag{2.9}$$

One of the earliest person-fit indices for probability models was G_i^{sqsr}, which is an individual squared standardised residuals measure (Wright and Stone, 1979). By defining

$$\upsilon_j(y_i) = [p \cdot g_j(x_j = 1|y_i;\theta_j) \cdot [1 - g_j(x_j = 1|y_i;\theta_j)]]^{-1} \tag{2.10}$$

and squaring the (signed) residual term in (2.9), we have

$$G_i^{sqsr} = \frac{1}{p}\sum_{j=1}^{p}\frac{[x_{i,j} - g_j(x_j = 1|y_i;\theta_j)]^2}{g_j(x_j = 1|y_i;\theta_j) \cdot [1 - g_j(x_j = 1|y_i;\theta_j)]}. \tag{2.11}$$

G_i^{sqsr} is the mean of the squared standardised residuals based on p observed variables, taking into account the conditional variances of the individual responses

$$\mathrm{Var}(x_j|y_i) = g_j(x_j = 1|y_i; \theta_j)[1 - g_j(x_j = 1|y_i; \theta_j)]. \tag{2.12}$$

Hence, larger values indicate large residuals and a more severe misfit. According to Wright and Stone (1979) and Wright and Masters (1982) we can transform G_i^{sqsr} to

$$G_i^{sqsr*} = \frac{\ln(G_i^{sqsr}) + G_i^{sqsr} + 1}{df/8}, \tag{2.13}$$

which is asymptotically standard normally distributed with $df = p - 1$ degrees of freedom.

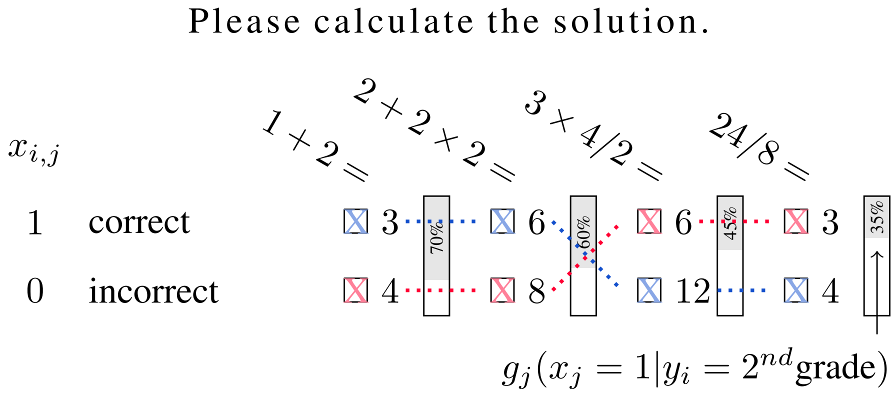
Individual Log-Likelihood as Person-Fit Another way of assessing a person-fit to the model is by drawing on the log-likelihood function used to derive ML estimators (MLE) of the model parameters $\boldsymbol{\theta}$ (Levine and Rubin, 1979). Conditional on y_i, the log-likelihood contribution for individual i is

$$\ell_i(\boldsymbol{\theta}) = \sum_{j=1}^{p}\{x_{i,j} \ln g_j(x_j = 1|y_i; \theta_j) + (1 - x_{i,j}) \ln[1 - g_j(x_j = 1|y_i; \theta_j)]\}. \tag{2.14}$$

The individual log-likelihood function $\ell_i(\boldsymbol{\theta})$ as a measure of person-fit was further developed and applied by others (e.g., Drasgow, Levine, and McLaughlin, 1991; Drasgow, Levine, and Williams, 1985; Levine and Drasgow, 1982, 1983).

Figure 2.2 shows an example comparing ℓ_i values for two different response patterns with given $g(\boldsymbol{x}|y_i; \boldsymbol{\theta})$ values for each item while both have same latent variable level $y_i =$ '2^{nd}grade$'$. In this context y_i can be referred to as ability. We can see that intuitively and numerically the response pattern $\boldsymbol{x}_b$ is less plausible given the model parameters. In this example, individual b answered questions that are ordered according to their difficulty level, such that the most difficult questions were answered correctly and, yet the easier questions answered incorrectly.

However, there are two caveats to this procedure: First, $\ell_i(\boldsymbol{\theta})$ is not standardised. Thus, a decision whether a response pattern is model conforming or model aberrant depends on the very y_i itself. Second, since the null distribution for $\ell_i(\boldsymbol{\theta})$ is usually



j	$g_j(x_j = 1\|y_i = 2)$
1	.70
2	.60
3	.45
4	.35

$$\boldsymbol{x}_a = \begin{pmatrix}1\\1\\0\\0\end{pmatrix}, \quad \boldsymbol{x}_b = \begin{pmatrix}0\\0\\1\\1\end{pmatrix},$$

hence,
$\ell_a(\boldsymbol{\theta}) = -1.896$,
and
$\ell_b(\boldsymbol{\theta}) = -3.969$.

Figure 2.2: Illustrative example of the individual log-likelihood contribution $\ell_i(\boldsymbol{\theta})$, for two response patterns labelled a and b.

unknown, it is difficult to actually classify a response pattern as model aberrant.

Therefore, Drasgow et al. (1985) proposed a standardised version of $\ell_i(\boldsymbol{\theta})$:

$$\ell_i^*(\boldsymbol{\theta}) = \frac{\ell_i(\boldsymbol{\theta}) - E(\ell_i(\boldsymbol{\theta}))}{[\mathrm{Var}(\ell_i(\boldsymbol{\theta}))]^{1/2}} \tag{2.15}$$

where the expectation of $\ell_i(\boldsymbol{\theta})$ is defined as

$$\begin{aligned} E(\ell_i(\boldsymbol{\theta})) = \sum_{j=1}^{p} \{ & g_j(x_j = 1|y_i;\theta_j)\ln[g_j(x_j = 1|y_i;\theta_j)] + \\ & [1 - g_j(x_j = 1|y_i;\theta_j)]\ln[1 - g_j(x_j = 1|y_i;\theta_j)]\} \end{aligned} \tag{2.16}$$

and the variance of $\ell_i(\boldsymbol{\theta})$ can be written as

$$\mathrm{Var}(\ell_i(\boldsymbol{\theta})) = \sum_{j=1}^{p} g_j(x_j = 1|y_i;\theta_j)[1 - g_j(x_j = 1|y_i;\theta_j)][\ln \frac{g_j(x_j = 1|y_i;\theta_j)}{1 - g_j(x_j = 1|y_i;\theta_j)}]^2. \tag{2.17}$$

The theoretical distribution of $\ell_i^*(\boldsymbol{\theta})$ under the true values of y_i is supposed to be standard normally distributed (Molenaar and Hoijtink, 1990, 1996). However, as was said before, any replacement of true parameter values by their respective maximum likelihood estimator generally has an impact on the distribution of person-fit statistics (Molenaar and Hoijtink, 1990; Nering, 1995, 1997; Reise, 1995). In this case, the variance of $\ell_i(\boldsymbol{\theta})$ usually is smaller than expected. Even attempts to correct a smaller empirical Type I error in contrast to the nominal one (e.g, using Warm's y_i estimator) could not account for overestimated positive and underestimated negative values of y_i (van Krimpen-Stoop and Meijer, 1999).

Ordered categorical Models

There are existing generalisations of binary-logistic model person-fit indices which are feasible for measuring a participant's misfit to ordinal categorical responses. The most commonly used model for such items is the ordinal logistic model (known in IRT literature as Graded Response Model, GRM; Samejima, 1970). Suppose there are U response categories $u = 1, \ldots, U$. In the GRM we model the probability of responding to an observed variable given y in or above a category u, i.e. $g_j(x_j \geq u|y; \alpha_j, \gamma_{j,u})$ as

$$g_j(x_j \geq u|y; \alpha_j, \gamma_{j,u}) = \frac{\exp[\alpha_j(y - \gamma_{j,u})]}{1 + \exp[\alpha_j(y - \gamma_{j,u})]}, \tag{2.18}$$

for $u = 2, \ldots, U$ and $g_j(x_j \geq 1|y; \alpha_j, \gamma_{j,u}) = 1$, thereby extending α_j to a slope parameter and $\gamma_{j,u}$ to a threshold parameter. Here the item parameters are $\theta_j = (\alpha_j, \gamma_{j,2}, \ldots, \gamma_{j,U})$. The joint distribution of the items given y is then given by

$$g(\boldsymbol{x}|y; \boldsymbol{\theta}) = \prod_{j=1}^{p} \Pr(X_j = x_j|y; \theta_j), \tag{2.19}$$

where the probabilities $\Pr(X_j = x_j|y; \theta_j)$ are derived from (2.18).

In a model with ordered categorical responses we can use a generalisation of (2.14), the individual log-likelihood contribution

$$\ell_i^{grm}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \sum_{u} \beta_u(x_{i,j}) \ln \Pr(X_j = u|y_i; \theta_j), \tag{2.20}$$

where $\beta_u(x_j) = 1$ if $x_{i,j} = u$ and $\beta_u(x_j) = 0$ otherwise (Drasgow et al., 1985). The expectation for $\ell_i^{grm}(\boldsymbol{\theta})$ is

$$E(\ell_i^{grm}(\boldsymbol{\theta})) = \sum_{j=1}^{p} \sum_{u} \Pr(X_j = u|y_i;\theta_j) \ln \Pr(X_j = u|y_i;\theta_j), \tag{2.21}$$

and the variance of $\ell_i^{grm}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} \mathrm{Var}(\ell_i^{grm}(\boldsymbol{\theta})) = &\sum_{j=1}^{p} \sum_{u,m} \Pr(X_j = u|y_i;\theta_j)\Pr(X_j = m|y_i;\theta_j) \\ &\ln \Pr(X_j = u|y_i;\theta_j) \ln[\frac{\Pr(X_j = u|y_i;\theta_j)}{\Pr(X_j = m|y_i;\theta_j)}]. \end{aligned} \tag{2.22}$$

2.1.3 Person-Fit for continuous Variables

Likelihood-based person-fit indices for categorical responses have become very sophisticated over time, e.g. by adjusting the sensitivity towards extreme factor scores. Many of the previously discussed indices for binary and categorical variables are generalisable to a multidimensional context (see Bartholomew, Knott, and Moustaki, 2011). In social sciences, observed variables are rarely the result of only a single underlying dimension. Hence, latent variable models allowing observed variables to be a function of several unobserved variables are often preferable and, as such, allow for more complex latent variables structures. So far we have seen person-fit indices for binary and categorical variables. For an exhaustive summary of person-fit indices, I shall also draw on log-likelihood estimations of individual response patterns for covariance-based models to further cover continuous observed variables.

For understanding the terminology used in this and following sections, we need to anticipate some of the notation required for the continuous treatment of latent variables. Throughout the thesis, I will use latent variable models in line with the common use of the structural equation modelling framework. Notation will be mostly in line with the unified approach in Bartholomew et al. (2011).

Defining the Latent Variable Framework

Let $\boldsymbol{x}$ be the $p \times 1$ random vector of observed variables and $\boldsymbol{y}$ the vector of q latent variables, then the factor model is given by

$$\boldsymbol{x} = \boldsymbol{\mu} + \Lambda \boldsymbol{y} + \boldsymbol{\epsilon}, \text{ where } \mathrm{E}(\boldsymbol{\epsilon}) = 0 \text{ and } \mathrm{Var}(\boldsymbol{\epsilon}) = \Psi, \tag{2.23}$$

where Λ is a $p \times q$ matrix of factor loadings $\lambda_{j,k}$ and Ψ is a diagonal matrix containing the error variances ψ_j. This implies that the covariance matrix of $\boldsymbol{x}$ is

$$\Sigma = \Lambda \Phi \Lambda^T + \Psi, \tag{2.24}$$

with Φ being the covariance matrix of $\boldsymbol{y}$ but without yet assuming any distributional properties of $\boldsymbol{y}$ or $\boldsymbol{x}$.

In this section, I will focus on a normal linear factor model. Assuming multivariate normality for $\boldsymbol{y}$ and for $\boldsymbol{\epsilon}$, which implies multivariate normal $\boldsymbol{x}$, we consider their joint density with following partitions

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} \sim N_{q+p}\left(\begin{bmatrix} \boldsymbol{\nu} \\ \boldsymbol{\mu} + \Lambda \boldsymbol{\nu} \end{bmatrix}, \begin{bmatrix} \Phi & \Phi \Lambda^T \\ \Lambda \Phi & \Sigma \end{bmatrix}\right), \tag{2.25}$$

where, conventionally and without loss of generality, I choose $E(\boldsymbol{y}) = \boldsymbol{\nu} = 0$ in all models used throughout this thesis.

The conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is then

$$\boldsymbol{y} | \boldsymbol{x} \sim N_q(\Phi \Lambda^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \Phi - \Phi \Lambda^T \Sigma^{-1} \Lambda \Phi). \tag{2.26}$$

(2.26) can be used in case we would like to make inferences about the latent variable on the basis of the observed/manifest variables.

Since Λ is indeterminate up to rotation some constraints on it need to be imposed. In addition, the latent scale for $\boldsymbol{y}$ is typically chosen to be

$$\boldsymbol{y} \sim N_q(0, \mathrm{I}). \tag{2.27}$$

Further constraints on the loadings, beyond what is needed for selecting a factor rotation, can be specified.

The most common method for the estimation of parameters, under (2.25), is maximum likelihood (ML) estimation. We can write the log-likelihood function as

$$\ell(\boldsymbol{\mu}, \Sigma) = \text{constant} + \frac{n}{2}[\ln|\Sigma^{-1}| - \text{trace}[\Sigma^{-1}S^*]], \tag{2.28}$$

where $S^* = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T/n$ and n is the sample size. ML estimates of the parameters are obtained by maximising (2.28) with respect to $(\boldsymbol{\mu}, \Lambda, \Phi, \Psi)$.

Individual χ^2 Contribution

Analogously to log-likelihood ratio tests for model fit, we can estimate the log-likelihood of a model at the level of an individual response pattern (Lange, Westlake, and Spence, 1976) contrasting two components with substitutes for Σ. This approach is similar to the likelihood-based person-fit index $\ell_i(\boldsymbol{\theta})$ introduced in Section 2.1.2. For this purpose I will redefine $D_i^2(\Sigma)$ for individual i, which is analogous to the Mahalanobis Distance used in (2.1):

$$D_i^2(\boldsymbol{x}_i; \Sigma, \boldsymbol{\mu}) = (\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{2.29}$$

where I will omit the notations $\boldsymbol{x}_i$ and $\boldsymbol{\mu}$ throughout this thesis such that $D_i^2(\boldsymbol{x}_i; \Sigma, \boldsymbol{\mu}) = D_i^2(\Sigma)$ to focus on Σ as subject to variation.

When $\boldsymbol{x}_i$ follow a multivariate normal distribution, we have

$$\begin{aligned}
\ell_i(\Sigma) = & \quad \ln\left(\frac{1}{\sqrt{(2\pi)^p \cdot |\Sigma|}} \times \exp^{-\frac{D_i^2(\Sigma)}{2}}\right) \\
= & \quad -\frac{1}{2} \cdot \ln\left[(2\pi)^p \cdot |\Sigma|\right] - \frac{1}{2} \cdot D_i^2(\Sigma) \\
= & -\frac{1}{2} \cdot \left[p \cdot \ln(2\pi) + \ln|\Sigma| + D_i^2(\Sigma)\right] = \quad C_i(\Sigma),
\end{aligned} \tag{2.30}$$

where p is the number of observed variables and I define the contrast component $C_i(\Sigma)$ as the log-likelihood for an individual response pattern $\boldsymbol{x}_i$ under theoretical Σ.

Reise and Widaman (1999) propose a contrast

$$\Upsilon_i(\Sigma, S) = -2[C_i(\Sigma) - C_i(S)] \tag{2.31}$$

where Σ is the model implied covariance matrix for $C_i(\Sigma)$ and S is the sample covariance matrix. This produces a value directly interpretable as an individual's contribution to the overall model χ^2. Large positive $\Upsilon_i(\Sigma, S)$ values indicate patterns with larger contributions to the overall model misfit.

2.1.4 Conclusions from the Review

In this section, we have seen that there are numerous ways and approaches for the identification of certain undesired response patterns. Although this review is not exhaustive since measures have been further developed and enhanced, I covered the most important concepts and aimed to provide general understanding. Outlier indices can be purely data-driven (e.g., multivariate outlier analysis) or fed with theory-driven information (e.g., individual consistency measures). As was discussed, both have their advantages and disadvantages. The simplest indices like the long string measure still remain unstudied in their effectiveness and often might be highly correlated to actual substantive preferences and traits of participants. Response time has often been shown to be a function of cognitive ability or training. Other individual consistency measures have substantial prerequisites, e.g. a reasonable amount of sub-scales for a consistency correlation coefficient to be meaningful. Furthermore, there is an issue of severe dependence on subjective judgements idiosyncratic to the study in question and, thus, neither universally applicable nor comparable. Cut-off values are often arbitrarily chosen or based on rules of thumb, e.g. graphical scree tests. Objective data acquired as response time shows only modest correlation with other identification measures and does not alter substantive findings in terms of marginal distributions and multivariate models.

More sophisticated attempts at identifying invalid responses are provided by person-fit indices. These vary in complexity and feasibility based on properties of observed variables (binary, ordinal, or continuous). However, the distinction between person-fit indices for categorical and continuous variables are somewhat artificial given that, for instance, we can use generalized linear IRT models (Mellenbergh, 1994).

In general, many of those person-fit indices have proven themselves somewhat useful in empirically distinguishing between uncooperative, cooperative and randomly generated classified groups (Birenbaum, 1985). Furthermore, these seem to be sensitive to detecting cheating, creative and careless responding, and lucky guessing (Meijer et al., 1996), as well as in cognitive diagnosis, trying to identify examinee misconceptions (K. K. Tatsuoka, 1996), or even curricular differences among schools (Harnisch and Linn, 1981). However, many of these indices might easily lead to the exclusion of valid responses because even valid but extreme factors scores can produce extreme person-fit values. Another study found that existing procedures are powerless in, for example, detecting careless responses (Woods, 2008). Furthermore, there is disagreement on whether existing indices reliably indicate all kinds of implausible response patterns (Li and Olejnik, 1997). For the interested reader, I would like to refer to more detailed discussions on this topic in Meijer and Sijtsma (2001) and Wise and Kong (2005). Ultimately, almost all of them have been developed for and within the IRT framework and primarily used in educational settings. Therefore, these are limited to mostly binary observed variables and only feasible in single latent variable (unidimensional) frameworks. This is a very unfortunate limitation since associations between several latent constructs would thereby be disregarded.

In light of this review, I will in the following chapters concentrate on the identification measure $\Upsilon_i(\Sigma, S)$ introduced in Section 2.1.3. It has been shown to correlate with many other of the established person-fit indices and, thus, carries the essential idea of those well-researched indices (Reise and Widaman, 1999). However, its potential remains still unstudied and might carry similar problems to other person-fit indices. Nonetheless, the idea behind $\Upsilon_i(\Sigma, S)$ might prove itself especially useful in a complex latent variable framework and, hence, also facilitate the development of a measure that is applicable to a much wider range of study settings. As a consequence of $\Upsilon_i(\Sigma, S)$ as the identification measure of choice, I will be focusing on data with continuous variables (i.e. not considering categorical variables).

2.2 Latent Class Approach

In this review chapter, I have discussed several types of measures that could be used to identify SpRPs. Previously I discussed how SpRPs can be detrimental for the analysis procedure and give us biased estimates by, for instance, increasing measurement error when invalid response strategies are not accounted for. In general, we would like to detect SpRPs in order to remove them from the sample if the assumption is that those cannot provide any information for the analysis of the valid response model, i.e. the constructs of interest. Another method to remove the influence of SpRPs for the estimation of parameters for the valid response model is to accommodate invalid response strategies into the model that result in SpRPs. In this section, I would like to review previous attempts in the literature to accommodate invalid response strategies into the model. However, the first step to do so is to have some indication of whether or not a respondent is a member of a valid or invalid response group. Because this membership variable is usually unobserved (latent) we need to use a method which is commonly referred to as Latent Class Analysis (LCA).

Latent Classes With latent variables, we usually like to assess unobserved underlying phenomena which we indirectly measure with observed indicator variables. The term LCA is used when latent variables are of categorical type. For instance, we would like to identify unmeasured class membership among participants. A latent class allows for variation in parameters of the measurement or structural model and can sufficiently be identified explaining the variation in parameters between different classes of responders (see Lubke and Muthén, 2005). A factor mixture design combines latent classes with confirmatory factor analysis (e.g., Bartholomew et al., 2011). Furthermore, we can define a latent class using other categorical or continuous observed variables, such as covariates. Hence, as with identification measures, we can gain an understanding of differences between different groups of responders, e.g. using demographic covariates. A major advantage of LCA is that we cannot only accommodate different groups of participants into the model, but we also obtain some form of probability measure for class membership given the hypothesised model.

LCA and SpRPs In our case, we have a scenario where the group membership, i.e. valid versus invalid responders, is not observed. We hope to infer group membership by using identification measures such as discussed in Section 2.1 or use LCA to accommodate invalid responses into the model. With LCA we do not necessarily need to classify each participant as valid or invalid but have the choice to do so using posterior class probabilities. I will introduce these concepts in more detail in Chapter 4 when I introduce a possible latent class model to accommodate an invalid response strategy. We will see that LCA can be an effective tool if we know the nature of by the participants employed response strategies. Disadvantages of LCA is that we require having some idea about the nature of SpRPs. For this purpose we can use, prior to the analysis of the model, person-fit indices to identify SpRPs (Meijer and Sijtsma, 2001) or deduct characteristics of SpRPs based on (standardised) residuals under the valid response model (Reiser and VandenBerg, 1994; Reiser, 1996). Ultimately, using identification measures to detect and LCA to accommodate SpRPs are both approaches that can be powerful instruments in dealing with SpRPS, when used in a combined fashion. As mentioned previously, I will implement such a combined approach in Chapter 6.

Similar Settings Most studies that can be found in the literature focus on three types of undesired response patterns, what is often referred to as nuisance data, based on disagreement versus agreement/acquiescence, extreme, or neutral response styles. However, using the term 'styles' in contrast to invalid response 'strategies' hints towards a similar but not identical research topic: Response styles are generally considered to be *content responsive tendencies* of participants to respond to items. Group-specific *extreme response styles* (ERS) are very problematic in, for example, cross-cultural research (e.g., Morren, Gelissen, and Vermunt, 2011) or with regards to socio-demographic differences (e.g., Moors, 2003).

These studies aim to adjust the information gained from the observed responses taking into account individual tendencies, which can influence responses alongside the constructs of interest. For example, the responses of participants who have a tendency to give extreme answers need to be adjusted before we are able to compare their scores with other participants in a survey. Often, we find that factor loadings for people with ERS styles are smaller when the valid response model is allowed to have different parameter estimates based on latent class membership (e.g., Moors, 2003). In these studies, there is little reason for classifying individuals to ERS and non-ERS participants. The goal is to adjust for these tendencies to have an accurate model to represent the data.

Studies of Interest In the following, I will focus on three studies that are representative of research efforts in the field of undesired response patterns. One study is an example for dealing with a content responsive response style (Meade and Craig, 2012), another an example for accommodating a *content non-responsive* invalid response strategy into the model (Moustaki and Knott, 2014), and the last is a study that fits a mixture of normals using a latent class to separate between pathological and non-pathological groups (Wall, Park, and Moustaki, 2015). I will briefly introduce their methodology and discuss findings that are related to the topic of this thesis. In Chapter 4, I will discuss specific aspects using LCA to deal with SpRPs in more detail, which I will do in reference to these studies.

In Meade and Craig (2012), the authors were concerned about inattentive or careless responses in their data. This study has been chosen for discussion because the paper covers a wide range of methods for detecting aberrant response patterns. The authors used data from a questionnaire for the assessment of personality traits. They argued that around 10% to 12% of the sample consists of careless respondents based on the application of a variety of methods for their detection. Alongside person-fit indices such as response consistency indices, (multivariate) outlier statistics, and the use of response time, the questionnaire also provided some bogus items as indicators of careless responding. The valid response model was a one-factor latent variable model that serves as measurement model for one of the Big Five personality factors. Parameters of the valid response model were allowed to vary between latent classes. Varying parameters are the factor loadings and the indicator error variances, where factor variances and indicator intercepts were fixed to ensure model identification. Hence, we have two valid response models for the two (unobserved) groups, i.e. valid versus careless responders. In order to support the formation of a class variable that can separate responders as intended, the latent class variable was defined as a function of previously mentioned (person-fit) indices (covariates in a factor mixture model). The results revealed that factor loadings are smaller for the careless response class. This suggests the presence of larger amounts of measurement error that is not explained by the latent variable when responses are careless. Using posterior probabilities (given the estimated latent class model), 45 were classified as careless responders and 336 as valid responders. The authors identified an issue of multicollinearity among person-fit indices (covariates for the latent class variable). However, when the same latent class model was analysed with different subsets of covariates, classifying responders lead to considerably different sets of responders that were classified as careless. Furthermore, the authors employed a logistic regression for class membership (based on the model with all covariates included) as the dependent variable and the covariates as predictor variables. This post hoc analysis suggests that the index psychometric synonym had the biggest influence on the formation of the latent class variable. The next best predictor was the even-odd consistency index and, unexpectedly, the sum of bogus items as an indicator for careless responding, which was only third best in explaining class membership after accounting for the other predictor variables. The long string index was least successful in explaining the class membership. These results suggest that data caused by careless responses can appear more plausible than what, for instance, a cursory screening of the data could detect.

In Moustaki and Knott (2014), the authors were concerned with response patterns to which they referred to as atypical. Atypical response patterns are assumed to be generated by a so-called *secondary* (invalid) *response strategy* which is different from the *primary* (valid) *response strategy.* In line with the majority of studies in this field, the authors refrain from using the label 'random' responses in this context, noting that atypical responses are not be seen as truly random. The authors state that they were motivated to accommodate atypical responses because hypothesised valid response models for the analysed data did not fit the data well. Undesired response patterns were identified as one cause for model misfit. The study is based on two datasets: data drawn from the Workplace Industrial Relations Survey (WIRS) from 1990 and British Social Attitudes (BSA) Survey of 2007. The data is of binary type and the observed variables serve as indicators for the one-factor latent variable model as the valid response model. The measurement model for the WIR data has six indicator variables, and the authors use five indicator variables for the BSA data. The valid response model is defined using two-parameter logistic links to the continuous latent variable. A mixture model for two classes is employed where an unobserved pseudo-item (latent class variable) models whether a chosen atypical response pattern is the result of a valid or invalid response strategy. Furthermore, the latent class variable is a function of the latent variables of interest and other covariates (e.g., socio-demographic variables) to help identify characteristics of invalid respondents. However, the valid response model is estimated free from the effects of the invalid response strategy and the covariates. The authors' unique approach consists in investigating possible atypical response patterns prior to the actual analysis of the model and has the advantage that no further factors need to be introduced into the model. Using this method, it is crucial to identify possible response patterns that can be the result of an invalid response strategy. Where there are many different ways in doing so, by for instance making use of person-fit indices, in this study response patterns are pre-flagged when their unstandardised residual value is greater 10. This residual is defined as the difference between observed and expected frequency under the estimated valid response measurement model. Hence, before the actual analysis, the measurement model for all responses is estimated without taking into account invalid response strategies. 9 response patterns where pre-flagged in that manner for the WIR data and 3 were pre-flagged for the BSA data. Based on these choices 9 and respectively 3 different analyses of the mixture model were run. This sophisticated design requires the fit of several models. Resulting goodness-of-fit indices and the estimated ratio between members of primary and secondary response classes (amongst other information) were used as judgement criteria for the identification of atypical response patterns. The authors conclude that if they had to choose one atypical response pattern (based on the results), it would be '101010' for the WIR data and '00000' for the BSA data. According to this model, both response patterns can also be the result of a valid response strategy. Given participants with these response patterns were using an invalid response strategy, we could interpret them as follows: a consistently alternating response strategy for the WIR survey and a long string response strategy for the BSA survey, e.g. consistently answering with 'no'.

In Wall et al. (2015), the authors were concerned about zero-inflated data, which is a manifested response pattern of participants who belong to a large non-pathological proportion of the sample. The main concern is that IRT models cannot represent both groups without taking into account that a large percentage of participants have non or few symptoms. The authors reject the assumption of normality for the underlying trait and, instead, allow the latent trait be a function of a mixture of normals including a degenerate component representing the non-pathological group.

The results reveal that incorrectly assuming normality leads to biased discrimination and severity estimates.

Conclusions from the Review Ultimately, the majority of in this context of undesired response patterns discussed studies did not only improve model fit via the use of factor mixture models. Most studies helped in gaining some understanding about the nature of invalid response strategies using LCA. Unfortunately, none of the studies dealing with invalid responses had an experimental design such that group membership is observed and, consequently, there was no possibility of assessing the actual accuracy of group allocations. Another important aspect is the accessibility of methods to a large non-expert audience: Where some studies are easily implementable using popular analysis software for latent variable models such as *Mplus* (Muthén and Muthén, 1998–2012), or *Latent GOLD* (Vermunt and Magidson, 2013), many require very case-specific specialist software or even analytical derivations and computational implementations in programming languages such as *R* (R Core Team, 2016). Furthermore, using the mixture model method often requires identifying specific invalid response patterns pre-analysis and can only account for mostly one type or a limited number of invalid response strategies. We cannot account for individual invalid response strategies, which can effectively all lead to SpRPs. The complex nature of the methodology discussed in this review also makes it difficult to establish a universally applicable setting for a wide range of research scenarios. For instance, factor mixture models can become computationally expensive accompanied by a large increase of free parameters per class, which makes estimated models less stable and can lead to model identification problems, especially when used for categorical data. Additionally, where the use of covariates can help to form the latent class variable more accurately (as intended), latent class variables defined in that manner can lead to different formations of class variables for different studies and make it difficult to generalise findings. For these reasons, our main goal is the identification of semi-plausible responders and less the investigation of solutions for an appropriate statistical modelling approach. However, because LCA and identification measures used in combination can be a powerful tool in detecting SpRPs, elaborating and evaluating its use in the context of SpRPs will play an important role in the following chapters of this thesis.

Chapter 3

Empirical Data and the basic Latent Variable Model

Previous chapters of this thesis defined the study subject, set terminology, and put SpRPs in relation to existing research. I further reviewed the most important methods that can help to either detect or accommodate SpRPs into the model. This chapter establishes a statistical framework for the analysis of valid response patterns. For this purpose, I will draw upon two empirical datasets underlying a well-established theoretical framework, namely, the Big Five Personality Factors. The goal is to acquire detailed and extensive knowledge of the empirical studies at hand and differences between experimental sub-groups within the valid response model. This is a first step towards understanding potential effects of SpRPs on parameter estimates and forming hypotheses about the statistical nature of SpRPs in a latent variable model.

3.1 The Big Five Personality Factors

The Big Five factors *Emotional Stability* (formerly referred to as *Neuroticism*), *Extraversion*, *Openness to Experience*, *Agreeableness* and *Conscientiousness* aim to describe a person's personality in all its facets. Empirical studies usually show weak to moderate correlations among the Big Five (Digman, 1997) although, theoretically, the Big Five are conceptualised as orthogonal/distinct latent factors (e.g., Costa and McCrae, 1995; Goldberg, 1993). However, this is not to the extent that would undermine the stable five-factor structure. When the goal is to use a framework that is empirically well known, the Big Five framework provides a solid base to validate or explore new statistical methods. Furthermore, personality assessment was one of the first scientific areas to use factor analytic procedures, and the Big Five framework itself is the result of multivariate analysis methods.

Among the Big Five, Openness to Experience is supposed to be the least stable and most controversial factor (e.g., Zuckerman, Kuhlman, Joireman, Teta, and Kraft, 1993). Especially studies seeking to investigate similarities and differences in the personality structure throughout different cultures reveal Openness to being the least distinct amongst the Big Five. Furthermore, the most successful applications of the Big Five are attributed to factors Emotional Stability and Conscientiousness. Both have proven to be very useful in organisational personnel or clinical disorder assessment as well as in predicting general career success, health, and even intelligence (e.g., Judge, Higgins, Thoresen, and Barrick, 1999; Friedman et al., 1995). It is noteworthy that Schmitt, Chan, Sacco, McFarland, and Jennings (1999) also found a connection between Conscientiousness and the choice of invalid response strategies. They found that test-taking motivation and conscientiousness were correlated moderately with person-fit indices for personality tests, and to a lesser extent, for cognitive tests. Furthermore, male participants had smaller person-fit values indicating higher misfit than female participants. However, when controlled for conscientiousness, this effect was eliminated. It was concluded that invalid response strategies explain the misfit of male participants.

As the relation of the Big Five factors to each other are relevant to the model specification, I will shortly introduce findings of a meta-analysis that seeks to explain the Big Five factors in a higher order framework. D. van der Linden, te Nijenhuis, and Bakker (2010) investigated the existence of a *General Factor of Personality* (GFP) in their meta-analysis. Although the existence of a GFP is still controversial in literature, D. van der Linden et al. could at least provide strong evidence for a higher order hierarchy of the Big Five (also see Digman, 1997). According to this, the Big Five factors Openness and Extraversion can be used as indicators of a so-called *Beta-Factor* and the remaining factors Conscientiousness, Agreeableness, and Neuroticism (or, inversely interpreted, Emotional Stability) are commonly affected by an *Alpha-Factor*. Thus, Big Five factors constituting the Alpha-Factor are more distinct from those that constitute the Beta-Factor than they are from each other. Nevertheless, even the higher order factors are still correlated.

3.2 Investigated Datasets

In order to empirically investigate statistical features of SpRPs, I draw upon two studies that collected data using the same assessment instrument. Therefore, in the following section, I will inform the reader about the assessment instrument and study specific design aspects.

3.2.1 Assessment Instrument

The *International Personality Item Pool* (IPIP; Goldberg, 1999) is built through international effort to develop and continually refine a set of personality inventories. These items are in the public domain, and the scales can be used for both scientific and commercial purposes. There is a large number of scientific publications validating the IPIP and its scales (for further reference, see http://projects.ori.org/lrg/).

The IPIP-NEO items are reliable measures of 30 personality facets (sub-scale factors) and are in line with scientifically acknowledged Big Five personality factor framework. The data at hand uses the most recent IPIP version. Each of the 30 sub-scales is formed by ten indicators (300 items). Each of the Big Five factors is measured by drawing on each six sub-scales which serve in turn as indicators for the global five factors. The lowest order indicators are observed variables measured via 5-point *Likert-type* scale answer options ranging from 1 (very inaccurate) to 5 (very accurate). Participants choose the answer category that applies to their personality as a response to a statement that describes a certain personality aspect.

3.2.2 Experimental Study (Huang et al., 2012)

The experimental design splits the IPIP's 300 items questionnaire into two halves, shaping a pseudo factor where different items are assessed (e.g., five items for a sub-scale in the first half and the other five items of the same sub-scale in the second half of the questionnaire). Hence, there are 150 items for each half of the questionnaire, basically collecting data for the same constructs but drawing on a different set of indicators.

Table 3.1: Instructions

Instructions	
1st half	2nd half
Normal instruction: There are no correct or incorrect answers. Describe yourself as you honestly see yourself. *Warning (additional):* Sophisticated statistical control methods are used to check for validity of responses and that responding without much effort would result in loss of credits.	*Continue:* Continue the instructions from the first half of the survey. *Cautionary IER:* Respond without much effort but pretend that you want your laziness in filling out this survey to remain undetected. *Outright IER:* Respond without effort with no risk of penalty: in fact, we request that you do so.

The research design is essentially based on a randomised 2x3 factor design. The first factor splits the sample into two groups with either normal instructions about how to fill out the questionnaire or normal instructions with an additional warning. The warning informs the respondents of the existence of statistical control methods which aim to check for validity of responses, and that responding without much effort would result in loss of credits. This first factor applies to the first half of the questionnaire. The second factor (partly) randomises the participants into three different groups before starting to fill out the second half of the questionnaire. One group is told to continue filling out the questionnaire as was instructed in the first half. The second group is to respond without much effort but to pretend as if they would like their laziness in filling out this survey to remain undetected (Cautionary IER). In the third group, the authors seek to induce the most extreme IER form by instructing the participants to respond without effort and with no risk of penalty (Outright IER). In fact, these are requested to do so. The instructions are summarised in Table 3.1.

The rationale for the 2x3 factor design was also to assess any difference between the first factor conditions warning versus normal instruction before inducing any IER conditions. Initial analyses and study results did not show any clear, meaningful pattern but somewhat fewer identification measures indicating occurrences of IER in the first half of the questionnaire. Carry-over effects of the second-factor conditions were found to be negligible but cannot be entirely excluded. In order not to further complicate following analyses, I will limit the focus on items of the second half of the questionnaire only and assume a one-factor design. Furthermore, the data presented in the next section has the purpose of providing us with more accurate parameter estimates. Analysis models should be kept free of design specific aspects such that sample differences are easily accessible. Nonetheless, we shall remain sensitive to differences between both first-half factor conditions and report where they give meaningful insights on study objectives.

Table 3.1 summarises the instruction for each of the cells 1 to 6. The cells are labelled according to unique conditions as shown in Table 3.2.

Table 3.2: Sub-samples

Groups	n_{Cell}	Conditions	
		1st half	2nd half
Cell 1	39	Warning	Continue
Cell 2	57	Warning	Cautionary IER
Cell 3	55	Warning	Outright IER
Cell 4	84	Normal instruction	Continue
Cell 5	64	Normal instruction	Cautionary IER
Cell 6	81	Normal instruction	Outright IER

$^{n_{\text{Cell}}}$ Sub-sample size.

The sample comprised 380 undergraduate students at a large Midwestern university (74 female, mean age = 21 years). A subset of respondents ($n = 39$) were students of one of the authors who volunteered to participate and were thought to be highly motivated to respond accurately and follow directions. This subgroup was assigned to Cell 1, partly compromising the otherwise randomised design.

3.2.3 Online Questionnaire (Johnson, 2005)

Johnson (2005) sought to estimate the relative incidence of invalid response patterns in online surveys versus by paper-and-pencil assessed personality measures. The sample for the web-based assessment comprises 23,994 participants of the IPIP's 300 items questionnaire. Approximately 3.8% of responses were judged as duplicates, about 3.5% as result of long string response strategies, and nearly 1% as invalid due to linguistic incompetence or inattentive responding. These classifications were conducted in a very conservative manner and validated in cursory investigations. Hence, I will use a sub-group of $n = 20,999$ responses which was cleared by the author.

3.3 Analysis assuming valid Responses only

In this section, I will present the statistical (theoretical) model to analyse all the data at hand under the assumption that the sample consists of valid responses only. In line with the model defined in Section 2.1.3, I will use the latent variable framework. However, I will not use the common factor analysis model of orthogonal factors to identify the model parameters. Instead, in the following section, I will set different constraints upon Λ in order to fit a model that is identified. The constraints on Λ outlined in the following section are in line with the theory about the Big Five personality factors and not all constraints are required for identification.

3.3.1 Theoretical Model

To give a detailed description of the model, I will make use of the structural equation modelling (SEM) framework. I am carrying out a confirmatory factor analysis, which is a special case of a SEM as there is no structural component to the model. Hence, I only employ so-called *exogenous* latent variables where there are no latent predictors in the model.

In accordance with the Big Five Framework and associated assumptions, I will treat the Big Five latent factor indicators as continuous items in a latent modelling framework. The models are estimated using maximum likelihood (ML) estimation. The Big Five framework was based on a linear factor model and are in fact the result of exploratory factor analyses. Hence, investigating structural and measurement parameter behaviour under the assumptions commonly made in the literature, and respective empirical findings, is in line with the global study objective. We would like to investigate the influence of invalid response strategies under a usually employed analysis context.

The focus lies on three of the Big Five factors, namely Emotional Stability (N), Extraversion (E), and Agreeableness (A). The selection was based on the mutual affiliation of N and A within the Alpha-Factor (see Section 3.1). Furthermore, I have chosen E in order to include a construct from within the Beta-Factor in anticipation of a more distinct factor structure. Hence, a valid structural model should reveal a negative association between the two latent factors N and A and small or no association between those (negative for N) and E. Remaining Big Five factors Conscientiousness (C) and Openness (O) are not taken into account for several reasons: First, we usually would not like to unnecessarily complicate the factor structure, especially since experimental plausibility conditions are represented only through medium sample sizes in the data. This way we support stable parameter estimates reducing the number of estimated model parameters versus sample size ratio. Secondly, as discussed in Section 3.1 the stability and validity of the Big Five factor Openness is still subject to a controversial debate, especially in intercultural settings and since this thesis employs and compares parameter estimates of two different studies (see Section 3.2), omitting this latent factor was judged to be a sensible step. Lastly as briefly mentioned in Section 3.1, the personality trait Conscientiousness might be related to some participants' choice of invalid response strategy or other construct-of-interest related aspects towards plausibility of response patterns. Associations in this idiosyncratic manner could unpredictably complicate model specification and resulting parameter estimates. Included latent variables and their parameters including estimated covariances between latent variables are listed in Table 3.3. Means and variances were fixed to 0 and 1, respectively, in the structural equation model. Hence, factor covariances in the last column are, more specifically, correlations between the latent variables.

The measurement model consisting of three latent variables formed by each six

Table 3.3: Parameter notations and labels related to the three latent variables

Big Five factor		y_k	ν_k	$\phi_{k,k}$	$\phi_{m,k}$
Neuroticism	(N)	y_1	$\nu_1 = 0$	$\phi_{1,1} = 1$	$\phi_{3,1}$
Extraversion	(E)	y_2	$\nu_2 = 0$	$\phi_{2,2} = 1$	$\phi_{2,1}$
Agreeableness	(A)	y_3	$\nu_3 = 0$	$\phi_{3,3} = 1$	$\phi_{3,2}$

^{y_k} Latent variable.

$^{\nu_k}$ Latent variable mean.

$^{\phi_{k,k}}$ Latent variable variance.

$^{\phi_{m,k}}$ Covariance of latent variables $m \neq k$.

observed variables (indicators) can be written as follows:

$$\boldsymbol{x} = \boldsymbol{\mu} + \Lambda \boldsymbol{y} + \boldsymbol{\epsilon} \tag{3.1}$$

Here, the factor loadings are further restricted following a simple loading structure such that

$$\begin{pmatrix} x_1 \\ \vdots \\ x_6 \\ x_7 \\ \vdots \\ x_{12} \\ x_{13} \\ \vdots \\ x_{18} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_6 \\ \mu_7 \\ \vdots \\ \mu_{12} \\ \mu_{13} \\ \vdots \\ \mu_{18} \end{pmatrix} + \begin{pmatrix} \lambda_{1,1} & 0 & 0 \\ \vdots & 0 & 0 \\ \lambda_{1,6} & 0 & 0 \\ 0 & \lambda_{2,7} & 0 \\ 0 & \vdots & 0 \\ 0 & \lambda_{2,12} & 0 \\ 0 & 0 & \lambda_{3,13} \\ 0 & 0 & \vdots \\ 0 & 0 & \lambda_{3,18} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_6 \\ \epsilon_7 \\ \vdots \\ \epsilon_{12} \\ \epsilon_{13} \\ \vdots \\ \epsilon_{18} \end{pmatrix}. \tag{3.2}$$

In a latent variable model with simple factor loading structure, each item serves as indicator for only one of the latent variables and the remaining factor loadings of the same observed variable are set to 0. Lastly, we can write the model implied covariance matrix defined in (2.24) as follows:

$$\Sigma = \Lambda \begin{pmatrix} \phi_{1,1} & & \\ \phi_{2,1} & \phi_{2,2} & \\ \phi_{3,1} & \phi_{3,2} & \phi_{3,3} \end{pmatrix} \Lambda^T + \begin{pmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_{18} \end{pmatrix}, \tag{3.3}$$

where ψ_j is the error variance for observed variable j that is not explained by the latent variables.

In line with the theory of the Big Five personality framework elaborated in Section 3.1 and the applied test instrument introduced in Section 3.2.1, I employ responses of the items listed in Table 3.4 as observed variables and allow them to be indicators of the respective Big Five latent variables.

I chose not to aggregate to sub-scale variables (e.g., by summing up item responses belonging to sub-scales) for three reasons: First, it is important that we preserve information derived from reversed coded items in raw form as these might help to identify long string responses. Moreover in general, we would like to preserve as much information as possible with regards to the research objective where aggregation is usually associated with loss of information. Second, the items where selected based on the subset of questions placed in the second half of the experimental study. Hence, the number of available items for sub-scale aggregation is significantly reduced (and varying between sub-scales). Furthermore, a pre-selection based on item reliability with regards to their corresponding Big Five personality factor further reduces the number of items available for aggregation. Third, missing responses where in large numbers eliminated following carefully derived list-wise exclusion criteria as described in detail in Johnson (2005) for the online questionnaire sample. However to disregard complications based on missing responses, items with zero to only a few missing responses were selected for analysis. The occasional missing response was simply replaced by the middle answer category, following the test instrument's normative guidelines (and author's procedures). The last selection criterion further reduced the amount of variables that could be used for sub-scale aggregation.

The measurement model includes only indicators from the second half of the questionnaire such that members of the three experimental response strategy conditions with resulting response patterns implausible, semi-plausible, and plausible

Table 3.4: List of IPIP observed variables as indicators for the three latent variables

Item	x_j	μ_j	$\sigma_{j,j}$	ϵ_j (ψ_j)	$\lambda_{k,j}$
Neuroticism					
VUL 056 (+)	x_1	μ_1	$\sigma_{1,1}$	ϵ_1 (ψ_1)	$\lambda_{1,1}$
ANX 008 (+)	.	.	.	.	.
ANX 006 (+)	.	.	.	.	.
ANX 009 (-)	.	.	.	.	.
DEP 030 (-)	.	.	.	.	.
VUL 060 (-)	x_6	μ_6	$\sigma_{6,6}$	ϵ_6 (ψ_6)	$\lambda_{1,6}$
Extraversion					
GRE 077 (+)	x_7	μ_7	$\sigma_{7,7}$	ϵ_7 (ψ_7)	$\lambda_{2,7}$
GRE 076 (+)	.	.	.	.	.
ASS 086 (+)	.	.	.	.	.
GRE 080 (-)	.	.	.	.	.
EXS 110 (-)	.	.	.	.	.
GRE 079 (-)	x_{12}	μ_{12}	$\sigma_{12,12}$	ϵ_{12} (ψ_{12})	$\lambda_{2,12}$
Agreeableness					
MOR 200 (+)	x_{13}	μ_{13}	$\sigma_{13,13}$	ϵ_{13} (ψ_{13})	$\lambda_{3,13}$
TRU 189 (+)	.	.	.	.	.
ALT 208 (+)	.	.	.	.	.
COO 220 (-)	.	.	.	.	.
MOR 198 (-)	.	.	.	.	.
ALT 209 (-)	x_{18}	μ_{18}	$\sigma_{18,18}$	ϵ_{18} (ψ_{18})	$\lambda_{3,18}$

^{x_j} Observed variable.
$^{\sigma_{j,j}}$ Observed variable variance.
$^{\epsilon_j(\psi_{j,j})}$ Observed variable error term (error variance).
$^{\lambda_{k,j}}$ Observed variable factor loading.
$^{(+/-)}$ Denoting the direction of the original coding. Those indicated '-' were recoded so that items were all positively coded.

(sub-samples) can be compared. Each latent variable is measured by, on affiliation with their respective latent construct, three positively associated and three negatively associated observed variables. However, theoretically negatively associated question responses were recoded such that positive factor loadings are expected throughout the measurement models.

The structural equation in (3.2) can also be represented graphically via a path diagram as shown in Figure 3.1.

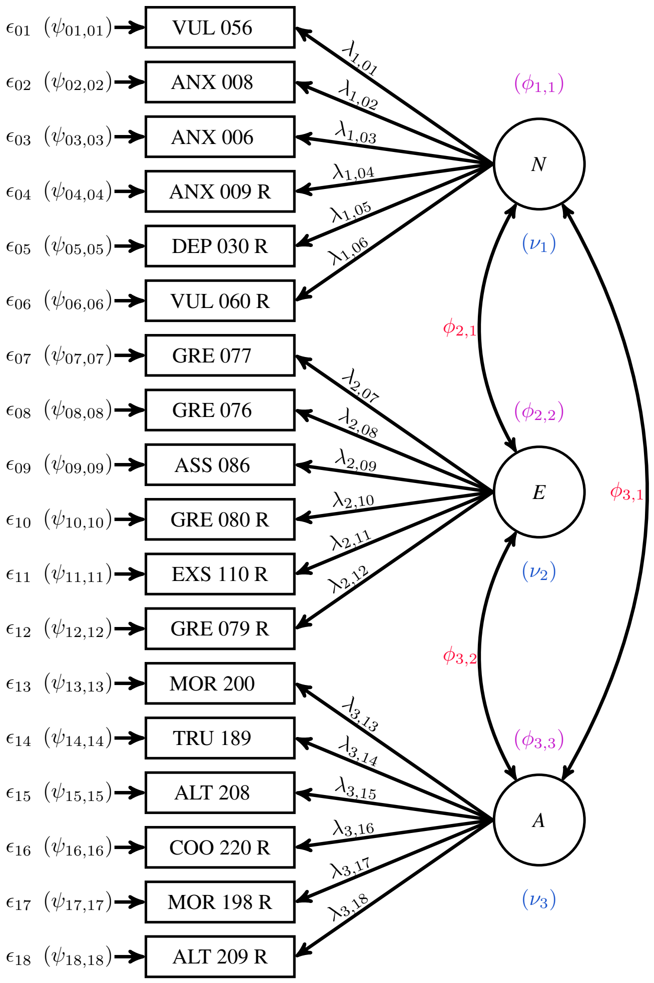

ϵ_{01} $(\psi_{01,01})$
ϵ_{02} $(\psi_{02,02})$
ϵ_{03} $(\psi_{03,03})$
ϵ_{04} $(\psi_{04,04})$
ϵ_{05} $(\psi_{05,05})$
ϵ_{06} $(\psi_{06,06})$
ϵ_{07} $(\psi_{07,07})$
ϵ_{08} $(\psi_{08,08})$
ϵ_{09} $(\psi_{09,09})$
ϵ_{10} $(\psi_{10,10})$
ϵ_{11} $(\psi_{11,11})$
ϵ_{12} $(\psi_{12,12})$
ϵ_{13} $(\psi_{13,13})$
ϵ_{14} $(\psi_{14,14})$
ϵ_{15} $(\psi_{15,15})$
ϵ_{16} $(\psi_{16,16})$
ϵ_{17} $(\psi_{17,17})$
ϵ_{18} $(\psi_{18,18})$
VUL 056
ANX 008
ANX 006
ANX 009 R
DEP 030 R
VUL 060 R
GRE 077
GRE 076
ASS 086
GRE 080 R
EXS 110 R
GRE 079 R
MOR 200
TRU 189
ALT 208
COO 220 R
MOR 198 R
ALT 209 R
$\lambda_{1,01}$
$\lambda_{1,02}$
$\lambda_{1,03}$
$\lambda_{1,04}$
$\lambda_{1,05}$
$\lambda_{1,06}$
$\lambda_{2,07}$
$\lambda_{2,08}$
$\lambda_{2,09}$
$\lambda_{2,10}$
$\lambda_{2,11}$
$\lambda_{2,12}$
$\lambda_{3,13}$
$\lambda_{3,14}$
$\lambda_{3,15}$
$\lambda_{3,16}$
$\lambda_{3,17}$
$\lambda_{3,18}$
$(\phi_{1,1})$
N
(ν_1)
$\phi_{2,1}$
$(\phi_{2,2})$
E
(ν_2)
$\phi_{3,1}$
$\phi_{3,2}$
$(\phi_{3,3})$
A
(ν_3)

Figure 3.1: Path diagram for the Big Three factor model.

3.3.2 Descriptive Statistics

In the previous section, I chose an appropriate valid response model to represent three of the Big Five personality factors and investigate semi-plausible response patterns. The next step is to fit the model described in Section 3.2. The main goal of this section is to check whether the model assumptions are met for the previously introduced datasets. Furthermore, I would like to establish comparability of the results for the experimental study sample and the online questionnaire sample.

Assumptions and Data Scaling

When we compare the factor analysis model and the data, it is apparent that one assumption of the model is not met, i.e. continuous multivariate normal observed variables as indicators for the latent variables. The data at hand is ordered categorical based on a 5-point Likert-type scaling and coded into ordered integer answer options ranging from 1 (very inaccurate) to 5 (very accurate). In this and the following chapter, I will treat these observed variables as if they were assessed on a continuous normal scale. Here, I shall assess the extent to which we deal with approximately normal data. I have chosen to proceed with the continuous treatment of observed variables mainly because of three reasons: First, the IPIP and other personality measurement instruments are the results of decades of research on the Big Five Personality Theory based on Likert-scale answer format and multivariate normality assumptions. Consequently, our results can be appropriately compared with those in previous literature within the same consistent framework. Secondly, it is crucial to ensure model parsimony when the model is used as an instrument for research of exploratory nature. I aim to extract patterns of differences between model parameter estimates based on different sub-samples with and without invalid responses. For instance, it would be more appropriate to fit a model where we assume an underlying distribution to the ordered categorical data which is captured by, in this case, 4 threshold parameters (5 answer options) for each of the $p = 18$ observed variables. This can be done with response function models, e.g. the graded response model (for an overview, ordinal variables in latent variable models, W. van der Linden and Hambleton, 1997). However, the results will be more difficult to interpret with a large number of parameter estimates, especially when comparing analyses outcomes for different samples. Lastly, through continuous treatment, we gain degrees of freedom for the analysis of more complex latent class analysis models, which we will need in the next Chapter 4. With latent class models, I seek to incorporate invalid response strategies into the model. As was discussed previously, SpRPs are difficult to detect and can seem very plausible on the surface of observed variables.

Hence, it is important to utilise the latent structure underlying the data. I argue, that the more sophisticated the latent structure (e.g., the number of latent variables) in the valid response model, the more information we gain about the nature of valid responses that is not shared with invalid responses and help discriminate SpRPs. Given the study objective, I decide in favour of stability and consistency of model estimates, comparability of findings, and flexibility in structural model definitions over accurate valid response model specification and maximising model fit to the data.

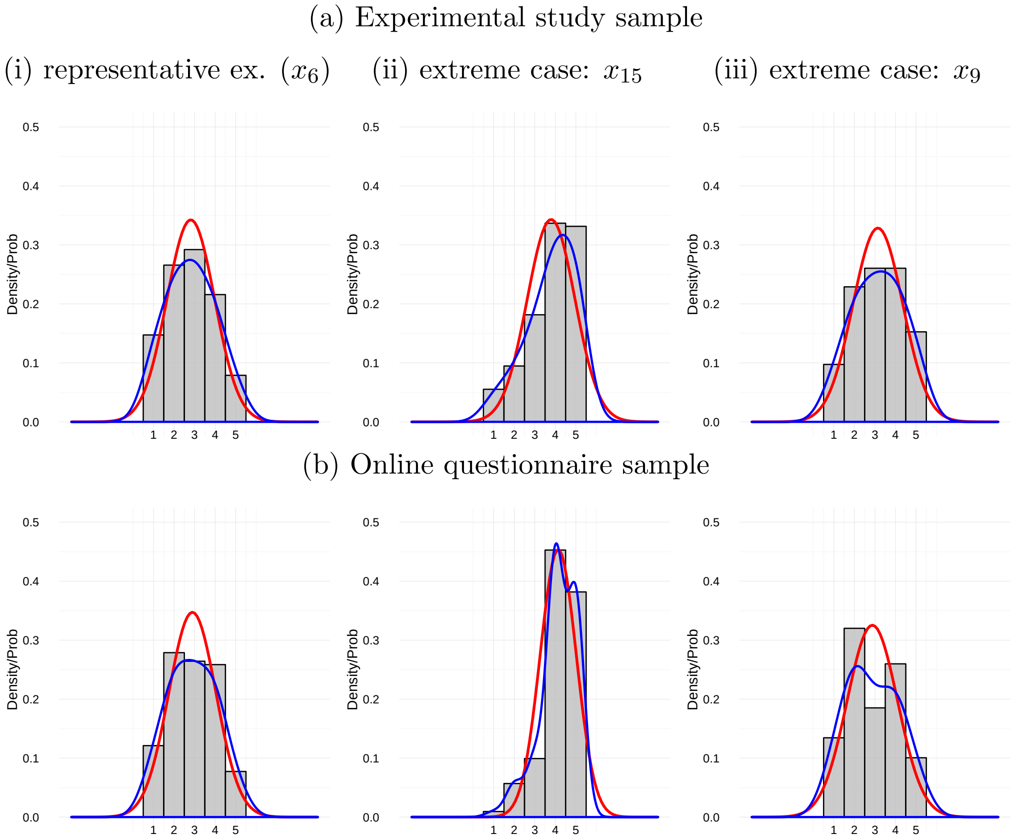
Having established the overall setting of the factor analysis model where we assume multivariate normality of the variables, we can compare the results and investigate the impact of semi-plausible responses and, hence, measurement error on the estimation. However, prior to the analysis, I shall investigate several descriptive statistics of the two datasets. We would like to gain a more explicit understanding of the data at hand. The correct interpretation of research outcomes requires knowledge about to what extent the study samples are suitable (e.g., approximate normality) in providing data for the analysis of the specified valid response model.

It is common for psychological constructs to be found roughly normally distributed in nature. Hence, we expect individuals to have mostly similar values symmetrically varying around a population mean where extreme difference are expected to be rare. There are different approaches towards measuring psychological constructs. Difficulties arise when the measurement is based on self-assessment, e.g. attitude questions, such as is the case for the IPIP. A popular answer format is the Likert scale. Where psychological literature provides uncountable examples to justify the assumption of normality, the ability of the Likert scale answer format to accurately capture the underlying distribution needs to be assessed individually in each study setting. Normality of univariate distributions is a prerequisite of multivariate normality. For this purpose, I will present univariate histograms comparing the data to a normal distribution. The parameters for the univariate normal distributions will be estimated using ML estimates of mean and variance. Furthermore, we can derive at graphically informed decisions of multivariate normality with multivariate distribution plots (e.g., bivariate or 3D-distribution plots). We can compare measures, such as the Mahalanobis distance, to their expected distributions with a QQ-plot. The theoretical distribution of the $D_i^2(\Sigma)$ is well known to be χ^2 distributed given the assumption of multivariate normality is met (see Section 5.4).

First, I will investigate the univariate distributions of the observed variables for both study samples. In general, a Likert scale answer format can lead to problems because the answer options can only capture a two-sided truncated version of a normal distribution. Mean and variance estimates for the underlying normal distribution may be slightly biased depending on the position of the population mean. For instance, a population mean that strongly deviates from the Likert scale middle answer category $x_j = 3$ can lead to skewed data when the variables are truncated. Extreme answers can only be captured by the largest (or smallest) answer options $x_j = 1$ and $x_j = 5$. The same issues arise when an observed variable's population variance is very large. Moreover, for the experimental study sample, I expect irregularities based on the large number (around 68%) of semi-/implausible versus plausible responders. In the large online questionnaire sample, it is reasonable to assume that the majority of responses are valid.

Figure 3.2 presents histograms for selected observed variable with their corresponding kernel density estimates and normal density curves based on ML estimates of mean and variance. Most of the univariate histograms for the observed variables in both study samples are represented by the histograms in the first column of Figure 3.2. The majority of univariate histograms and kernel density estimates suggest that a normal distribution can capture most important characteristics of the distributions of the observed variables. Especially, a consistent approximate normality assumption for all observed variables seems to be a sensibly parsimonious choice. We can see in the representative example shown in the first column of Figure 3.2, for the indicator variable x_6 (`VUL 060`) of the latent factor y_1 (N) that the kernel density estimates are moderately reproducible by a normal distribution with mean and variance estimated from the data. Even in the experimental data sample, with a majority of semi/-implausible response patterns, both density curves can be sufficiently approximated using a normal distribution. For the online questionnaire data, the middle category is not chosen as frequently as we would expect for approximately normal distributed variables. However, we can see that the kernel density estimates produces unimodal curves which are declining on both ends.

Histograms shown in the second and third column of Figure 3.2 are selected for print because these represent the two most extreme cases throughout all observed


(a) Experimental study sample

(i) representative ex. (x_6)

(ii) extreme case: x_{15}

(iii) extreme case: x_9

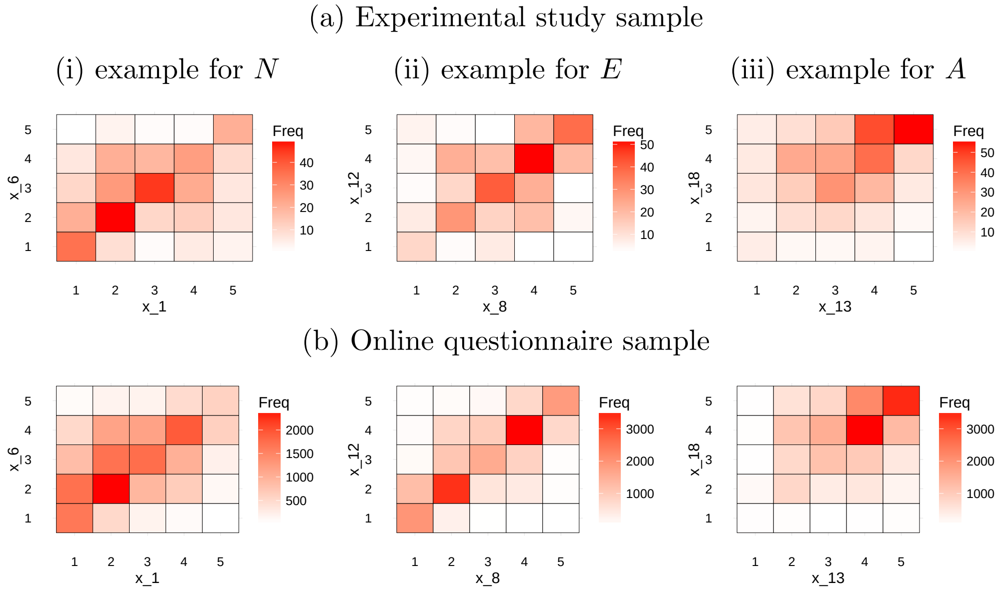
(b) Online questionnaire sample

Figure 3.2: Histograms for a representative observed variable (i) and selected observed variables which are the two extreme cases (ii) and (iii), for the experimental study and online questionnaire samples with their corresponding normal density curves and kernel density estimates.

variables in both study samples. In the second column, we see an extreme case, x_{15} of the latent factor y_3, where the observed variable shows some degree of skewness. The left-skewed data in (ii) might be due to the presence of a large number of invalid responses. However, we can see a similarly left-skewed distribution for (ii) in the online questionnaire samples, suggesting that this deviation from normal is item specific. Another extreme case is shown in the third column of Figure 3.2 for observed variable x_9 of the latent factor y_2. For the experimental study sample, we have an approximately normal kernel density curve. However, the middle category is not as pronounced as it would be expected if it was approximately normal distributed. This is even more extreme for the corresponding variable (iii) based on the online questionnaire sample: the histogram suggests a bi-modal distribution.

Secondly, Figure 3.3 shows several equivalents of bivariate scatter plots for two ordinal categorical variables (mosaic plot). There are three plots for each study sample for three pairs of two selected variables, where each pair of variables is a pair of indicator variables for the same latent variable. I selected the same variables for both study samples, and these were chosen such that they are representative for (visually most similar to) the remaining 14 combinations of indicator variables within the same factor. The cell frequencies are indicated by their colour similar to a heat map, ranging from white (lowest frequency) to red (highest frequency). When we look for bivariate normality in bivariate scatter plots for continuous variables, we would usually like to identify an elliptic shape, where points should be densest in the epicentre and become less dense the further away they are from the epicentre. In this ordinal mosaic plot, we would expect the equivalent form of an elliptic shape such that, e.g. one cell is the densest (dark red) and the next densest cells (fading red) would be at the adjacent top-right corner and bottom-left corner (positive relationship). Further away from the epicentre we would expect white or strongly fading red cells for the ordinal equivalent of a bivariate normal distribution. In the experimental study sample, we can see slightly elliptic colour patterns for the selected variable pairs x_1 and x_6 (indicator variables for N) and x_8 and x_{12} (indicator variables for E). In (ii), we can see a clear epicentre where in (i) we have two similarly dense red cells. The corresponding variable pairs for the online questionnaire sample, are similarly elliptic. However, the densest cells seem to be at the corners of (surrounding) a less dense cell, which we would expect to be the epicentre. This is in line with the slightly bimodal tendencies for the online questionnaire sample, which we have seen in some univariate histograms in Figure 3.2. Lastly, the bivariate mosaic plots for observed variables x_{13} and x_{18} (indicator variables for latent variable A) serve as examples for variables that are left-skewed. We can see that the colour pattern is similar to an elliptic shape, where the top right corner is truncated.

Thirdly, a good indicator of multivariate normality is the Mahalanobis distance. Given multivariate normality, we expect the empirical values to be approximately χ^2 distributed with degrees of freedom $df = 18$ (number of observed variables). The analogous QQ-plots for the $D_i^2(\Sigma)$ as defined in (2.29) can be found in Figure 3.4. Empirical values based on $D_i^2(\hat{\Sigma})$ are plotted against the theoretical quantiles of


(a) Experimental study sample

(i) example for N (ii) example for E (iii) example for A

(b) Online questionnaire sample

Figure 3.3: Bivariate mosaic plots for example indicator variables for each latent variable, for the experimental study and online questionnaire samples, where cell frequencies are represented by colours (heat map).

the respective χ^2 distribution, on the abscissa. The first plot is based on the entire experimental study sample, whereas the second plot shows the results for the plausible sub-sample. The last plot shows respective quantiles for the online questionnaire data. Conceivably, $D_i^2(\hat{\Sigma})$ for the experimental study sample with predominantly

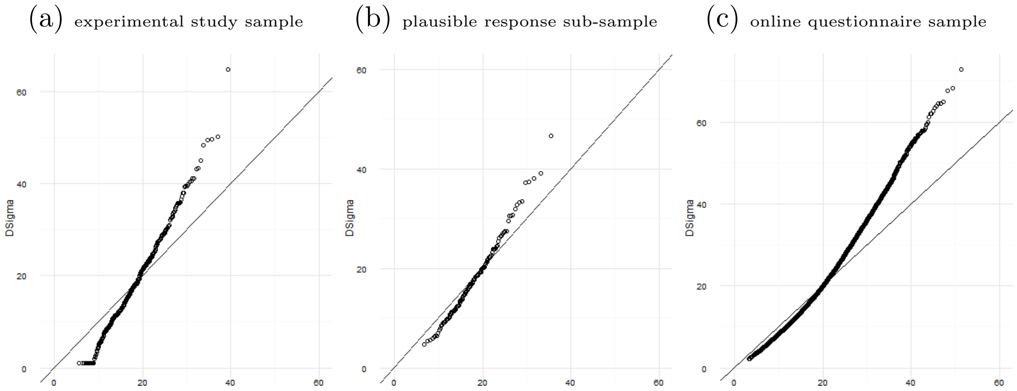

(a) experimental study sample (b) plausible response sub-sample (c) online questionnaire sample

Figure 3.4: QQ-Plots for the Mahalanobis distance for the experimental study sample, the plausible response sub-sample and online questionnaire sample against the theoretical $\chi^2(18)$ distribution

semi-/implausible responses does not follow the theoretical distribution. These results can be a combination of invalid responses and, consequently, measurement error in the $\hat{\Sigma}$ estimate. Hence, we should take a closer look at the QQ-plot for the plausible sub-sample only. Most points in the plot follow the diagonal line with some departures at the more extreme quantiles. These results suggest approximate multivariate normality for the observed variables for valid responses only. The latter plot for the online questionnaire data follows the findings for the univariate distributions: skewed data for the observed variables exhibit many more extreme values in $D_i^2(\hat{\Sigma})$ then we would expect under the theoretical χ^2 distribution. These results will be integrated and further discussed in Section 5.3.1.

Lastly, there are several test measures available which vary in their sensitivity towards sample size. Univariate normality tests can be performed with, e.g. Kolmogorov-Smirnov/Lilliefors test statistics (Kolmogrov, 1933; Lilliefors, 1967). Lilliefor's test for univariate normality lead to unambiguous conclusions: The null hypothesis of normality is consistently rejected ($p < .01$) for all observed variables, for both study samples, as well as the experimental plausible and semi-/implausible sub-samples. A procedure for a multivariate normality test was proposed by Mardia (1970), which compares empirical and expected values for multivariate extensions of skewness and kurtosis. If the empirical distributions show a good fit (i.e. individual test results and visual comparisons) to the theoretical distributions, then this would suggest that the assumption of multivariate normality for the observed variables is met. Mardia's multivariate normality test also rejects the null hypothesis for all samples ($p < .01$).

Comparability of the two Study Samples

To compare the analysis results in the next sections for the experimental study and online questionnaire, I shall establish a degree of comparability of valid responses in both study samples. For this purpose, Table 3.5 summarises mean and variance results of the observed variables for the experimental study and its sub-samples as well as the online questionnaire data. The first row gives averaged values over observed variables' means and observed variables' standard deviations. The second row lists standard deviations of the individual summary statistics. Both study samples show similar overall mean values around 3.28 with a standard deviation of mean values around 1.19. However, observed variables' standard deviations are smaller for the experimental study sample with mean 0.36 ($SD = 0.06$) in comparison to the online questionnaire results with mean 0.50 ($SD = 0.14$). The latter results, for the online questionnaire study, are similar to the results for the plausible sub-sample of the experimental study. However, the overall mean 3.44 is slightly larger for the plausible sub-sample. The semi-/implausible sub-sample has the smallest overall mean 3.20 and overall standard deviation 0.30. The predominantly semi-/implausible experimental study sample is accordingly heavily impacted by invalid responses. The larger differences between the plausible sub-sample and the online questionnaire data could be the result of a similar mixed composition of the latter: the online questionnaire sample may consist of a mixture of valid and invalid responses, as well.

Table 3.5: Mean and standard deviations of observed variables' means and standard deviations for the sub-groups of the experimental study and online questionnaire samples

	Experimental sub-samples						Online questionnaire	
	All		Plausible		Semi-/impl.			
	$\bar{x}_j$	s_j	$\bar{x}_j$	s_j	$\bar{x}_j$	s_j	$\bar{x}_j$	s_j
Mean	3.28	0.36	3.44	0.53	3.20	0.30	3.29	0.50
SD	1.23	0.06	1.17	0.15	1.23	0.05	1.19	0.14

Conclusions

In summary, graphical illustrations of the univariate distributions show that normal distributions can capture the most important characteristics of the empirical distributions, but univariate normality tests consistently rejected the normality assumptions. However, the non-normal characteristics such as the tendency to left-skewed distributions are consistent between both study samples. It is reasonable to assume that findings here are comparable to distributional characteristics of the respective IPIP items and latent variables analyses in the literature. Therefore and although the data at hand might no be perfectly suited for above latent variable model definitions, results between both study samples at hand and between those and samples in previous literature allow comparisons, because these model assumptions are consistently made with similar prerequisites. Comparing different analyses results based on different samples is critical for the exploratory investigation of semi-plausible response patterns.

3.3.3 Goodness of Fit

The theoretical latent variable model was fitted to the experimental study sample in total. The experimental sub-group membership was ignored entirely assuming the sample consists of valid responses only. The same model was fitted to the online questionnaire sample, as well.

The sample sizes throughout the different sub-groups of analysis are larger than the number of observed variables and, hence, a minimum requirement for model estimation met (MacCallum, Browne, and Sugawara, 1996). No convergence problems, Heywood cases nor negative variance estimates or the like occurred in or as result of the estimation process. Thus, further standard requirements for measurement models with moderate to small sample sizes are met (Chen, Bollen, Paxton, Curran, and Kirby, 2001).

Before I report and discuss model parameter estimates, I shall evaluate the overall model fit indices for both samples. However, interpretations based on rule-of-thumb cut-off criteria are arbitrary and should not be taken too seriously. Conclusions drawn by model fit indices can also be the result of model misspecification, small-sample bias, effects of a violation of normality and independence, and estimation method effects (Hu and Bentler, 1998). Table 3.6 on page 78 summarises a selection of model fit indices retrieved from the Mplus output file.

Test of Goodness-of-Fit

The χ^2-Test statistic allows for an inferential judgement about whether the model implied covariance matrix is significantly different from the unconstrained sample covariance matrix (i.e. the covariance matrix from a saturated model). For the online questionnaire study sample, we have $\chi^2 = 11857$ and $\chi^2 = 524$ for the experimental study sample, respectively. Both indicate highly significant values ($p < .01$) with

$df = 132$ degrees of freedom. These usually are interpreted as poor model fit to the data (e.g., Schermelleh-Engel, Moosbrugger, and Müller, 2003). Hence, we would reject the null hypothesis that the model is correct, or more precisely, that the model can reproduce the observed covariance matrix. Large sample sizes tend to produce large χ^2 values and vice versa. Hence, it seems sensible to not only rely on this statistics for the evaluation of the model fit for the online questionnaire sample with a sample size of $n = 20993$. However, for the experimental study sample with a total sample size of $n_{\text{total}} = 380$, there is no convincing argument to render our inference as strongly impacted by sample size. The test for the null model is also significant with $p < .01$ for both samples with values $\chi^2 = 112657$ and $\chi^2 = 1466$ for $df = 153$ for the online questionnaire and experimental study sample, respectively.

Descriptive Model-Fit Indices

χ^2 statistic can further be used as a descriptive goodness-of-fit index by setting the value in relation to the number of degrees of freedom (Jöreskog and Sörbom, 1993). The ratios χ^2/df for both samples are greater than 2, which according to Jöreskog and Sörbom (1993) indicates a bad model fit.

A descriptive measure which is regarded as relatively independent of sample size is the *Root Mean Square Error of Approximation* (RMSEA; Steiger, 1990). It is a measure of approximate fit in the population and taking account for the discrepancy due to approximation. The RMSEA is usually in favour of more parsimonious models (e.g., Browne and Cudeck, 1993; Kaplan, 2009). Smaller values indicate a better fit. The estimates for the experimental study sample is RMSEA = .088 with a 90% confidence interval $CI_{\text{RMSEA}} = [.081; .096]$ and the online questionnaire study sample with RMSEA = .065 and $CI_{\text{RMSEA}} = [.064; .066]$ indicate an acceptable fit based on the criteria RMSEA $< .10$. In the case of the online questionnaire study sample, RMSEA $\leq .8$ can be interpreted as mediocre model fit (MacCallum et al., 1996). A CI lower boundary with values smaller than .05 would have indicated a good model fit (Schermelleh-Engel et al., 2003).

Another descriptive measure is the *Standardised Root Mean Square Residual* (SRMR; Bentler, 1995), which is an overall badness-of-fit measure that is based on the standardised residual matrix. The SRMR values for both samples exceed the criteria SRMR $<$ .05 for a good fit. However, both values .06 for the online questionnaire and .091 for the experimental study meet the requirement SRMR $<$ .10 for an acceptable fit (Hu and Bentler, 1995). Other sources also accept values $\leq$ .08 as good model fit, which is fulfilled by the model based on the online questionnaire sample (Hu and Bentler, 1999).

Comparative Model-Fit Indices

The *Comparative Fit Index* (CFI; Bentler, 1990) and *Tucker-Lewis Index* (TLI/N-NFI; Tucker and Lewis, 1973) are indices that compare the fit of a model of interest with the fit of some baseline model. These measures are also of purely descriptive nature. For both indices, the baseline model is the independence model which assumes that observed variables are uncorrelated. Furthermore, these indices are supposed to be relatively less sensitive to sample size, and they penalise less parsimonious models. Both indices generally range from 0 to 1, larger values indicating better fit. None of the CFI and TLI values for both samples are above the thresholds .95 or .97 and, according to literature, indicate a poor fit relative to the independence model (Schermelleh-Engel et al., 2003). Where online questionnaire values of CFI and TLI are close to a critical value of .9, in the case of the experimental study sample, comparative model fit indices CFI = .701 and TLI = .654 seem to be amongst the most affected by the high ratio of semi-/implausible relative to plausible response patterns.

Table 3.6: Model fit indices for different samples assuming valid responses only

Sample	Model fit indices					
	χ^2/df	RMSEA	CI_{RMSEA}	SRMR	TLI	CFI
Online questionnaire	89.828	.065	[.064;.066]	.060	.879	.896
Experimental study	3.970	.088	[.081;.096]	.091	.654	.701
Sub-sample						
Plausible	5.396	.104	[.089;.119]	.108	.717	.756
Semi-plausible	3.845	.074	[.056;.091]	.086	.724	.762
Implausible	4.193	.077	[.061;.093]	.084	.710	.749

Model-Fit Indices for Sub-samples

Table 3.6 also shows model-fit indices for the analyses with plausible, semi-plausible, and implausible sub-samples only. Where the model-fit indices are not directly comparable, we can see that there are no significant changes in comparison to model-fit indices based on the entire experimental study sample. However surprisingly, the fit indices for the plausible sub-sample show worse model fit in (rough) comparison to those indices based on the entire experimental study sample. In fact, the semi-plausible sub-sample seems to fit best to the respective model with medium correlated latent variables. It seems some participants even in the plausible conditions do not follow a distinct factor structure model as is estimated based on the plausible sub-sample.

Discussion

The sample size sensitive χ^2 model fit statistics give a significant result indicating poor model fit for both samples. For the model estimated with the online questionnaire sample, the fit indices tend towards a mediocre to good model fit. The fit indices produce sensible values for the evaluation of model fit.

Expectedly, model fit based on the experimental study sample resulted in acceptable model fit indicators, at best. However, amongst reported indices, CFI and TLI values were most affected by the existence of 68% for respondents for whom we might expect semi-/implausible response patterns in the sample. Under normal circumstances, these comparative indices suffer from a null model that is not too bad regarding fit to the data. In other words, the more variables with little correlation exist, the less accurate will CFI/TLI be able to evaluate overall model fit. Hence, it would be of interest to incorporate null model information into statistics developed for the detection of SpRPs.

The reader should be reminded that the validation of the model is of relatively low importance with regards to the general study objective. Even more so, in general, the usual interpretations of model fit can be arbitrary because most indices are based on rule-of-thumb cut-off criteria. As was mentioned before, fit indices are usually affected by model misspecification, small-sample bias, effects of a possible violation of normality and independence, and estimation method effects (Hu and Bentler, 1998). Therefore, it is always possible that a model may fit the data although one or more fit measures suggests bad fit. In light of this and because of the already strong empirical validation of the IPIP items and the Big Five framework, I will further assume a correct model specification.

3.3.4 Parameter Estimates

The methodology for the comparison of estimates throughout different sub-groups will be implemented as follows: First, free parameters as defined in Section 3.3.1 will be estimated using the online questionnaire sample of the Johnson (2005) study introduced in Section 3.2.3. These shall serve as anchors for estimating the model under usual conditions and with sufficient sample size. Secondly, the same estimation procedure will be applied separately to sub-groups in the experimental study, with Cell 1 and 4 combined (plausible response patterns), Cell 2 and 5 combined (semi-plausible response patterns), and Cell 3 and 6 combined (implausible response patterns).

Online Sample

Figure 3.5 displays estimates of observed variables' residual variances, factor loadings as well as latent variable variances, means, and covariances for the online questionnaire sample of the Johnson (2005) study in a path diagram.

All factor loadings are highly significant on a $p \leq .01$ level and have the right sign in the hypothesised direction. Ranging between $[0.38; 1.10]$ such that participants who endorse questions x_1–x_3, x_7–x_9, and x_{13}–x_{15} are associated with larger values in their respective latent variable and vice versa on observed variables x_4–x_6, x_{10}–x_{12}, and x_{16}–x_{18}. Regarding the latent variable parameters, we have highly significant negative covariances (correlations) between factor N and each other latent variable, where $\hat{\phi}_{2,1} = -.24$ is slightly higher than $\hat{\phi}_{3,1} = -.17$, due to N and E's affiliation to the Alpha-Factor. Also as expected, the remaining covariance between E and A is positive and, although significant, negligible with $\hat{\phi}_{3,2} = .02$. All estimated parameters have very small standard errors due to the large sample size (≤ 0.01).

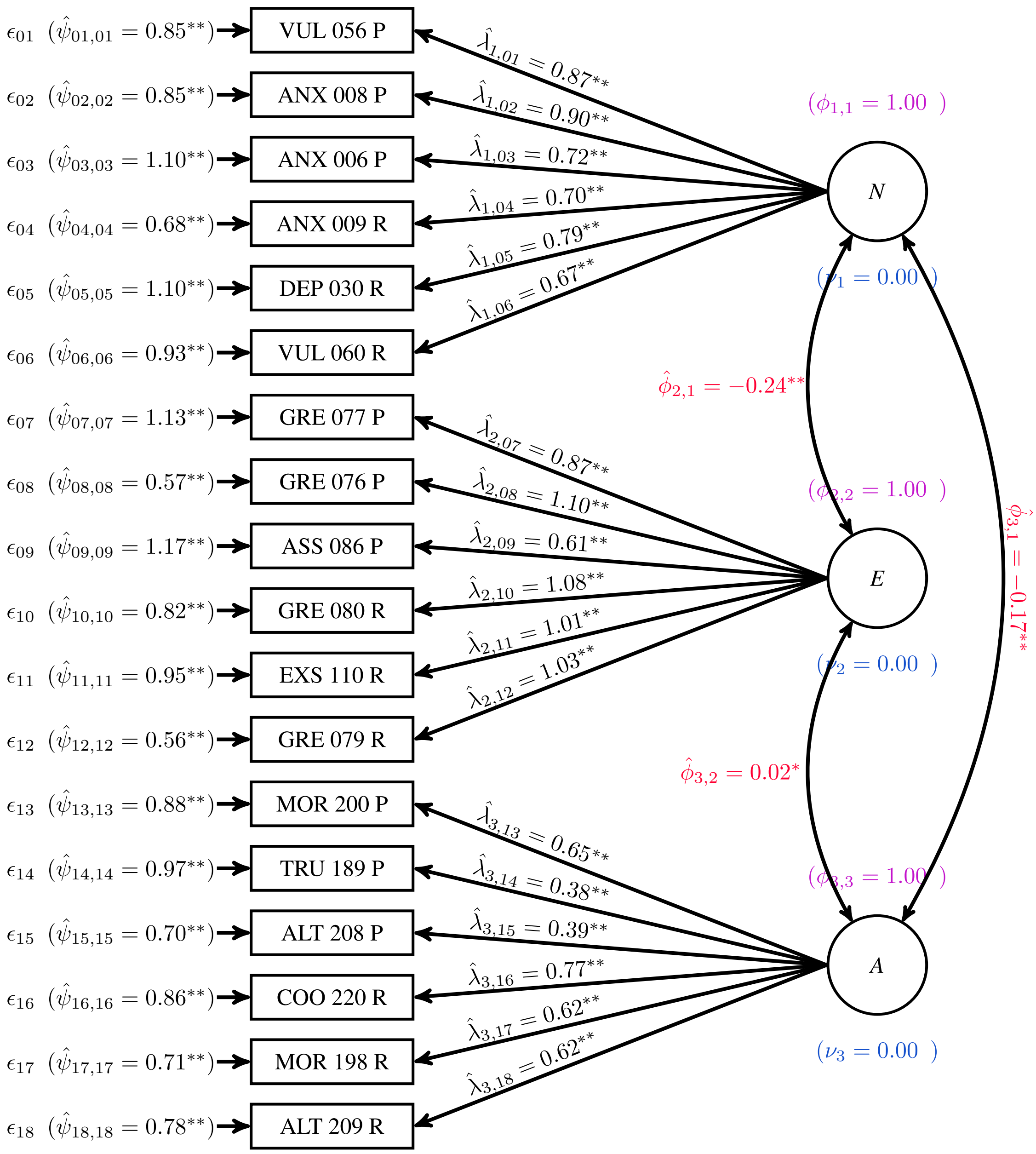

ϵ_{01} ($\hat{\psi}_{01,01} = 0.85^{**}$)
ϵ_{02} ($\hat{\psi}_{02,02} = 0.85^{**}$)
ϵ_{03} ($\hat{\psi}_{03,03} = 1.10^{**}$)
ϵ_{04} ($\hat{\psi}_{04,04} = 0.68^{**}$)
ϵ_{05} ($\hat{\psi}_{05,05} = 1.10^{**}$)
ϵ_{06} ($\hat{\psi}_{06,06} = 0.93^{**}$)
ϵ_{07} ($\hat{\psi}_{07,07} = 1.13^{**}$)
ϵ_{08} ($\hat{\psi}_{08,08} = 0.57^{**}$)
ϵ_{09} ($\hat{\psi}_{09,09} = 1.17^{**}$)
ϵ_{10} ($\hat{\psi}_{10,10} = 0.82^{**}$)
ϵ_{11} ($\hat{\psi}_{11,11} = 0.95^{**}$)
ϵ_{12} ($\hat{\psi}_{12,12} = 0.56^{**}$)
ϵ_{13} ($\hat{\psi}_{13,13} = 0.88^{**}$)
ϵ_{14} ($\hat{\psi}_{14,14} = 0.97^{**}$)
ϵ_{15} ($\hat{\psi}_{15,15} = 0.70^{**}$)
ϵ_{16} ($\hat{\psi}_{16,16} = 0.86^{**}$)
ϵ_{17} ($\hat{\psi}_{17,17} = 0.71^{**}$)
ϵ_{18} ($\hat{\psi}_{18,18} = 0.78^{**}$)
VUL 056 P
ANX 008 P
ANX 006 P
ANX 009 R
DEP 030 R
VUL 060 R
GRE 077 P
GRE 076 P
ASS 086 P
GRE 080 R
EXS 110 R
GRE 079 R
MOR 200 P
TRU 189 P
ALT 208 P
COO 220 R
MOR 198 R
ALT 209 R
$\hat{\lambda}_{1,01} = 0.87^{**}$
$\hat{\lambda}_{1,02} = 0.90^{**}$
$\hat{\lambda}_{1,03} = 0.72^{**}$
$\hat{\lambda}_{1,04} = 0.70^{**}$
$\hat{\lambda}_{1,05} = 0.79^{**}$
$\hat{\lambda}_{1,06} = 0.67^{**}$
$\hat{\lambda}_{2,07} = 0.87^{**}$
$\hat{\lambda}_{2,08} = 1.10^{**}$
$\hat{\lambda}_{2,09} = 0.61^{**}$
$\hat{\lambda}_{2,10} = 1.08^{**}$
$\hat{\lambda}_{2,11} = 1.01^{**}$
$\hat{\lambda}_{2,12} = 1.03^{**}$
$\hat{\lambda}_{3,13} = 0.65^{**}$
$\hat{\lambda}_{3,14} = 0.38^{**}$
$\hat{\lambda}_{3,15} = 0.39^{**}$
$\hat{\lambda}_{3,16} = 0.77^{**}$
$\hat{\lambda}_{3,17} = 0.62^{**}$
$\hat{\lambda}_{3,18} = 0.62^{**}$
($\phi_{1,1} = 1.00$)
N
($\nu_1 = 0.00$)
$\hat{\phi}_{2,1} = -0.24^{**}$
($\phi_{2,2} = 1.00$)
E
($\nu_2 = 0.00$)
$\hat{\phi}_{3,1} = -0.17^{**}$
$\hat{\phi}_{3,2} = 0.02^{*}$
($\phi_{3,3} = 1.00$)
A
($\nu_3 = 0.00$)

Figure 3.5: Big Three factors model path diagram with parameter estimates for the online questionnaire sample.

Experimental Sub-Samples

Table 3.7 summarises selected parameter estimates for three further analyses with the same model as discussed previously but each experimental sub-group serving as a separate sample for the parameter estimation. Furthermore, estimates are based on standardised observed variables with $\sigma_j = 1$ and $\mu_j = 0$. In doing so, it is easier to compare between sub-groups throughout all estimates. First, the goal is to investigate the model under optimal conditions with purely valid responses (cells 1 and 4) and compare these estimates with estimates from a sample that mainly consists of semi-plausible response patterns (cells 2 and 5). Secondly, we would like to explore results for a sample that produced responses by only drawing upon invalid response strategies (cells 3 and 6). Lastly, the first column also reports estimates for the analysis in the previous section with the online questionnaire data that are also based on standardised observed variables. As such we can compare estimates for plausible responses only sample with the online questionnaire data.

Table 3.8 simplifies the comparison further by giving summary statistics for the three different parameter types: factor loadings, factor covariances, and residual variances. Minimum, mean, and maximum absolute differences between two corresponding parameter estimates based on the plausible sub-group and each of the three other sub-groups are listed.

First, I shall compare the estimation for the plausible condition and the estimation based on the online questionnaire data. We would like to regard estimates for the online questionnaire sample as being closest to the true parameters for the model. Hence, small differences are a reassuring fact of the validity of the plausible sub-group as well as the online questionnaire sample. Mean difference between factor loadings, covariances, and residual variances are close to the mean standard error estimates of the corresponding parameters for the plausible sub-group. Hence, apart from some larger absolute differences (e.g., the maximum difference of .25 for residual variances), we can cautiously assume that analyses have led to similar estimated parameter values. Secondly, I shall further investigate differences between the plausible and the semi-plausible conditions. Here, we have larger differences between estimates based on the two sub-groups. Absolute differences between factor correlations are amongst the most apparent, ranging from at least .34 to .56. Whereas, estimated factor covariances for the first are highly significant but small enough to represent distinct factors and match with the statistical properties of the three included Big Five latent variable constructs reported in the literature. The latter sample leads to larger estimated values of the factor covariances. However, in comparison with the plausible

Table 3.7: Parameter estimates and standard errors for different samples

		Experimental study sub-samples		
	Online questionnaire	Plausible cells 1&4	Semi-plausible cells 2&5	Implausible cells 3&6
Factor loadings		$n = 123$	$n = 121$	$n = 136$
$\hat{\lambda}_{1,1}$	0.68 (0.01)**	0.76 (0.05)**	0.51 (0.11)**	0.55 (0.07)**
.	0.70 (0.01)**	0.59 (0.07)**	0.60 (0.09)**	0.57 (0.07)**
.	0.57 (0.01)**	0.53 (0.08)**	0.23 (0.11)**	-0.32 (0.09)**
.	0.65 (0.01)**	0.66 (0.07)**	0.32 (0.11)**	0.48 (0.08)**
.	0.60 (0.01)**	0.55 (0.08)**	0.58 (0.10)**	0.57 (0.07)**
$\hat{\lambda}_{1,6}$	0.57 (0.01)**	0.68 (0.06)**	0.46 (0.12)**	0.58 (0.07)**
$\hat{\lambda}_{2,7}$	0.63 (0.01)**	0.50 (0.09)**	0.38 (0.11)**	-0.18 (0.08)**
.	0.82 (0.01)**	0.80 (0.06)**	0.47 (0.09)**	0.35 (0.08)**
.	0.49 (0.01)**	0.34 (0.10)**	0.24 (0.10)**	0.50 (0.08)**
.	0.76 (0.01)**	0.58 (0.08)**	0.48 (0.10)**	-0.32 (0.08)**
.	0.72 (0.01)**	0.64 (0.08)**	0.47 (0.09)**	-0.22 (0.09)**
$\hat{\lambda}_{2,12}$	0.81 (0.01)**	0.82 (0.06)**	0.67 (0.09)**	0.50 (0.07)**
$\hat{\lambda}_{3,13}$	0.57 (0.01)**	0.71 (0.06)**	0.56 (0.08)**	0.38 (0.09)**
.	0.36 (0.01)**	0.23 (0.10)**	0.49 (0.08)**	-0.24 (0.10)**
.	0.42 (0.01)**	0.47 (0.08)**	0.65 (0.07)**	0.57 (0.08)**
.	0.64 (0.01)**	0.78 (0.06)**	0.60 (0.07)**	0.55 (0.08)**
.	0.60 (0.01)**	0.64 (0.07)**	0.58 (0.08)**	0.52 (0.08)**
$\hat{\lambda}_{3,18}$	0.57 (0.01)**	0.72 (0.06)**	0.54 (0.08)**	0.54 (0.08)**
Factor covariances/correlations				
$\hat{\phi}_{2,1}$	-0.24 (0.01)**	-0.15 (0.11)**	-0.49 (0.13)**	-1.14 (0.10)**
$\hat{\phi}_{3,1}$	-0.17 (0.01)**	-0.16 (0.11)**	-0.57 (0.12)**	-0.78 (0.09)**
$\hat{\phi}_{3,2}$	0.02 (0.01)**	0.12 (0.11)**	0.68 (0.10)**	0.95 (0.11)**

▨ Denoting the direction of the original coding. Those indicated with grey row color were recoded so that items were all positively coded.

data results, the semi-plausible data reveals clearly that the independent latent variable structure is no further a valid representation of data that predominantly consists of semi-plausible response patterns. The strongly negative and positive inter-correlations between latent variables suggest a more global factor structure

Table 3.8: Absolute differences between parameter estimates of different samples summarised for different types of parameters

Contrast	Absolute summary statistics for type of parameter								
Plausible	Factor loadings			Factor covariances			Residual variances		
vs.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
Online data	.01	.09	.18	.01	.07	.10	.01	.11	.25
Semi-plausible	.01	.17	.34	.34	.44	.56	.02	.19	.42
Implausible	.02	.34	.90	.62	.81	.99	.01	.22	.52

than the three distinct factors that are expected under valid response patterns.

Lastly, comparing the plausible versus implausible conditions reveals the most distorted and least consistent picture. Differences are large and reach a maximum absolute difference of .99. However, residual variances seem to differ in a similar way as the semi-plausible condition differs from the plausible condition. Each factor in the implausible condition has at least one-factor loading with shifted (negative) sign. This might be caused by the existence of recoded observed variables and, hence, revealing that some invalid response strategies are predominantly independent of the actual question content. Most interestingly, the implausible data does not seem to fit the structural model specification well. Factor covariances are estimated such that values have the right sign: negative covariances with N, $\hat{\phi}_{2,1} = -1.139$ and $\hat{\phi}_{3,1} = -0.775$ and positive for $\hat{\phi}_{3,2} = 0.951$. However, these covariances are unexpectedly large and, hence, inconsistent with the structural model specification for independent latent variables. The employed analysis software also provides a warning about possible linear dependence between latent variables N and E. The factor covariance matrix is not positive definite. A manifestation of this problem can further be seen by thoroughly interpreting the covariance $\hat{\phi}_{2,1} = -1.139$, which is hardly interpretable given the fact that both latent variables are defined to have variances $\phi_{1,1} = \phi_{2,2} = 1$. Such results are often referred to as Heywood cases. The iterative maximum likelihood estimation method converged to a numerical parameter solution that is smaller than a reasonable lower-bound value of -1 for a covariance that is defined as correlation. Heywood cases can occur when the model is misspecified, e.g. too many latent variables extracted. Given the large correlations between all of the three factors, the results suggest a misspecified single-factor model. However, even though the model for implausible response patterns suggests severe misspecification, we still have negative covariance estimates associated with the latent variable N. It suggests that even though invalid response strategies do not seem to be represented well by the three distinct factor model, altogether these implausible response patterns do tend to show some plausible tendencies with regards to the valid response model.

3.3.5 Discussion

In this section, I discussed analysis results for a three latent variable model with simple factor structure based on different analysis samples. I compared results for the online questionnaire study sample with results based on the experimental study sample. I further compared standardised estimates of model parameters for different sub-groups of the experimental study sample. Model fit indices suggest a bad fit for the experimental study sample and a good to mediocre fit for the online questionnaire sample. In general, parameter estimates based on the online questionnaire sample were in line with information reported in previous studies which used the IPIP to assess the Big Five personality factors. The experimental study sample produced sensible and interpretable results, but those were far off from what we would expect. Especially, results obtained using only implausible group members as sample lead to estimates that are largely opposite to the theory. Factor loadings of observed variables, which serve as reversed indicators for the latent variables, switched sign although these were recoded such that only positive factor loadings are expected. The theoretical distinct factor structure disappeared, even suggesting a one (single) global-factor solution as a better fit to the data. The ratio of explained versus residual variance in the model decreases the more SpRPs in the analysis sample. This pattern which is incongruent with outcomes cited in the literature is less severe but similar for the semi-plausible sub-group and also to be found in the analysis results based on the entire experimental study sample. However, said estimation bias pattern disappears when only the plausible sub-group is used as analysis sample.

Chapter 4

A Latent Class Model accommodating invalid Responses

In previous chapters, I have looked at the latent variable model that we would take to be true if there were no invalid responders present. Furthermore, I estimated model parameters for separate samples. However, in this section, I would like to illustrate the implications for a statistical model which does not meet these assumptions and introduce a multi-group model which allows for the presence of invalid responders. In a discussion of this approach I will focus on three essential questions:

(1) Are SpRPs also a function of the construct(s) of interest?

(2) How many different groups of invalid responders do we need to account for?

(3) Can we assume conditional independence of invalid responses?

Ultimately, I will draw conclusions from actual estimation under a latent class approach, and determine whether this approach is feasible in the context of SpRPs.

4.1 Multi-Group and Latent Class Models

I will continue with introducing a generic multi-group approach with extension to a latent class approach, where group membership is unobserved. This is followed by a brief reference to two related applications which were reviewed in Section 2.2.

4.1.1 The Multi-Group Model

Let η_z be the prior probability that a randomly chosen respondent i is in group $z = 0, \ldots, c-1$ with $\sum_{z=0}^{c-1} \eta_z = 1$ and $g(\cdot)$ denote probability density function of $\boldsymbol{x}$. Then in a multi-group model with observed z or in latent class model with unobserved z and c latent classes, we can write

$$g(\boldsymbol{x}) = \sum_{z=0}^{c-1} \eta_z \cdot g(\boldsymbol{x}|z), \tag{4.1}$$

where η_z is the mixture component for the different multivariate distributions of $\boldsymbol{x}$ given z. In a simple scenario with latent variables $\boldsymbol{y}$ we usually model a mixture design such that

$$\boldsymbol{x}|\boldsymbol{y}, z \sim N(\boldsymbol{\mu}^{(z)} + \Lambda^{(z)}\boldsymbol{y}, \Psi^{(z)}), \tag{4.2}$$

where we allow means, factor loadings, and error variances to vary between groups. (4.2) assumes that the different classes z are defined by essentially the same latent variables $\boldsymbol{y}$, however, potentially different parameter values.

Applying a multi-group model to our case needs further adaptation. So far, we only know that there are different groups of responders and that valid responses follow the latent variable model defined in Section 3.3.1. Assuming we have one group that represents valid responders ($z = 0$) and the corresponding probability η_0 and a group representing semi-plausible response patterns ($z = 1$) with $\eta_1 = (1 - \eta_0)$, therefore $c = 2$ groups, we can write

$$g(\boldsymbol{x}|\boldsymbol{y}, z; \boldsymbol{\theta}^{(z)}) = (1-z)\prod_{j=1}^{p} g_j(x_j|\boldsymbol{y}, z = 0; \theta_j^{(0)}) + z \cdot g(\boldsymbol{x}|\boldsymbol{y}, z = 1; \boldsymbol{\theta}^{(1)}), \tag{4.3}$$

where $\theta_j^{(0)}$ for $j = 1, \ldots, p$ and $\boldsymbol{\theta}^{(1)}$ are vectors of specific parameters for the measurement model for the items which are allowed to vary between groups or even represent an entire different set of parameters for any z. Here, we are assuming conditional independence of the x_j given $\boldsymbol{y}$ for the valid responders, as e.g. implied by the diagonal error covariance matrix in (3.3), but we are avoiding making that assumption for the invalid responders at this stage, returning to it in Section 4.1.4.

Following Rudas, Clogg, and Lindsay (1994), a mixture model similar to (4.3) can serve as an evaluation of overall model fit, where η_0 can be used as another form of model fit index. Hypothetically, if we were able to estimate above model and allow $g(\boldsymbol{x}|\boldsymbol{y}, z = 1; \boldsymbol{\theta}^{(1)})$ to be a representation of any invalid response, $\hat{\eta}_0$ would indicate how successfully we can represent the entire sample with the valid response model. Unfortunately, such a model is not generally identified. Therefore, it is necessary to make some assumption about $g(\boldsymbol{x}|\boldsymbol{y}, z = 1; \boldsymbol{\theta}^{(1)})$, hypothesising the statistical nature of invalid responses.

4.1.2 Independence of Construct(s) of Interest and invalid Responses

The first question that arises is whether we can assume for all of the observed variables x_j that

$$g(\boldsymbol{x}|\boldsymbol{y}, z = 1; \boldsymbol{\theta}^{(1)}) = g(\boldsymbol{x}|z = 1), \tag{4.4}$$

such that an invalid responders' response is not a function of the constructs of interest $\boldsymbol{y}$ at all. We might further hypothese that $g(\boldsymbol{x}|z = 1)$ follows a known distribution, for instance a multivariate uniform distribution, where we have random response patterns and for each $x_j \sim U(\min(u), \max(u))$, where u represents the answer categories to choose from.

The literature provides diverse attempts towards the integration of undesired response patterns into the statistical model. Many studies infer class membership from differences in the construct(s) of interest. Thereby these studies also assume that the construct(s) of interest not just have an influence on the valid responses but also affect invalid responses as well as class membership itself. I will briefly refer to two previously proposed approaches for integrative modelling and discuss these with regards to the independence assumption.

As introduced in Section 2.2, Moustaki and Knott (2014) integrates atypical response patterns into the model by allowing for a latent two-class system. $g(\boldsymbol{x}|\boldsymbol{y}, z = 1, \boldsymbol{x}^c; \boldsymbol{\theta}^{(1)})$ is formulated as a function of covariates $\boldsymbol{x}^c$ as well as the class depending on the latent variable. Therefore, it is assumed that there is an association between the probability of being an invalid responder and the latent variable presenting the construct of interest itself. In a comparable manner in Meade and Craig (2012), the same response model was defined for both classes of responders where only the parameters of the valid response model were allowed to vary between the classes. However, we could probably assume independence unless the construct of interest is a variable that might clearly or empirically proven to be related to the use of invalid response strategies (e.g., Big Five personality factor Conscientiousness, Schmitt et al., 1999).

4.1.3 Diversity of invalid Responders

The second important question is whether we can assume only one class of invalid responders or whether we need to assume $c > 2$, e.g. long string responders, responding as a function of positive versus negative *question wording*, and responding as a function of a graphically chosen response strategy. If there are several classes of invalid responders, the measurement model $g(\boldsymbol{x}|\boldsymbol{y}, z; \boldsymbol{\theta}^{(z)})$ is

$$\prod_{j=1}^{p} g_j(x_j|\boldsymbol{y}, z = 0; \theta_j^{(0)}) \tag{4.5}$$

for $z = 0$, and $g(\boldsymbol{x}|\boldsymbol{y}, z; \boldsymbol{\theta}^{(z)})$ for $z = 1, \ldots, c-1$. Invalid response patterns can be the result of numerous idiosyncratic response strategies. Hence, the only appropriate integrative approach would require more than two latent classes. For instance, taking account of a response strategy based on more or less random answers would ignore response strategies like long string responses. In theory, the number of latent classes that must be included would exceed the capacity of degrees of freedom necessary for model identification, based on the availability of response patterns resulting from each response strategy and the actual number of possible response strategies.

The extreme perspective is a scenario in which every invalid responder has his/her own unique response strategy. In this scenario $c-1$ would equal the number of these invalid responders and $\eta_z = (1 - \eta_0)/(c-1)$ for all $z = 1, \ldots, c-1$. Hence, each invalid response strategy alone would have little effect on the conditional distribution of $\boldsymbol{x}$. However, all $c-1$ invalid classes combined could have a severe impact.

The most likely scenario to which I would like to draw the reader's attention is halfway between the two extremes discussed previously. We can neither assume that semi-plausible response strategies are purely random (see Section 1.2), nor can we simply ignore invalid response strategies, assuming that they have little to no influence on the parameter estimation based on the assumption that each invalid responder has his or her unique response strategy. Even more so in the case of semi-plausible responses, I argue that there is not an unlimited number of response strategies from which a semi-plausible responder can choose. Hence, individual semi-plausible responses will most likely be the result of a small set of $c - 1$ invalid response types, potentially causing a significant amount of measurement error if not taken into account.

4.1.4 Conditional Independence and Method Factors

There is a further assumption that we can make about the distribution of invalid responses, further simplifying the estimation process. Once we have identified a possible invalid response strategy we may be able to represent it in a parsimonious way accounting for a latent (method) variable w (or a set of latent variables $\boldsymbol{w}$) that is not part of the constructs of interest $\boldsymbol{y}$ and is independent of $\boldsymbol{y}$. Capturing invalid response strategies in this way may justify the assumption of conditional independence of invalid responses such that the measurement model $g(\boldsymbol{x}|\boldsymbol{y}, w, z; \boldsymbol{\theta}^{(z)})$ is

$$\prod_{j=1}^{p} g_j(x_j|\boldsymbol{y}, z = 0; \theta_j^{(0)}) \tag{4.6}$$

for $z = 0$,

$$\prod_{j=1}^{p} g_j(x_j|w, z = 1; \theta_j^{(1)}) \tag{4.7}$$

for $z = 1$ where w is often referred to as method factor with parameter(s) $\theta_j^{(1)}$, and $g(\boldsymbol{x}|z > 1)$ for $z = 2$. Omitting the last part $g(\boldsymbol{x}|z > 1)$ can provide an identified model for further analysis. However, the reader should be aware that this implies the assumption that there are no other invalid response strategies present than those actually accounted for by the model, e.g., in this case, $g(\boldsymbol{x}|w, z = 1; \boldsymbol{\theta}^{(1)})$.

In the next section, I will give an example of such a latent class model for the data at hand as an attempt to accommodate invalid responses.

4.2 A Model incorporating a Method Factor

After having laid out a generic latent class framework, I will continue to define a possible latent class model extending the theoretical model in Section 3.3.1 in an attempt to accommodate the semi-plausible and implausible sub-groups of the experimental data. I will assume $g(\boldsymbol{x}|w, z = 1; \boldsymbol{\theta}^{(\mathbf{1})})$ to be a conditionally independent multivariate distribution given the method factor w, in a single invalid responders' class $z = 1$, and w independent from the constructs of interest $\boldsymbol{y}$. Further constraints on the invalid responders' model will be discussed in detail.

As discussed in the previous section, if invalid response strategies are not taken into account the conditional independence is not met when semi-plausible responders are not excluded from the sample. An inflation of spurious cross-correlations in the sample covariance matrix are to be expected based on any $g(\boldsymbol{x}|z \neq 0)$ that is not accounted for in the statistical model. In other words, we will see interdependencies between observed variables that are based on commonly used invalid response strategies. Despite the fact that in latent factor analysis random measurement error is more or less captured by the model, these spurious cross-correlations are of systematic nature and, hence, not accounted for if not specifically included.

In order to account for invalid response strategies, we need to hypothesise the statistical nature of $g(\boldsymbol{x}|z \neq 0)$. Results of separate estimations for sub-groups of the experimental data in Section 3.3.4 suggest that the three-factor model is not an appropriate representation of the invalid response patterns. The three latent variables were highly correlated suggesting a single global factor solution. Furthermore, measurement error estimates were larger and factor loadings smaller than the respective estimates using the Johnson (2005) sample (see Figure 3.5) or the valid responders' sub-sample of the experimental study (see Table 3.7).

A single global factor model shall serve as representative of the data of the class of $z = 1$ invalid responders. Furthermore, I restrict the number of $c - 1$ unknown latent classes representing invalid response strategies to one. This model is identified.

Let $w \sim N(0,1)$ denote a latent variable capturing the individual tendency of an invalid responder to favour a specific range of answer options independent from actual item content (e.g., question intend in survey) but as a function if item wording (i.e. responding to a more superficial layer of information through the wording of items). Therefore, we assume that invalid responders' responses are not a function of the actual constructs of interest ($\boldsymbol{y}$) and no relation between $\boldsymbol{y}$ and the individual tendency w. We can write

$$g(\boldsymbol{x}|\boldsymbol{y}, w, z; \boldsymbol{\theta}^{(z)}) = (1-z)\prod_{j=1}^{p} g_j(x_j|\boldsymbol{y}, z=0; \theta_j^{(0)}) + z\prod_{j=1}^{p} g_j(x_j|w, z=1; \theta_j^{(1)}). \tag{4.8}$$

Here $g(\boldsymbol{x}|\boldsymbol{y}, z = 0; \boldsymbol{\theta}^{(0)})$ follows the three-factor model defined in Section 3.3.1. However, I set different constraints on $g(\boldsymbol{x}|w, z = 1; \boldsymbol{\theta}^{(1)})$ and define the measurement model for invalid responses as follows:

$$\boldsymbol{x} = \boldsymbol{\mu}^{(1)} + \Lambda^{(1)} w + \boldsymbol{\epsilon}^{(1)} \tag{4.9}$$

$$\begin{pmatrix} x_1 \\ \vdots \\ x_3 \\ x_4 \\ \vdots \\ x_6 \\ \vdots \end{pmatrix} = \begin{pmatrix} \mu_{w,1} \\ \vdots \\ \mu_{w,1} \\ \mu_{w,2} \\ \vdots \\ \mu_{w,2} \\ \vdots \end{pmatrix} + \begin{pmatrix} \lambda_{w,1} \\ \vdots \\ \lambda_{w,1} \\ \lambda_{w,2} \\ \vdots \\ \lambda_{w,2} \\ \vdots \end{pmatrix} w + \begin{pmatrix} \epsilon_{w,1} \\ \vdots \\ \epsilon_{w,3} \\ \epsilon_{w,4} \\ \vdots \\ \epsilon_{w,6} \\ \vdots \end{pmatrix} \tag{4.10}$$

Restrictions are only applied on the vector of observed variable means $\boldsymbol{\mu}^{(1)}$ and on the vector of factor loadings $\Lambda^{(1)}$. Observed variables x_1, x_2, x_3, x_7, x_8, x_9, and x_{13}, x_{14}, x_{15} are assigned a single mean parameter $\mu_{w,1}$ and recoded observed variables x_4, x_5, x_6, x_{10}, x_{11}, x_{12}, and x_{16}, x_{17}, x_{18} are assigned $\mu_{w,2}$. Hence, their respective means are restricted to be equal within groups of recoded versus remaining observed variables. Analogously, we have only two factor loading parameters $\lambda_{w,1}$ and $\lambda_{w,2}$, where the latter restricts factor loadings for recoded observed variables to be equal and vice versa. This is equivalent to a scenario where half of the items are worded in the other direction than the remaining half of the items.

Hence in comparison to the valid responders' model, not each observed variable is assigned an individual intercept. Instead, we have only two intercept parameters, $\mu_{w,1}$ and $\mu_{w,2}$. This is because the hypothesised invalid response strategy does not entail capturing question content but merely assess whether questions have positive versus negative wording. Answer options are reversed with regards to item wording for recoded items. Allowing intercepts to be different for recoded versus unrecoded observed variables provides us with the opportunity to compare and judge whether we actually have $\mu_{w,1} = \mu_{w,2}$. In case $\mu_{w,1} - 1 = \max(u) - \mu_{w,2}$ the invalid answer strategy might only be the result of constantly picking the same numerical answer option (e.g., long string answer strategy). Following the same logic, we have only two factor loadings $\lambda_{w,1}$ and $\lambda_{w,2}$. Hypothesised answer strategy would be indicated by similar factor loadings, but more importantly, they should have the same sign. Opposite signing would, once again, indicate a simple long string strategy. Lastly, factor loadings associated with the latent variable w could also turn out to be 0. Because item content is hypothesised to be irrelevant for the invalid response strategy we have only a single latent variable w as method factor, indicating the individual tendency towards a specific, meaningful range of answer options (in contrast to the numeric representation of answer option in the original questionnaire). Constructs of interest and the method factor are assumed to be independent. Therefore, this method factor only applies to invalid responders, whereas observed variables serve as indicators for the three of the assessed Big Five personality factors only for the valid responders' group. Accommodating the invalid response strategy into the model may help to estimate parameters of the construct of interest more accurately. To be in line with the Big Five personality factors theory discussed in Section 3.1, we would expect significant factor loadings, weak correlation between latent variables y_1 and y_3 (Emotional Stability and Agreeableness are affiliated with the higher order Beta-Factor), and little to no correlation between those and y_2 (Extraversion is affiliated with the higher order Alpha-Factor). Lastly in contrast to the strict constraints for factor loadings and intercepts, the invalid responders' model still incorporates item specific error variances $\boldsymbol{\psi}^{(1)} = (\psi_1^{(1)}, \ldots, \psi_{18}^{(1)})$. In this way, we will be able to assess whether error variances are similar, which is what we would expect if question content does not matter apart from item wording.

Following this measurement model we can write

$$\Sigma^{(1)} = \Lambda^{(1)}\phi_w\Lambda^{(1)T} + \begin{pmatrix} \psi_{w,1} & & \\ & \ddots & \\ & & \psi_{w,18} \end{pmatrix} \tag{4.11}$$

for the model implied covariance matrix given $z = 1$, where we set $\phi_w = 1$ as variance for latent variable w.

By modelling a two latent class model, I do not only seek more accurate estimation for parameters of the constructs of interest. We will also have an estimate on the percentage of valid (η_0) versus invalid responders ($1 - \eta_0$) in the sample. Furthermore, we are provided with estimates about the probability of an individual being a member of any of the two classes by drawing on the posterior distribution of z. By simple application of the Bayes' theorem, we can write

$$\Pr(z = 1|\boldsymbol{x}_i) = \frac{(1 - \eta_0) \cdot g_i(\boldsymbol{x}_i|z = 1)}{g_i(\boldsymbol{x}_i)}, \tag{4.12}$$

for the probability of any i being member of the invalid responders' group $z = 1$ given the observed data defined by the estimated model parameters. In practice, a simple allocation rule may serve as an indicator for class membership such that individuals are placed in the class for which the posterior probability of class membership (given the response pattern and model parameters) is the greatest.

For easy access of the model specification, the path diagram for the two-class model is provided in Figure 4.1.

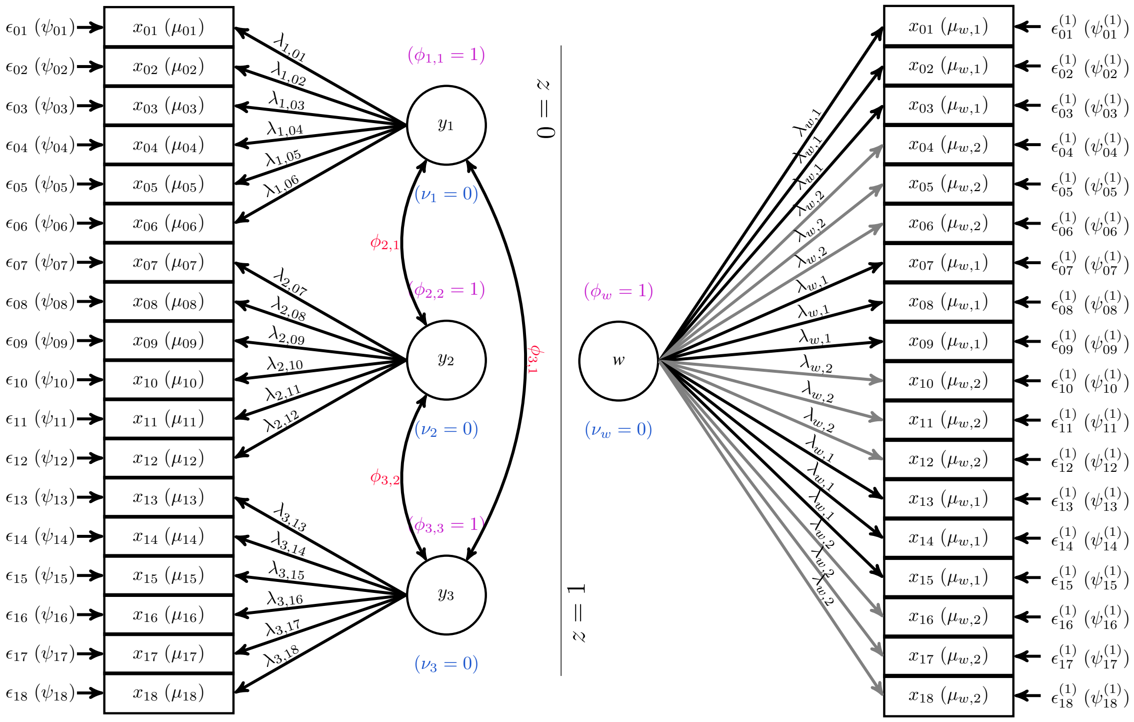


Figure 4.1: Path diagram for the two-class model with the valid ($z = 0$) and invalid ($z = 1$) response models.

4.3 Computational Implementation

A standard implementation of the analysis procedures for previously introduced latent class models is not straightforward or in many cases not at all possible, even in specialist software like Mplus. In Mplus it is not possible to define an entirely unrelated latent variable structures for the different manifestations of a latent class. For instance, the invalid response model must be defined in reference to the valid response model. Hence, I chose R as a more flexible programming language to analyse the latent class model with above specified valid and invalid class measurement models. However, implementing a more idiosyncratic analysis procedure is often error prone and can be very cumbersome. Hence, we should always test written code with the results of available software implementations using a simpler model design.

Fortunately, we can use a workaround to estimate above specified latent class model in Mplus. This is done by making use of mathematically equivalent model specifications when one measurement model is a special case of the other. For the invalid response measurement model, we can fix the latent trait variable covariances to be 1 (perfect correlation) ultimately emulating a single latent variable design. Furthermore, we set all factor loadings of reversed indicator variables to be equal and the remaining factor loadings to be equal, as well. Lastly, we would set intercepts for reversed and remaining indicators to be equal in the same manner. The software output will show some incorrect fit statistics because these are a function of the number of parameters in the model. However, the parameter estimates and the model log-likelihood value retrieved from the R implementation with correct model specifications were exactly replicated with Mplus using the workaround for the online questionnaire sample.

I will now continue to introduce the computational implementation of the latent class model using R. First, a basic numeric return function was coded, defining the negative log-likelihood function:

$$\begin{aligned} -\sum_{i=1}^{n}\{\ln[\quad \eta_0 \quad &\cdot\ g(\boldsymbol{x}_i|z=0;\boldsymbol{\mu}^{(0)},\Lambda^{(0)},\Phi^{(0)},\Psi^{(0)}) \quad + \\ (1-\eta_0) \quad &\cdot\ g(\boldsymbol{x}_i|z=1;\boldsymbol{\mu}^{(1)},\Lambda^{(1)},\Phi^{(1)},\Psi^{(1)}) \quad]\} \end{aligned} \tag{4.13}$$

This function takes values for parameters that we would like to estimate and returns the calculated negative log-likelihood value for the data given the parameters.

Secondly, we need to run a maximiser/optimiser algorithm to find parameter values that maximise values returned by the likelihood function. I chose the *optimx* package for R, which is a wrapper function for a variety of optimisation methods (see package optimx, Nash and Varadhan, 2011; Nash, 2014). optimx can handle multidimensional fitting problems. These procedures are very generic solutions for a wide range of problem scenarios and not specifically tailored to their use as analysis methods for latent class models. I chose the so-called *L-BFGS-B* method (Byrd, Nocedal, and Schnabel, 1994). BFGS is a variable metric method. There are numerous methods from which we can draw, but some others are generally more fragile. L-BFGS-B is a modification of the quasi-Newton method for limited-memory. The algorithm does not require analytic gradients, which is very helpful since it is easy to make mistakes when providing analytic gradients. However, the algorithm requires reasonable lower and upper bounds for parameter values (so-called *box* *constraints*) and starting values. Various starting values were tested and converged to the same solutions. A general disadvantage is that optimisations used for latent class analyses can become computationally very expensive. A further drawback is that L-BFGS-B always requires finite return values. Hence, lower and upper bound choices for parameters are to be chosen carefully. All methods require initial/starting values which must also satisfy constraints when lower and upper bounds are employed. There exists an alternative L-BFGS-B based wrapper function called *mle* which is part of the R base package *stats4*. mle can provide additional information, for instance, standard errors for parameters. This is done by retrieving an approximate covariance matrix for the parameters which is obtained by inverting the Hessian matrix at the optimum. However, the algorithm took much more time to converge. Because of this, I chose a two-step approach: First, I retrieve results from optimx and then input those as starting values for mle. This way we can validate the first solution and retrieve standard errors.

Lower and upper boundaries for parameters are chosen such that they do not violate model assumptions. Here, η_0 is the probability of valid group membership and as such has only support for $\eta_0 \in [0; 1]$. All variances are positive and residual variances of observed variables must not be larger than their respective total variances. Further constraints can be set based on easily acquirable sample information. Squared factor loadings can, in our case, not exceed the corresponding observed variable variances. This is because the latent variables are given a metric to have variance of 1 and the measurement model is based on simple factor loading structure. Means cannot be smaller or larger than the range of empirical values of the respective observed variables. Covariances (or, here, more accurately correlations) between standard normal latent variables can only take on values between -1 and $+1$. These constraints are only examples, which can be even further constrained based on properties of the defined model and its assumptions if the algorithm returns errors for non-finite likelihood values.

Furthermore, it is required to define a set of sensible initial values for the parameters to be estimated. The choice of starting values can be crucial for the outcome. In latent class models, local maxima of the likelihood function are amongst the most prevalent problems (Bartholomew et al., 2011; see also Aitkin, Anderson, and Hinde, 1981; Uebersax, 2000). Maximisation algorithms can converge on a set of parameter values without reaching the optimal solution. This is because most algorithms searching for a global maximum of the likelihood function rely on generic properties of maxima that cannot differentiate between local and global type. In general, it is recommended to employ several different sets of starting values and compare the results to check for local maxima solutions. The analysis of latent class models is often a follow-up step after having defined a global latent variable or another type of measurement model. Previously retrieved parameter estimates or sample statistics can be used as starting values as well as randomly generated parameter values not exceeding their respective lower and upper bounds. If the sample size is small, the model too large or over-parametrised, and the data very noisy, the global maximum should not be the only criterion to chose from different sets of results. In these cases, maxima values close to each other can be the mere result of bad data conditions and the choice should be enriched by theoretical deliberations of the study at hand. In our case, there were no different solutions based on numerous sets of randomly generated starting values (within box constraints). However, the algorithm frequently did not converge for some sets of starting values.

Lastly, standard errors and significance levels for parameters were retrieved using the R wrapper mle and model goodness-of-fit indices were calculated using various analysis results.

4.4 Comparing the different Analysis Results

After having defined the statistical mixture model design, I will analyse this model using the experimental data as a sample with unknown class membership. Some discussion will include the same analysis using the online questionnaire sample. Model fit statistics that allow for a comparison of the basic latent variable analysis assuming no invalid responses, and the latent class analysis, accommodating an invalid response strategy into the model, will be discussed. Furthermore, valid responders' model parameters in this latent class model will be compared with parameter estimates from the analyses in the previous section Section 3.3 (without accounting for invalid responders in the statistical model) where they provide insight about model validity. This should give us sufficiently exhaustive information about whether a mixture design can help to reduce the impact of SpRPs on parameter estimates for the constructs of interest. A comprehensive table of all the estimates can be found in Table in Appendix A.3. Lastly, we will use the analysed two latent class model to allocate individuals to either class and compare the percentage of flagged invalid responders in each of the experimental sub-samples.

Goodness of Fit

The most important question regarding model fit of the latent class models is whether the accommodation of an invalid response strategy into the model leads to a better representation of the data than was the case for a latent variable model assuming no invalid responses in Chapter 3. First, I will discuss model fit indices for the latent class model. However, we cannot directly compare them between our two models. For this purpose, I will, secondly, draw on information criteria to evaluate which model is a better fit to the data.

The model fit indices for the latent class model were calculated using the definitions given in Hu and Bentler (1999). All of the Mplus output for the latent variable model assuming valid responses only were reproduced first to validate the computations.

The calculation of many of the following statistics require the recovery of the overall model implied covariance matrix and mean values. Following the mixture model, the overall mean is defined as

$$\boldsymbol{\mu} = \eta_0 \cdot \boldsymbol{\mu}^{(0)} + (1-\eta_0) \cdot \boldsymbol{\mu}^{(1)}, \tag{4.14}$$

and the overall model implied covariance matrix can be written as

$$\begin{aligned} \Sigma = \qquad & \eta_0 \left[\Sigma^{(0)} + (\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu})^T\right] + \\ & (1-\eta_0) \left[\Sigma^{(1)} + (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu})^T\right]. \end{aligned} \tag{4.15}$$

The goodness-of-fit statistics that follow can be calculated using the usual formulas, once above stated Σ and its degrees of freedom (compared to the saturated model, i.e. the sample covariance matrix) are available.

I calculated the commonly used χ^2 statistic which is a likelihood ratio test between the constrained model using the results in (4.14) and (4.15) and the saturated model (using sample means and covariance matrix). For the online questionnaire study sample, we have $\chi^2 = 6296.734$ and $\chi^2 = 54.167$ for the experimental study sample, respectively. This is a highly significant value ($p < .01$) for the online questionnaire sample. However, for the experimental study, we have a non-significant result with a value close to $p \approx 1$. Results are based on $df = 107$ degrees of freedom because we have to estimate an additional number of 25 parameters for the invalid class measurement model, in comparison to the latent variable analysis model assuming no invalid responses. Based on these results, we would reject the null hypothesis that the two-class model can reproduce the observed covariance matrix in the case of the online questionnaire study sample. However, this is not the case for the experimental study sample. Reiterating the discussion of model fit, the χ^2 test is very sensitive towards sample size and can to produce significant results for large sample sizes.

The ratio χ^2/df for the online questionnaire sample is greater than 2 indicating a bad model fit. However, this is not the case for the experimental study where a value of $\chi^2/df = 0.506$ speaks for a good model fit. The RMSEA model fit index estimates for the experimental study sample is RMSEA = 0 with no confidence interval because the value is set to 0 by definition of the index. This is to be interpreted as a very good model fit. In the case of the online questionnaire study sample the result is RMSEA = .052 with a confidence interval of $CI_{\text{RMSEA}} = [.051; .054]$ in which the lower boundary is very close to the .05 mark which would indicate a good model fit. In both cases, we have better model fit results for the RMSEA than was the case when we did not account for invalid response patterns. This is even more so important considering that the RMSEA is usually in favour of more parsimonious models. The SRMR value for the online questionnaire sample no longer exceeds the criteria SRMR $< .08$ and, hence, speaks in favour of a good model fit. However, for the experimental study sample, we have SRMR $> .10$, which speaks for a bad model fit. It seems the overall badness-of-fit measure that is based on the standardised residual matrix leads to a (slightly) worse judgement of model fit for the latent class analysis than for the latent variable model assuming no invalid responses. The comparative fit indices TLI and CFI reveal better model fit for the latent class analysis model. Where values CFI = .92 and TLI = .95 for the online questionnaire are no longer below the critical value of .9 and even reach the threshold of .95 for good model fit, in the experimental study sample the corresponding values (both set to 1) indicate very good model fit. These indices are evaluating model fit relative to the independence model. Both were severely affected when we did not account for invalid responses in the experimental study sample. It seems we have successfully accounted for response patterns which led to CFI and TLI values around .67 in the latent variable model assuming no invalid responses.

Table 4.1: Model fit indices for the latent variable and latent class model for two different samples

Model fit indices	Sample			
	Online questionnaire		Experimental study	
	LVA	LCA	LVA	LCA
χ^2/df	89.828	58.848	3.970	0.506
RMSEA	.065	.052	.088	0
CI_{RMSEA}	[.064;.066]	[.051;.054]	[.081;.096]	0
SRMR	.060	.066	.091	.126
TLI	.879	.921	.654	1
CFI	.896	.945	.701	1

$^{\text{LVA}}$ Latent variable model assuming valid responses only.
$^{\text{LCA}}$ Latent class model accommodating an invalid response strategy.

To directly evaluate which model, the latent class or the latent variable model assuming no invalid responses is a better fit to the data, I will draw on so-called information criteria. The *Akaike Information Criterion* (AIC) is a measure of the goodness of fit of a model that adjusts for the number of estimated parameters and can be used to compare competing models that need not be nested. However, all calculations ought to be based on the same sample of data. The model with the smaller AIC value is regarded as the better fitting model. The AIC seeks to select the model which serves best as an approximation to reality (or the sample data). It also penalises a high number of estimated parameters and, hence, rewards parsimony. *Bayesian Information Criterion* (BIC) is comparable in form to the AIC with a larger penalty term for the number of parameters. Similar to the BIC, the SABIC places a penalty for adding parameters based on sample size based on $n^* = (n + 2)/24$ (Muthén and Muthén, 1998–2012). However, SABIC does not penalise as strongly as

Table 4.2: Model fit comparison indices for two study samples

Information criteria	Sample			
	Online questionnaire		Experimental study	
	LVA	LCA	LVA	LCA
Akaike (AIC)	1114618	1109107	21333	20913
Bayesian (BIC)	1115071	1109759	21558	21236
Sample-size adjusted BIC	1114890	1109499	21377	20976

$^{\text{LVA}}$ Latent variable model assuming valid responses only.
$^{\text{LCA}}$ Latent class model accommodating an invalid response strategy.

the BIC. The three model fit indices are estimated based on following definitions:

$$\text{AIC} = 2s - 2\ell \tag{4.16}$$

$$\text{BIC} = s \ln[n] - 2\ell \tag{4.17}$$

$$\text{SABIC} = s \ln[(n+2)/24] - 2\ell \tag{4.18}$$

where s is the number of parameters to be estimated, n is the sample size, and ℓ is the log-likelihood of the data under the model. Table 4.2 shows the AIC, BIC, and SABIC statistic, for the latent variable and latent class model fit for each dataset. The table entries for the basic latent variable analyses are based on the Mplus output. These entries were successfully reproduced with the same methods that are used for the computation of the table entries for the latent class analyses. We can see that although the latent class model has a larger number of parameters to be estimated, all three information criteria indicate the latent class model to better fit to the online questionnaire data. The same conclusions can be drawn when we compare the fit information criteria for the experimental study sample even after we allow for the increased complexity of the model. The latent class model is a better fit to the highly contaminated experimental study sample. It seems even for the experimental study, we were able to incorporate measurement error based on invalid responses as successfully into the model as it is the case for the online questionnaire.

Table 4.3 shows a selection of estimated parameter values and corresponding standard errors for the two different measurement models.

Table 4.3: Two-class model parameter estimates and standard errors for the experimental study sample

Parameter	Latent class		Parameter
	valid	invalid	
$\hat{\lambda}_{1,1}$	1.07 (0.10)**		
.	0.94 (0.10)**	-0.21 (0.04)**	$\hat{\lambda}_{w,1}$
.	0.51 (0.11)**		
.	1.03 (0.11)**		
.	0.74 (0.08)**	-0.17 (0.04)**	$\hat{\lambda}_{w,2}$
$\hat{\lambda}_{1,6}$	0.96 (0.09)**		
$\hat{\lambda}_{2,7}$	1.19 (0.11)**		
.	0.34 (0.12)**	-0.21 (0.04)**	$\hat{\lambda}_{w,1}$
.	0.23 (0.12)*		
.	1.11 (0.10)**		
.	0.88 (0.11)**	-0.17 (0.04)**	$\hat{\lambda}_{w,2}$
$\hat{\lambda}_{2,12}$	0.37 (0.11)**		
$\hat{\lambda}_{3,13}$	0.38 (0.09)**		
.	0.11 (0.11)	-0.21 (0.04)**	$\hat{\lambda}_{w,1}$
.	0.35 (0.06)**		
.	0.50 (0.07)**		
.	0.45 (0.06)**	-0.17 (0.04)**	$\hat{\lambda}_{w,2}$
$\hat{\lambda}_{3,18}$	0.38 (0.06)**		
$\hat{\phi}_{2,1}$	-0.07 (0.10)		
$\hat{\phi}_{3,1}$	-0.31 (0.09)**		
$\hat{\phi}_{3,2}$	0.00 (0.11)		
$\hat{\eta}_0$	0.40 (0.03)**	0.60 (0.03)**	$\hat{\eta}_1$

▨ Factor loadings for recoded variables.

Class Membership Probability

The estimated probability of a random participant being a member of the valid responder group, $\hat{\eta}_0 = .4$, seems to represent the percentage of plausible responders in the experimental data more accurately than expected. Cells 1 and 4 together represent the plausible responders in the experimental study sample, which equals

$(n_{\text{Cell 1}} + n_{\text{Cell 4}})/n = 32.37\%$ of the entire analysis sample. The respective estimate for the online questionnaire data is $\hat{\eta}_0 = .9$. This is a reassuring fact and shows that parameter estimates of the online questionnaire can serve as anchors for the evaluation of the accuracy of parameter estimates.

Factor Loadings for Method Factor

Factor loadings for recoded versus not recoded observed variables do not differ significantly and, most importantly, even have the same sign. This is validating the initial hypotheses as is the case for factor loadings for the valid response model which also have the same sign. Moreover, these findings substantiate a valid interpretation of w as method factor representing invalid responders' tendency to favour a certain range of meaningful (with regard to item wording) answer options, regardless of item content. Factor loadings for the invalid response model are small but significantly different from 0.

Intercepts for Invalid Class

In line with the findings about the factor loadings, the two intercept parameter estimates $\hat{\mu}_{w,1} = 3.11\ (0.03)^{**}$ and $\hat{\mu}_{w,2} = 3.13\ (0.03)^{**}$ are similar enough, such that they could be constraint to be equal. Hence, these intercepts represent a (positive) meaningful (with regards to item wording, not item content) middle answer category.

Error Variances

Error variances $\boldsymbol{\psi}^{(1)}$ have range $[1.20; 1.59]$ with mean 1.31 and standard deviation 0.13, hence, they are more alike than error variances $\boldsymbol{\psi}^{(0)}$ with respectively descriptive values $[0.13; 1.65]$, 0.80, and 0.48. These results substantiate the hypothesis of similar error variances regardless of item content for the invalid responders models and also validates the interpretation of w as method factor.

Factor Loadings and Intercepts

Factor loadings between different analyses models are not directly comparable because latent variables means and variances are fixed to 0 and 1, respectively. This applies to the mixture model as well as to the valid responders only model. However, factor loading patterns within designs can still give insight about the validity of estimates. In general, within each set of 6 indicator variables per latent variable, smaller loadings on parameters estimated with the online questionnaire sample are also amongst the smaller loading estimates of the experimental study sample and vice versa.

Factor loadings $\hat{\lambda}_{1,3} = 0.51$, $\hat{\lambda}_{2,9} = 0.23$, and $\hat{\lambda}_{3,14} = 0.11$ are the smallest in each set of indicators in the valid response model (latent class model). The same is true for the respective parameter estimates for latent variables models based on the online questionnaire sample (0.57/0.49/0.36, see Table 3.7) and the plausible sub-groups of the experimental study (0.53/0.34/0.23). Largest factor loadings in the two-class model are $\hat{\lambda}_{1,1} = 1.07$, $\hat{\lambda}_{2,7} = 1.19$, and $\hat{\lambda}_{3,16} = 0.50$, which is mostly similar to the respective latent variable model estimates based on the online questionnaire sample and the plausible sub-groups of the experimental study sample. The factor loadings in the set of indicators for latent variable y_2 (Extraversion) show a slightly but not meaningfully different pattern here.

A similarly reasonable pattern can already be seen with the entire experimental study data as the sample when we do not account for invalid responses in the model (see Figure 3.5). Only factor loadings of the two-class model for the experimental study data related to y_3 do not reveal the same simple pattern, e.g. $\hat{\lambda}_{3,16} = 0.11$ (0.11), which is not significantly different from 0.

Explained Variances

To gain a more clear insight into measurement model accuracy, I would like the draw the readers' attention to differences in explained variance versus error variances of indicator variables. The percentage of explained versus error variance is easily accessible under this analysis design. Each indicator only measures one of the constructs of interests, which in turn have a fixed variance of $\phi_{1,1} = \phi_{2,2} = \phi_{3,3} = 1$. Hence, we can calculate the reliability of indicators as follows: $\lambda^2_{k,j}/(\lambda^2_{k,j} + \psi_j)$. Table 4.4 shows averaged reliabilities of indicators per corresponding latent variable and in total for all indicators. The first two columns provide a very distinct picture about the increase of measurement accuracy for the experimental study sample after having accounted for an invalid responders' class. This is especially true for measurement accuracy of indicators for the first two latent variables y_1 (Emotional Stability) and y_2 (Extraversion). We see little to no change in measurement accuracy for indicators of y_3 (Agreeableness). Nonetheless, on hypothetical higher order factors, measurement accuracy has clearly increased for latent variables corresponding to both the Beta-Factor and the Alpha-Factor. Lastly, the two-latent-class model also seems to be slightly beneficial for measurement accuracy when applied to the online questionnaire study sample.

Table 4.4: Means of explained variances based on observed variables' factor loadings on either of the three factors for the valid response model in the latent variable versus latent class model

Factors	Label	Sample			
		Experimental study		Online questionnaire	
		LVA	LCA	LVA	LCA
y_1	N	.28	.49	.40	.43
y_2	E	.22	.34	.51	.56
y_3	A	.31	.32	.29	.37
		.27	.40	.40	.43

$^{\text{LVA}}$ Latent variable model assuming valid responses only.
$^{\text{LCA}}$ Latent class model accommodating an invalid response strategy.

Factor Covariances

Two of the covariance estimates between the three constructs of interest show no significant correlation and one a significant negative correlation between y_3 and y_1. These results are closer to the theoretical and empirically shown three distinct factors model than the respective results when no invalid response strategy is accounted for. Where we may expect some correlation between y_1 and y_3 because of their mutual affiliation within the Alpha-Factor, we have no significant correlation between them and y_2. y_2 is the only latent variable affiliated with the higher order Beta-Factor. Ignoring group membership or separately estimating the valid measurement model for semi-/implausible conditions was not in line with estimates reported in the literature.

However, we can still observe significant discrepancy between estimates using only plausible responders as the sample and the corresponding parameter estimates in a model that accounts for invalid response strategies.

4.5 Predicted Class Membership

LCA can also be used not only to incorporate invalid responses into the model but also to detect invalid responders. In this section, I will compare the model predicted class memberships with group membership for the experimental study sub-samples. Hence, this section has the purpose of assessing classification performance. In doing so, we can also investigate how well the latent class model represents the experimental study setting.

In this sections reported results are based on predicted class membership using the posterior latent class membership probabilities $\Pr(z = 1|\boldsymbol{x}_i)$ and $1 - \Pr(z = 1|\boldsymbol{x}_i)$ given the observed variables as defined in (4.12). When participants have posterior probabilities greater than .5, they are allocated to the invalid response class and vice versa. Table 4.5 summarises the percentages of individuals in each of the experimental conditions to be allocated to the invalid responders' class. Using the posterior latent class membership probabilities, we would correctly classify about 70% of them as invalid responders. However, we would also incorrectly flag on average about 36% of participants in the plausible response conditions. Hence, we would correctly classify 64% of valid responders and on average 68% of all responders.

In models that incorporate latent classes, it is useful to evaluate how well a measurement model identifies the latent classes. One such measure is the entropy. It is generally used to assess the quality of the measurement instrument as a whole. The entropy of the latent class variable z with probability mass function $\Pr(z)$, is defined by $\text{Ent} = -\sum_{z=0}^{c-1} \Pr(z) \log \Pr(z)$, where c is the number of classes. The entropy is a measure of uncertainty of a random variable and has its maximum when $\text{Ent} = \log c$, which is the case for uniformly distributed variables. It is always non-negative $\text{Ent} \geq 0$ and can be standardised such that $\text{Ent}/\log c \leq 1$. Consequently, the entropy of the latent class variable z given the observed variables can be written as $\text{Ent}^* = -\sum_{z=0}^{c-1} \Pr(z|\boldsymbol{x}) \log \Pr(z|\boldsymbol{x})$. For a particular choice, we would like $\Pr(z|\boldsymbol{x})$ to be 1 or 0. In latent class analysis, the meaning of the standardised entropy is

Table 4.5: Percentage of response patterns identified as members of the invalid class based on their posterior class membership probabilities

Condition	Cell(s)	Proportion allocated to $z = 1$
Plausible	1	.38
	4	.35
Semi-plausible	2	.58
	5	.67
Implausible	3	.82
	6	.74
Plausible	1,4	.36
Semi-plausible	2,5	.63
Implausible	3,6	.77
Semi-/implausible	2,3,5,6	.70

Note Allocated class membership according to largest posterior probability.

traditionally reversed such that latent class entropies approaching 1 indicate a clear delineation of classes (Celeux and Soromenho, 1996). We can calculate the latent class entropy based on a sample of size n using

$$\text{Ent}_{\text{LC}} = 1 + \frac{1}{n \ln c} \sum_{i=1}^{n} \sum_{z=0}^{c-1} \Pr(z|\boldsymbol{x}_i) \ln \Pr(z|\boldsymbol{x}_i). \tag{4.19}$$

Ent_{LC} is a standard output measure in latent class models when results are retrieved via Mplus (Asparouhov and Muthen, 2014). For this particular case of binary class membership, we only need the posterior probability for the invalid class membership as was derived in (4.12) to calculate the latent class entropy, resulting in $\text{Ent}_{\text{LC}} = .879$. Entropy values lower than .8 are often considered problematic if the aim is to predict and use class membership of individuals in further analyses. It seems that the latent class model provides a clear enough separation between classes.

The entropy does not indicate classification performance based on actual group membership. With this latent class analysis, I sought to capture the group membership of valid versus invalid responders with the latent class. In order to evaluate the success of capturing group membership within the latent class variable, we need to compare actual group membership with the associated predicted class membership.

Table 4.6: 2x2 count table of predicted latent class versus experimentally induced/ observed sub-sample membership

	Latent class		
Sub-sample	Valid	Invalid	Sum
Plausible	79	44	123
Semi-/implausible	76	181	257
Sum	155	225	380

We can test the null hypothesis that group and class membership variables are independent. A successful classification should reflect the experimental group membership and, hence, the test for independence should fail. The test is based on a 2×2 count table of class and group membership as shown in Table 4.6. Table 4.6 is a reduced version of Table 4.5. However, Table 4.6 gives the cell counts. By comparing the observed and expected cell frequencies we can perform a standard Pearson χ^2 test of independence. The resulting value $\chi^2 = 39.945$ with degrees of freedom $df = 1$ is significant on a significance level of $p < .01$. Therefore, we reject the null hypothesis of independence between the latent class membership and experimental group membership. However, a test for simple random allocation is not very informative and can be enriched by estimating the effect size $\sqrt{\chi^2/n} = .32$, which is similarly interpretable as a correlation coefficient. These results suggest that the latent class allocation is a mediocre indicator of the experimentally induced plausible versus semi-/implausible conditions.

Ultimately, the latent class analysis provided a clear class separation based on the two different measurement models for valid and invalid responses. Classification performance is significantly better than random allocation, but 36% of incorrectly flagged valid responses clearly exceeds any justifiable tolerance level. It seems reasonable to define a more conservative cut-off criterion for the estimated posterior probabilities of invalid class membership.

4.6 Discussion

A two latent class model, in which we assume a single global (method) factor measurement model for invalid responders, suggests improvement in the accuracy of parameter estimates for the constructs of interest. Furthermore, in our case, we were able to have an accurate estimation of the percentage of invalid responders in the sample. Although the plausible sub-group constitutes $(n_{\text{Cell 1}} + n_{\text{Cell 4}})/n = 32.37\%$ of the total sample and the valid class membership probability was estimated to be $\hat{\eta}_0 = .4$, we can assume that some participants in the experimentally induced semi-plausible response group (Cell 2 and Cell 5) might have had difficulties implementing such a semi-plausible response strategy without partly reverting to the choice of valid responses. Using the posterior class membership probabilities does not seem to provide us with a very precise way to identify invalid responders in the sample. Incorrectly classifying more than 10% valid responders as invalid responders does not represent a sensible level of risk. I will further elaborate on risk levels and discuss what thresholds for incorrectly classifying valid responses is defined as justifiable throughout this thesis in the following chapter when I will shift focus to identification measures. Furthermore, the latent class model was not able to accurately hypothesise the nature of all present invalid response strategies. This might be partly due to the limitations of a latent class approach. It becomes more and more difficult the more invalid responder classes we must account for. At the same time, we need to be able to allow for enough parameters to accommodate less obvious, more complex invalid response strategies. For instance, some response strategies are more complex than merely a long string strategy or an invalid response strategy: that is, an idiosyncratic tendency to favour a specific range of answer options regardless of item content. In the model considered above, I accounted for only one possible invalid response strategy and assumed that there are no other invalid responses present than for those accounted.

For these reasons, our primary goal is the identification of semi-plausible responders and less the investigation of solutions for an appropriate statistical modelling approach. This has been the primary focus of this thesis so far and will continue to be the main research goal. To have a universally applicable approach, the identification procedure should be data-driven but without the necessity of data-customised solutions. This is because users of applied statistics usually do not have sophisticated knowledge about implementing case-customised estimation procedures incorporating complex latent class models. As an example, the latent class model used in this chapter is not directly implementable even in expert statistical analysis software solutions for latent variable models, such as Mplus (for an indirect approach see Section 4.3). Furthermore, latent class analysis can become computationally very expensive. The convergence of algorithms towards local maxima in the likelihood function is another of such complications, and we must exercise caution. For these reasons, most topics in this thesis tend to cover the field of person-fit instruments in more detail (see Section 2.1). As long as semi-plausible response patterns do not exceed a specific impact level on parameter estimation when the statistical model is not accounting for invalid response strategies, the resulting estimates could contain enough information for their identification and subsequent exclusion from the sample.

Chapter 5

A new Measure for Detection

Having set the statistical, theoretical, and empirical framework in a latent variable environment, this chapter will focus on identifying SpRPs. For this purpose, I will integrate knowledge about the statistical properties of SpRPs that I investigated in the former chapter and test and propose a modification of the most promising identification measure presented in Section 2.1.3.

5.1 Modifying the Covariance-based Index

As I will show in the following sections, although the covariance-based person-fit measure $\Upsilon_i(\Sigma, S)$ as defined in (2.31) seems to be the most promising and flexibly applicable index in a variety of research settings, it performs poorly in detecting SpRPs. Therefore, I propose a modification to detect SpRPs which is based on the knowledge gained from investigating the statistical properties of those patterns in the former chapters.

5.1.1 Maximum Penalty Conditions

Υ_i consists of contrasting components, which incorporate in its original form in $\Upsilon_i(\Sigma, S)$ the model-implied covariance and the saturated (unrestricted) covariance matrix. Therefore, it is sensible to investigate what possible other components might be most effective in detecting SpRPs.

In Section 2.1 we saw that there are innumerable ways to detect undesired response patterns, and although they often produce correlated results they reveal and capture different aspects. Where some indices only flag based on prior model specifications and whether or not response patterns fit to subjectively theory driven or empirically tested models, other indices simply allow the data exploratory to define outliers. I have outlined the theoretical framework behind plausibility of response patterns in Section 4.1 and discussed plausibility of response patterns as subject to many causes and their respective statistical manifestations. Therefore, the most sensible approach is not necessarily detecting all the different types of semi-plausible response patterns but rather identifying valid response patterns instead and categorising the remaining as invalid. These two options might seem the same at first glance, but they entail entirely different strategies. This is because the latter approach is much easier, as we usually have an idea, theory, or even empirically researched information about underlying mechanisms of valid responses. Even more so, study instruments (e.g., interview questions, cognitive tests, or biological measurements) are designed to capture information about the underlying model behind valid responses. However, semi-plausible response patterns can be the result of numerous possible mechanisms that usually do not represent the focus of the study objective but undifferentiated noise. Here is why a contrasting approach like Υ_i comes in very handy.

Theoretically, a simple outlier or person-fit measure can only have a limited variability to differentiate between valid and invalid responses. Capturing differences based on characteristics of one type of invalid response strategy often happens at the expense of not measuring properties of other kinds of invalid response strategies. Another limiting factor is the number of possible response combinations or, in other words, possible response patterns. Optimally, we would like to have a variable that assigns different values to each of the possible response patterns to ensure maximum variance/sensitivity. However, this measure would then be a function of many unknown variables, e.g. the actual instrument employed, valid response mechanisms, invalid response strategies, and potentially the latent model underlying the observed data, thus a composite measure that elides complexity within the data. Therefore, a single one-dimensional measure, even if sensitive enough to flag implausible response patterns by assigning them extreme values, is hardly sufficient to identify semi-plausible response patterns. Considering a second measure that is similarly affected by the unknowns but is further a function of one more distinct aspect, say the fit to the theoretical model, is very useful when contrasted to the first. Contrasting could help to statistically partial out all confounding information but the one valid response pattern defining aspect. This way person-fit values are centralised and rendered comparable throughout different studies (or more accurately sets of observed variables). Therefore, to keep the valid response defining aspect as accurate as possible, we would optimally like to find a first contrast component that is minimally affected by this very aspect.

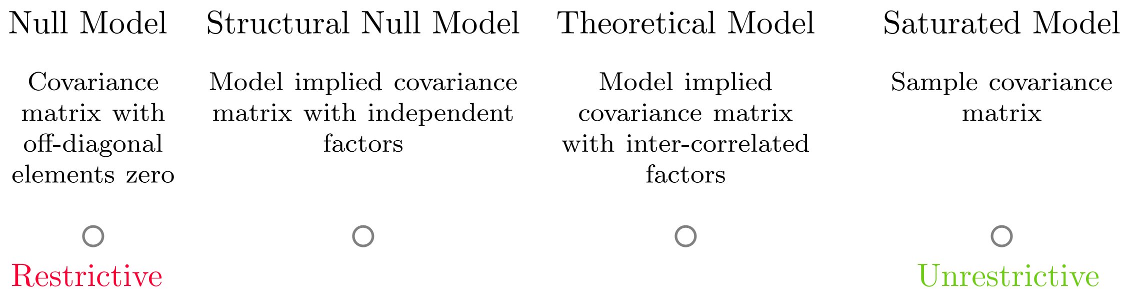

Null Model
Structural Null Model
Theoretical Model
Saturated Model
Covariance matrix with off-diagonal elements zero
Model implied covariance matrix with independent factors
Model implied covariance matrix with inter-correlated factors
Sample covariance matrix
Restrictive
Unrestrictive

Figure 5.1: Possible contrast components ordered with respect to model restrictiveness.

Υ_i provides us exactly with this mechanism. However, the components contrasted are not optimally chosen for our purposes of SpRP detection. The most important quantities apart from the individual response pattern are the provided covariance matrix and mean values for observed variables in each component. In order to discuss other possible contrast components, let us for the sake of the argument span a continuum indicating the degrees of freedom for a certain model. In Figure 5.1 we can see that the most restricted model is represented by the so-called null model. The least restricted model is by nature often referred to as the saturated model, which allows for as many parameters as independent, observed statistics are available. In other words, the saturated model produces a perfect replica of the data covariance matrix. The theoretical model can be allocated in between these two extreme poles. The actual continuum location then depends on how parsimonious we expect and define the valid responses underlying mechanisms to be. Furthermore, Figure 5.1 illustrates another possible scenario that could in our actual case be allocated between the null and theoretical model. This structural null model would be a more restricted theoretical model that assumes no correlation between latent variables. In Chapter 3, we saw that less plausible conditions suggest a single global latent variable model. Covariance estimates between latent variables were unexpectedly large and contradicted the distinct Big Five factor structure in literature. Hence, restricting the model to independent latent variables might help to separate valid responses from invalid response strategies. However, to maximise power and variability in the resulting person-fit measure, we would like to choose the first contrast component to be most distinct from the theoretical model. Hence, I suggest the null model as a contrast to the theoretical model. Ultimately, using the covariance matrix with off-diagonal elements 0 for the first component $\mathcal{N} = \operatorname{diag}(\Sigma)$ of $\Upsilon_i(\mathcal{N}, \Sigma)$ and the model implied covariance matrix as the other contrast component in $\Upsilon_i(\mathcal{N}, \Sigma)$, we should enable differentiation between valid and invalid response patterns. This differentiation is then based on plausibility regarding defining aspects of the theoretical model. However, the reader should keep the discussion in Section 4.1 in mind: The strength of this approach depends upon the assumption that valid responses have, proportionally, the largest impact on model estimation. This means, the parameter estimates are, by tendency, a representation of the valid model and not to be heavily contaminated by SpRPs.

5.1.2 Key Quantities

The modification of the original contrast in (2.31) that I propose

$$\begin{aligned}\Upsilon_i(\mathcal{N}, \Sigma) &= -2[C_i(\mathcal{N}) - C_i(\Sigma)] \\ &= \ln|\mathcal{N}| - \ln|\Sigma| + D_i^2(\mathcal{N}) - D_i^2(\Sigma) \qquad (5.1)\\ &= \text{constant} + D_i^2(\mathcal{N}) - D_i^2(\Sigma)\end{aligned}$$

can be further simplified, by dropping the constant term, into two quantities that depend on i:

$$\mathcal{T}_i = D_i^2(\mathcal{N}) - D_i^2(\Sigma) \qquad (5.2)$$

$D_i^2(\Sigma)$ was defined in (2.29) as

$$D_i^2(\Sigma) = (\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}). \tag{5.3}$$

The contrasted components are variations of the Σ term in the Mahalanobis distance as defined in (2.29), where $\Sigma = \text{diag}(\Sigma) = \mathcal{N}$ in $D_i^2(\mathcal{N})$. Matrix $\mathcal{N}$ represents the covariance matrix under the null model with off-diagonal elements of the theoretical covariance matrix Σ fixed to 0. Σ is the matrix to which estimates of the covariance matrix from the fitted model would converge. It is equal to the true covariance matrix of the variables if the model is correct. In use, the statistic is $D_i^2(\hat{\mathcal{N}}) - D_i^2(\hat{\Sigma})$ using $\hat{\boldsymbol{\mu}}$, which in this case is estimated using the mean vector $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}_i$. The reader should be reminded that we are assuming here that $\Sigma = \Lambda\Phi\Lambda^T + \Psi$ as defined in (2.24). Lastly, the estimated elements of $\mathcal{N}$ are the sample variances of the variables. These are equal to $\text{diag}(\hat{\Sigma})$ if the model is such that it does not impose further constraints on the variances (which is usually the case, and can be assumed here).

Mahalanobis Distance under a Common Factor Analysis Model

Skinner (2014) has shown that we can use properties of Σ^{-1} and $\boldsymbol{y}|\boldsymbol{x}$ under the factor analysis model with $\Phi = \text{I}$ and simple factor loading structure to further decompose the Mahalanobis distances $D_i^2(\mathcal{N})$ and $D_i^2(\Sigma)$ (see 5.15). In the following, I will generalise these results for the unconstrained factor analysis model introduced in Section 2.1.3. This will allow us to derive a detailed interpretation of the components involved in $\mathcal{T}_i$.

First for this purpose, I will use the fact that

$$\Lambda^T(\Lambda\Phi\Lambda^T + \Psi)^{-1} = (\text{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}\Lambda^T\Psi^{-1} \tag{5.4}$$

which can be proven as follows:

$$(\text{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}(\text{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)\Lambda^T(\Lambda\Phi\Lambda^T + \Psi)^{-1} =$$
$$(\text{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}\Lambda^T\Psi^{-1}(\Lambda\Phi\Lambda^T + \Psi)(\Lambda\Phi\Lambda^T + \Psi)^{-1}.$$

Secondly,

$$\Sigma^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda\Phi(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}\Lambda^T\Psi^{-1}, \tag{5.5}$$

as can be seen by post-multiplying both sides of (5.5) by Σ:

$$\begin{aligned}\Sigma^{-1}\Sigma = & \left[\Psi^{-1} - \Psi^{-1}\Lambda\Phi(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}\Lambda^T\Psi^{-1}\right](\Lambda\Phi\Lambda^T + \Psi) = \\ = & \ \Psi^{-1}\Lambda\Phi\Lambda^T + \mathrm{I} - \Psi^{-1}\Lambda\Phi(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)\Lambda^T = \ \mathrm{I}\end{aligned}$$

Using the posterior distribution of $\boldsymbol{y}$ as defined in (2.26) and (5.4) we may write

$$\begin{aligned}\boldsymbol{y}_i^* = E(\boldsymbol{y}_i|\boldsymbol{x}_i) = & \ \Phi\Lambda^T(\Lambda\Phi\Lambda^T + \Psi)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ = & \ \Phi(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}\Lambda^T\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\end{aligned} \tag{5.6}$$

as the $q \times 1$ vector of factor scores. This is interpretable as a transform of the original vector into a vector of expected values of the latent variables defined by the latent variable model.

It follows from (2.29) and (5.5) that

$$D_i^2(\mathcal{N}) = (\boldsymbol{x}_i - \boldsymbol{\mu})^T\mathcal{N}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}). \tag{5.7}$$

whereas the second contrast component

$$\begin{aligned}D_i^2(\Sigma) = & \ (\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ = & \ (\boldsymbol{x}_i - \boldsymbol{\mu})^T[\Psi^{-1} - \Psi^{-1}\Lambda\Phi(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}\Lambda^T\Psi^{-1}](\boldsymbol{x}_i - \boldsymbol{\mu}) \\ = & \ (\boldsymbol{x}_i - \boldsymbol{\mu})^T\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) - (\boldsymbol{x}_i - \boldsymbol{\mu})^T\Psi^{-1}\Lambda\Phi(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda\Phi)^{-1}\Lambda^T\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}).\end{aligned} \tag{5.8}$$

Furthermore using the definition of $\boldsymbol{y}_i^*$ in (5.6)

$$\begin{aligned}\boldsymbol{y}_i^{*T} = & \ (\boldsymbol{x} - \boldsymbol{\mu})^T(\Psi^{-1})^T(\Lambda^T)^T((\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda)^{-1})^T(\Phi)^T \\ = & \ (\boldsymbol{x} - \boldsymbol{\mu})^T\Psi^{-1}\Lambda(\mathrm{I} + \Lambda^T\Psi^{-1}\Lambda)^{-1}\Phi,\end{aligned} \tag{5.9}$$

and

$$\boldsymbol{y}_i^{*T}\Phi^{-1}(\mathrm{I}+\Lambda^T\Psi^{-1}\Lambda) = (\boldsymbol{x}-\boldsymbol{\mu})^T\Psi^{-1}\Lambda, \tag{5.10}$$

it turns out to be convenient to express parts of (5.8) in terms of factor scores as

$$\begin{aligned} D_i^2(\Sigma) = & \quad (\boldsymbol{x}_i-\boldsymbol{\mu})^T\Psi^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})- \qquad\qquad (\boldsymbol{x}_i-\boldsymbol{\mu})^T\Psi^{-1}\Lambda\boldsymbol{y}_i^* \\ = & \quad (\boldsymbol{x}_i-\boldsymbol{\mu})^T\Psi^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})- \quad \boldsymbol{y}_i^{*T}\Phi^{-1}(I+\Phi\Lambda^T\Psi^{-1}\Lambda)\boldsymbol{y}_i^*. \end{aligned} \tag{5.11}$$

Hence, we can write $\mathcal{T}_i$ as a function of observed and latent residuals:

$$\begin{aligned} \mathcal{T}_i = & \quad \boldsymbol{\delta}_i^T\mathcal{N}^{-1}\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^T\Psi^{-1}\boldsymbol{\delta}_i + \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* + \boldsymbol{y}_i^{*T}(\Lambda^T\Psi^{-1}\Lambda)\boldsymbol{y}_i^* \\ = & \qquad \boldsymbol{\delta}_{\boldsymbol{i}}^T(\mathcal{N}^{-1}-\Psi^{-1})\boldsymbol{\delta}_{\boldsymbol{i}} + \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* + \boldsymbol{y}_i^{*T}(\Lambda^T\Psi^{-1}\Lambda)\boldsymbol{y}_i^* \end{aligned} \tag{5.12}$$

where $\boldsymbol{\delta}_i = (\boldsymbol{x}_i-\boldsymbol{\mu})$. Given the fact that $\Psi^{-1} = \mathrm{diag}(\hat{\psi}_j)$ and $\mathcal{N}^{-1} = \mathrm{diag}(\hat{\sigma}_{jj})$, we can express this as

$$\mathcal{T}_i = \sum_{j=1}^{p}(\frac{1}{\sigma_{jj}} - \frac{1}{\psi_j})\delta_{ij}^2 + \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* + \boldsymbol{y}_i^{*T}(\Lambda^T\Psi^{-1}\Lambda)\boldsymbol{y}_i^*. \tag{5.13}$$

Lastly, because $\Lambda^T\Psi^{-1}\Lambda$ is a symmetric matrix, we may write

$$\begin{aligned} \mathcal{T}_i = & \quad \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* + \sum_{k=1}^{q}\sum_{m=1}^{q} y_{ik}^* y_{im}^* \sum_{j=1}^{p}\frac{\lambda_{jk}\lambda_{jm}}{\psi_j} - \sum_{j=1}^{p} z_{ij}^2\frac{\sigma_{jj}}{\psi_j} + \sum_{j=1}^{p} z_{ij}^2 \\ = & \quad \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* + \sum_{k=1}^{q}\sum_{m=1}^{q} y_{ik}^* y_{im}^* \sum_{j=1}^{p}\frac{\lambda_{jk}\lambda_{jm}}{\psi_j} - \sum_{j=1}^{p} z_{ij}^2(\frac{\sigma_{jj}}{\psi_j} - 1) \qquad , \end{aligned} \tag{5.14}$$

where $z_{ij} = \delta_{ij}/\sqrt{\sigma_{jj}} \sim N(0,1)$, e.g. standard normalised scores of x_{ij}.

In (5.14) I wrote $\mathcal{T}_i$ such that we have two terms involving latent variable scores, y_{ik}^*, and one term involving observed-variable scores, x_{ij}, for a given i.

All of the terms except for the first term, which only involves a squared distance of $\boldsymbol{y}_i^*$ from their multivariate distribution, are subject to weighting terms. As a side note, $\boldsymbol{y}_i^*$ can be interpreted as latent variable scores because approximately these can be seen as an estimate of the extremeness of the observation's values of $\boldsymbol{y}_i$. We can see that the ratios σ_{jj}/ψ_j and $\lambda_{jk}\lambda_{jm}/\psi_j$ can quickly becoming the defining elements of $\mathcal{T}_i$ if $\psi_j \rightarrow 0$. Hence, a latent variable model with little noise will affect the weighted terms such that unweighted squared latent variable scores become negligible.

The most interesting feature is that the penalty for being a univariate outlier in $\boldsymbol{x}$ and the penalty for being a multivariate outlier in $\boldsymbol{y}^*$ are of opposite signs. The observed variable term is always non-positive because $\sigma_{jj} \geq \psi_j$ whereas the first latent variable term will always be non-negative because Φ^{-1} is positive definite. The second latent variable term can take on negative values in certain situations which I will elaborate on further below.

A parsimonious interpretation of this composition is that when i has a highly aberrant response pattern with regards to observed-variable scores, then i has larger likelihood of being member of the semi-plausibly responding group, e.g. extreme values in $\mathcal{T}_i$. This is the case unless the aberrance captured in $\boldsymbol{x_i}$ is due to extreme (by the model correctly estimated) latent variable levels captured in $\boldsymbol{y}_i^*$. Therefore, we would expect values near 0 for valid response patterns.

On the other hand, if we have a bad measurement model with large elements of noise we would even for invalid response patterns expect values near 0. For example, if all elements of Λ are close to 0, so are also $\boldsymbol{y}^*$ from (5.6) and $\frac{\sigma_{jj}}{\psi_j} - 1$, so all terms of (5.12) are close to 0.

Therefore, any inconsistency of either extreme latent variable scores or extreme observed-variable scores will only lead to extreme scores in $\mathcal{T}_i$ for model aberrant response patterns but only given that the model for valid responses reflects a good measurement in the first place.

Mahalanobis Distance under a Factor Analysis Model with independent Factors

In a factor analysis model with independent factors we have each observed variable only serving as an indicator for one of the defined latent variables (simple factor structure) and uncorrelated factors (e.g., as suggested in 2.27). Consequently, we have $\Lambda^T \Psi^{-1} \Lambda$ to be a diagonal matrix such that the latent variables are also independent a posteriori given the observed variables.

Given a simple factor structure, we can write (5.14) as

$$\begin{aligned}\mathcal{T}_i = \quad & \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* + \sum_{k=1}^{q} y_{ik}^{*2}\sum_{j=1}^{p}\frac{\lambda_{jk}^2}{\psi_j} + 2\sum_{k\neq m}^{q} y_{ik}^{*}y_{im}^{*}\sum_{j=1}^{p}\frac{\lambda_{jk}\lambda_{jm}}{\psi_j} & -\sum_{j=1}^{p} z_{ij}^2(\frac{\sigma_{jj}}{\psi_j}-1) \\ = \quad & \sum_{k=1}^{q} y_{ik}^{*2} + \sum_{k=1}^{q} y_{ik}^{*2}\sum_{j=1}^{p}\frac{\lambda_{jk}^2}{\psi_j} & -\sum_{j=1}^{p} z_{ij}^2(\frac{\sigma_{jj}}{\psi_j}-1)\end{aligned} \tag{5.15}$$

because $\lambda_{jk}\lambda_{jm} = 0$ given $k \neq m$ and $\Phi = \mathrm{I}$. The eliminated third term can be negative. However given we can drop that therm, we have a clear contrast between positive penalty with regards to the latent variables and negative penalty with regards to the observed variables. Furthermore, it follows from the simple factor structure and $\Phi = \mathrm{I}$ that

$$\sigma_{jj} = \mathrm{Var}(x_j) = \sum_{k=1}^{q}\lambda_{jk}^2 + \psi_j = \lambda_j^2 + \psi_j, \tag{5.16}$$

because, in a simple factor structure model, $\lambda_{jk}^2 \neq 0$ only for one k.

For illustrational purposes, given a very homogeneous measurement model with regards to standard normal observed variables' ($x_j \sim N(0,1)$), factor loading ($\sum_{k=1}^{q}\lambda_{jk}^2 = r + (q-1)*0, \forall j$), error variance ($\psi_j = 1 - r, \forall j$), and p/q number of indicators per latent variable constant, we can write

$$\begin{aligned}\mathcal{T}_i = \quad & \sum_{k=1}^{q} y_{ik}^{*2} + \sum_{k=1}^{q} y_{ik}^{*2}\frac{p}{q}\frac{r}{1-r} - \sum_{j=1}^{p} z_{ij}^2\frac{r}{1-r} \\ = \quad & \sum_{k=1}^{q} y_{ik}^{*2} + \frac{r}{1-r}\left[p\sum_{k=1}^{q}\frac{y_{ik}^{*2}}{q} - \sum_{j=1}^{p} z_{ij}^2\right],\end{aligned} \tag{5.17}$$

where r can be interpreted as a measure of reliability. Once again, we can see that the ratio $r/(1-r)$ dominates as r increases. Hence, the equal contrast of deviation from $\boldsymbol{x}$ and deviation in $\boldsymbol{y}$ (averaged squared factor scores multiplied by p) plays the most important role in $\mathcal{T}_i$ and raw squared latent variable scores $\sum_{k=1}^{q} y_{ik}^{*2}$ become increasingly negligible with respect to i.

Ultimately, T_i seems to be a parsimonious index that accounts for easily accessible quantities like the variance, error variance and individual levels of i in $\boldsymbol{x}$ and $\boldsymbol{y}$. However, to the same extent, T_i is sensitive towards reliability of manifest variables and complexity (averaged over q) of the measurement model.

Alternative Interpretation

In order to gain a clearer understanding of which components of $\mathcal{T}_i$ are affected by the extent of error variance in the valid response model, I will give an alternative interpretation of the identification measure. We may define

$$\begin{aligned} \boldsymbol{\delta}_i^* &= (\Lambda\Phi\Lambda^T)\Sigma^{-1}\boldsymbol{\delta}_i \\ &= \Lambda\boldsymbol{y}_i^* \qquad\qquad , \end{aligned} \tag{5.18}$$

i.e. as fitted values for the observed variables $\boldsymbol{\delta}_i$ from the factor model if $\boldsymbol{y}_i^*$ was the value of the factor. Substituting $\Lambda\boldsymbol{y}_i^*$ in (5.12) yields

$$\begin{aligned} \mathcal{T}_i = &\ \boldsymbol{\delta}_i^T\mathcal{N}^{-1}\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^T\Psi^{-1}\boldsymbol{\delta}_i + \boldsymbol{y}_i^{*T}\Lambda^T\Psi^{-1}\Lambda\boldsymbol{y}_i^* &&+\boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* \\ = &\ \boldsymbol{\delta}_i^T\mathcal{N}^{-1}\boldsymbol{\delta}_i + \boldsymbol{\delta}_i^{*T}\Psi^{-1}\boldsymbol{\delta}_i^* - \boldsymbol{\delta}_i^T\Psi^{-1}\boldsymbol{\delta}_i &&+\boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* \\ = &\ \boldsymbol{\delta}_i^T\mathcal{N}^{-1}\boldsymbol{\delta}_i + (\boldsymbol{\delta}_i^* - \boldsymbol{\delta}_i)^T\Psi^{-1}(\boldsymbol{\delta}_i^* + \boldsymbol{\delta}_i) &&+\boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^*. \end{aligned} \tag{5.19}$$

Using the symmetry of matrices N^{-1} and Ψ^{-1}, we can further simplify such that

$$\begin{aligned} \mathcal{T}_i = &\ \boldsymbol{\delta}_i^T\mathcal{N}^{-1}\boldsymbol{\delta}_i + (\boldsymbol{\delta}_i^* - \boldsymbol{\delta}_i)^T\Psi^{-1}(\boldsymbol{\delta}_i^* + \boldsymbol{\delta}_i) &&+\boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* \\ = &\ \sum_{j=1}^{p} z_{ij}^2 - \sum_{j=1}^{p}\frac{\delta_{ij}^2 - \delta_{ij}^{*2}}{\psi_j} &&+\boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^*, \end{aligned} \tag{5.20}$$

and interpret the term in middle as weighted residual term that becomes increasingly dominant with $\psi_j \to 0$. Furthermore, the middle term is either negative or 0. In order to prove this, it is required to show that

$$\begin{aligned} &(\boldsymbol{\delta}_i^* - \boldsymbol{\delta}_i)^T\Psi^{-1}(\boldsymbol{\delta}_i^* + \boldsymbol{\delta}_i) \\ = &(\boldsymbol{\delta}_i^* - \boldsymbol{\delta}_i)^T\Psi^{-\frac{1}{2}}\Psi^{-\frac{1}{2}}(\boldsymbol{\delta}_i^* + \boldsymbol{\delta}_i) \\ = &(\Psi^{-\frac{1}{2}}\boldsymbol{\delta}_i^* - \Psi^{-\frac{1}{2}}\boldsymbol{\delta}_i)^T(\Psi^{-\frac{1}{2}}\boldsymbol{\delta}_i^* + \Psi^{-\frac{1}{2}}\boldsymbol{\delta}_i) \ \le 0. \end{aligned} \tag{5.21}$$

Before we further proceed I temporarily define

$$\Psi^{-\frac{1}{2}}\boldsymbol{\delta}_i^* = \boldsymbol{\alpha}_i^* \qquad \text{and} \qquad \Psi^{-\frac{1}{2}}\boldsymbol{\delta}_i = \boldsymbol{\alpha}_i \tag{5.22}$$

for ease of presentation. First, we multiply out the left hand side and write

$$(\boldsymbol{\alpha}_i^* - \boldsymbol{\alpha}_i)^T(\boldsymbol{\alpha}_i^* + \boldsymbol{\alpha}_i) = \boldsymbol{\alpha}_i^{*T}\boldsymbol{\alpha}_i^* - \boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_i^* + \boldsymbol{\alpha}_i^{*T}\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i^{*T}\boldsymbol{\alpha}_i^* - \boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_i. \tag{5.23}$$

Secondly, we reiterate the definition of $\boldsymbol{\delta}_i^*$,

$$\boldsymbol{\delta}_i^* = (\Lambda\Phi\Lambda^T)\Sigma^{-1}\boldsymbol{\delta}_i, \tag{5.24}$$

and rewrite $\boldsymbol{\delta}_i$ such that

$$\boldsymbol{\delta}_i = (\Lambda\Phi\Lambda^T + \Psi)\Sigma^{-1}\boldsymbol{\delta}_i. \tag{5.25}$$

Since $\Sigma = \Lambda\Phi\Lambda^T + \Psi$ we can then write (5.23) as

$$\begin{aligned}
(\Psi^{-\frac{1}{2}}\Lambda\Phi\Lambda^T\Sigma^{-1}\boldsymbol{\delta}_i)^T(\Psi^{-\frac{1}{2}}\Lambda\Phi\Lambda^T\Sigma^{-1}\boldsymbol{\delta}_i) - (\Psi^{-\frac{1}{2}}\Sigma\Sigma^{-1}\boldsymbol{\delta}_i)^T(\Psi^{-\frac{1}{2}}\Sigma\Sigma^{-1}\boldsymbol{\delta}_i) \\
= \boldsymbol{\delta}_i^T\Sigma^{-1}\Lambda\Phi\Lambda^T\Psi^{-1}\Lambda\Phi\Lambda^T\Sigma^{-1}\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^T\Sigma^{-1}\Sigma\Psi^{-1}\Sigma\Sigma^{-1}\boldsymbol{\delta}_i \\
= \boldsymbol{\delta}_i^T\Sigma^{-1}(\Lambda\Phi\Lambda^T\Psi^{-1}\Lambda\Phi\Lambda^T - \Sigma\Psi^{-1}\Sigma)\Sigma^{-1}\boldsymbol{\delta}_i \\
= \boldsymbol{\delta}_i^T\Sigma^{-1}\left[\Lambda\Phi\Lambda^T\Psi^{-1}\Lambda\Phi\Lambda^T - (\Lambda\Phi\Lambda^T + \Psi)(\Psi^{-1}\Lambda\Phi\Lambda^T + \Psi^{-1}\Psi)\right]\Sigma^{-1}\boldsymbol{\delta}_i \\
= (-1)\cdot\boldsymbol{\delta}_i^T\Sigma^{-1}\left[\Lambda\Phi\Lambda^T + \Lambda\Phi\Lambda^T + \Psi\right]\Sigma^{-1}\boldsymbol{\delta}_i \\
= (-1)\cdot\boldsymbol{\delta}_i^T\Sigma^{-1}\left[\Lambda\Phi\Lambda^T + \Sigma\right]\Sigma^{-1}\boldsymbol{\delta}_i
\end{aligned} \tag{5.26}$$

The next step is to demonstrate that the following inequality holds:

$$\boldsymbol{\delta}_i^T\{\Sigma^{-1}\Lambda\Phi\Lambda^T\Sigma^{-1} + \Sigma^{-1}\}\boldsymbol{\delta}_i \geq 0. \tag{5.27}$$

For this inequality to hold, the $p \times p$ matrix term within the curly brackets needs to be positive semi-definite. By definition, Σ is positive definite and, consequently, its inverse, too. Furthermore, Φ is positive definite. It then follows that

$$\begin{aligned} &\boldsymbol{\delta}_i^T(\Sigma^{-1}\Lambda\Phi\Lambda^T\Sigma^{-1})\boldsymbol{\delta}_i \\ &= (\Lambda^T\Sigma^{-1}\boldsymbol{\delta}_i)^T\Phi(\Lambda^T\Sigma^{-1}\boldsymbol{\delta}_i) \\ &= a^T\Phi a \qquad\qquad \geq 0 \end{aligned} \tag{5.28}$$

and is only 0 if $\boldsymbol{\delta}_i = 0$ or $\Lambda = 0$. Hence, $(\Sigma^{-1}\Lambda\Phi\Lambda^T\Sigma^{-1})$ is also positive semi-definite. Lastly, the sum of two positive definite matrices results in a positive semi-definite matrix. Hence, the resulting matrix within the curly brackets in a quadratic form with a non-zero column vector yield only positive values.

The key result here is the middle term in (5.20). We can conclude that it is the only negative term in this alternative representation of $\mathcal{T}_i$ as a sum of three components. For applications where we would like to pursue a conservative testing procedure, we should limit a cut-off for negative values only. The penalties for outlier responses in latent scores and observed values are of positive sign. Hence, if a valid response pattern is highly aberrant based on an extreme response pattern, this would only be caught by positive values. However, if there is an additional discrepancy of fitted versus actual observed values, this might indicate that there is a semi-plausible response strategy behind a response pattern. This could result in negative values in $\mathcal{T}_i$, however, only given that the valid response model is reliable in the first place (captured in Ψ). If we have a measurement model for valid responses that has large amounts of noise, we have little information and, consequently, even invalid response patterns cannot be assigned large negative penalties.

Lastly, I would like to further elaborate on the middle term of the alternative interpretation of $\mathcal{T}_i$ in (5.20). Where the interpretation of the first term in (5.20) (sum of standardised squared observed $\boldsymbol{x}$) and the last term (multivariate extremeness of predicted values of $\boldsymbol{y}^*$) are fairly straightforward, the interpretation of the difference $\delta_{ij}^2 - \delta_{ij}^{*2}$ is of a more complex nature. It is helpful to be reminded that $E(\boldsymbol{y}) = 0$, so we can write $\boldsymbol{\delta}_i^* = \Lambda\boldsymbol{y}_i^* - \Lambda 0$, where $\Lambda 0 = 0$. Hence, $\boldsymbol{\delta}_i^*$ is also a difference similar to $\boldsymbol{\delta}_i$. However, $\boldsymbol{\delta}_i = (\boldsymbol{x}_i - \boldsymbol{\mu})$ is a measure of (univariate) extremeness of observed $\boldsymbol{x}$ where $\boldsymbol{\delta}_i^* = \Lambda\boldsymbol{y}_i^*$ represents a value on the $\boldsymbol{x}$-scale implied by predicted values of $\boldsymbol{y}$. The result of $\delta_{ij}^2 - \delta_{ij}^{*2}$ can then be interpreted as a residual of the observed scores for participant i that remains even after we have accounted for the participant's individual (model implied) extremeness in the latent variables. This residual is then weighted by ψ_j, which is a measure of the measurement accuracy or a model's capacity to capture the differences between participants' observed values with latent variables.

Throughout the thesis, I will focus on $\mathcal{T}_i$'s interpretation as discussed in the previous sections because it offers a more intuitive understanding of negative and positive values when used as a test measure with two-sided cut-offs. Where it helps comprehension of results, we will draw links to the interpretation used in this sections, nonetheless.

5.2 Methods for Detection using the new Measure

In the previous section, I defined the identification measure $\mathcal{T}_i$ as a new detection instrument for SpRPs. This test measure is a function of $\mathcal{N}$, Σ, and $\boldsymbol{\mu}$. $\mathcal{N}$ is defined as the covariance matrix of observed variables under the independence model. Therefore, it equals the model implied covariance Σ, where off-diagonal elements are set to 0 and we can simply define $\mathcal{N} = \text{diag}(\Sigma)$.

In this section, I will outline methods of using $\mathcal{T}_i$ to detect SpRPs. The simplest approach to resolving problems arising from SpRPs is to exclude them from further analyses once they have been identified. Therefore, further analyses will be based on a sub-group of smaller size consisting only of valid responses that follow the population model defined through Σ and $\boldsymbol{\mu}$.

5.2.1 Information Sources for $\mathcal{T}_i$

The first step towards identifying SpRPs is to derive $\mathcal{T}_i$ for each i'th response pattern in the sample of size n. Here, we are confronted with two possible scenarios: First, the valid response population model is known, or more precisely, Σ and $\boldsymbol{\mu}$ are known. In this case, we can simply calculate values of $\mathcal{T}_i$ under the valid response model. Second, we do not know parameters defining the valid response model, and it is required to derive estimates of parameters that define Σ and $\boldsymbol{\mu}$.

A known valid response model represents the optimal but a rather less likely scenario. This is specifically true for social statistics. Nonetheless, we might be able to acquire information of the valid response model from diverse information sources. For instance, coming back to the example study setting of personality assessment, we can draw information from results of norming procedures that large-scale implemented test instruments often are required to undergo. The IPIP and the Big Five personality framework have a long history of quantitative analyses in numerous cultures and specific population groups. Even more so, personality assessment is still consistently used by many professional sectors as well as often even implemented as a by-product for loosely related study questions.

Outcomes of other Studies In light of this, latent variable models provide the instrument of choice: In Chapter 3, I used information from previous literature about the latent variable structure and set numerous constraints to the analysis model, such as loading patterns of observed variables. This way, I defined an analysis model that is much closer to the response model that valid responses originate from than the diverse invalid response strategies. Admittedly, we saw that resulting estimates seem to be strongly biased when based on a sample that consists not only of valid responses. However surprisingly, the results did not produce entirely unreasonable results in spite of a majority of invalid responses in the analysis sample. In the following sections, I am going to use the estimated parameters $\hat{\Sigma}$ and $\hat{\boldsymbol{\mu}}$ as information source for $\mathcal{T}_i$. I will demonstrate that even biased estimates can provide useful information for the discrimination of invalid responses with $\mathcal{T}_i$.

Measurement equivalent Samples Benefits of previous studies based on the same measurement instrument are not only limited to extraction of latent structural knowledge. In fact given that certain requirements are met, results from other samples that are free of SpRPs or are known only to have a small fraction of the sample not following the valid response model can be fully adapted. $\hat{\Sigma}$ and $\hat{\boldsymbol{\mu}}$ derived from valid responders only samples, can be used as information source for $\mathcal{T}_i$ to derive values of response patterns in other samples of more problematic valid versus invalid response ratios. This approach is reasonable if study samples are comparable, i.e. samples from the same population, hence, measurement equivalence is given.

The combined Approach In some cases, the sample might be heavily contaminated with invalid responses, as is the case for our experimental study sample. Although even in this case $\mathcal{T}_i$ proves to be very useful in identifying SpRPs (see Section 5.3), it seems sensible to implement other methods to further reduce estimation bias prior to detection. I expect that more accurate estimates as information to $\mathcal{T}_i$ will improve its discrimination power even further. In Chapter 6, I will evaluate a combined approach towards detection. We have seen in the previous Chapter 4, that accommodating an invalid response strategy into the latent class model severely reduced measurement error in the valid response model. Furthermore, valid response model parameter estimates were more sensible and closer to the corresponding estimates of the online questionnaire study. Invalid response strategies are uncountable and have idiosyncratic components based on individual invalid response strategies. However, because of limitations with regards to model identifiability, we would not possibly be able to account for all different types of invalid response patterns. Nonetheless, the invalid response strategy item wording seems to reflect a broad range of invalid response patterns just well enough to derive more accurate estimates of the valid response model parameters. In the combined approach, the estimated valid model implied covariance matrix $\hat{\Sigma}^{(0)}$ and mean values $\hat{\boldsymbol{\mu}}^{(0)}$ serve as information source to $\mathcal{T}_i$.

5.2.2 Deriving Cut-off Values for $\mathcal{T}_i$

After we have derived $\mathcal{T}_i$ either with known parameters Σ and $\boldsymbol{\mu}$ or estimates of them, a cut-off criterion or cut-off criteria need to be defined. Based on the interpretations of $\mathcal{T}_i$, discussed in Section 5.1, we either flag extreme negative values below a certain threshold or extreme positive and negative values outside a certain range as extreme response patterns in $\mathcal{T}_i$.

Percentiles of $\mathcal{T}_i$ In an optimal scenario, we would like to have knowledge about two things: the theoretical distribution of $\mathcal{T}_i$ under the theoretical valid response model with known Σ and $\boldsymbol{\mu}$ and the percentage of SpRPs in the sample. In Section 5.4, I will derive said distribution. Having the theoretical distribution enables us to choose cut-off values based on percentiles of its distribution. As a conservative example with one cut-off criterion, we can decide to set the risk of excluding valid responses to a certain level, such as 1%. Based on the theoretical distribution, we can derive the first percentile of $\mathcal{T}_i$ under the theoretical valid response model for valid responses. This would serve as the cut-off value, such that any response pattern i with a $\mathcal{T}_i$ value below that cut-off criterion can be classified as invalid. A less conservative approach would be to derive the .5th and 99.5th percentiles as two-sided cut-off criteria.

Ratio of valid versus invalid Responses The most important aspect in defining the risk level of flagging valid responses is the percentage of SpRPs in the sample. The risk of incorrectly identified valid and correctly detected invalid responses need be kept in a sensible proportion. However, a scenario in which the percentage of SpRPs is known is unlikely. In Chapter 1, we extracted estimates varying between 5 and 15 percent of invalid responses in online studies. In Chapter 4, we have seen that latent class analysis can be used to derive a satisfactory estimate of the percentage of valid responses, η_0, in the sample. For the online questionnaire study sample, the estimated percentage of invalid responses was $1 - \eta_0 = .10$. In these scenarios, a risk level of 10% flagged valid responses can represent a maximally tolerable threshold, if we would be willing to sacrifice less than one valid response for detecting one potentially very influential invalid response. Consequently, throughout this thesis, I will use this reasoning to interpret outcomes based on a maximally tolerable risk level of 10% incorrectly classified valid responses. It is apparent that a risk level of 10% changes meaning with the ratio of valid versus invalid responses in the sample as well as the total sample size. For instance, the experimental study sample consists of nearly 68% semi-/implausible responders and excluding 10% of valid responses may seem a drastic approach. However, the severity of contamination in this sample may justify a risk level of 10% incorrectly flagged valid responses because otherwise, we may not expect to gain any reasonable information from valid response model estimates (cf. Chapter 3). It is my intention to elaborate more on these risk levels throughout different evaluation scenarios and have a discussion about this matter in the last chapter of this thesis, taking into account previous and following findings. Nonetheless, I believe it is of value to define a consistent level of tolerable risk throughout the thesis to compare outcomes of different approaches with regards to their potential for detecting SpRPs.

Empirically enriched Decisions The previous paragraph referred to scenarios where we only have estimates $\hat{\Sigma}$ and $\hat{\boldsymbol{\mu}}$ as an information source for the calculation of $\mathcal{T}_i$. We do not know the theoretical distribution of $\mathcal{T}_i$ under the valid response model, either, and, hence, can only derive cut-off values of $\mathcal{T}_i$ for valid responses based on the estimated valid response model. Using theoretical percentiles based on the estimated valid response model as cut-off criteria becomes a less tangible approach. Nonetheless, I will still use percentiles as cut-criteria throughout the thesis to evaluate the discriminatory power of $\mathcal{T}_i$ based on this approach. In the last chapter of this thesis, I will discuss the appropriateness of percentiles as cut-off value based on results of a large-scale simulation study. When in doubt, I strongly suggest investigating the empirical cumulative distribution of $\mathcal{T}_i$ in detail. However, percentiles derived on the basis of the biased theoretical distribution of $\mathcal{T}_i$ can provide anchors for an individual choice of cut-off values. The empirical distribution is a good instrument to identify areas of large extreme values that do not follow a typical distribution shape of $\mathcal{T}_i$ under the theoretical valid response model.

A step-wise Approach Lastly, a computationally intensive but conservative approach can be followed by step-wise estimating the valid response model parameters and excluding only the most extreme values of $\mathcal{T}_i$. This way each new estimation of the valid response model will be less and less affected by SpRPs. Exclusion criteria will then be based on more accurate information of the valid response model. Step-wise estimation and exclusions can be followed until a satisfactory congruence of empirical and theoretical distribution of $\mathcal{T}_i$ is reached.

5.2.3 Summary

In most cases, we will have some information about the valid response model, but it will be required to derive estimates for parameters as information source for the new measure. In samples also consisting of SpRPs, we will need to use analysis results as estimates in spite of potential measurement error. In samples with a large proportion of invalid response patterns, we will need to set model restrictions such that the estimation process will produce results that represent the valid response model more accurately than invalid response strategies. Therefore, it is important to have diverse observed variables as indicators for a complex latent variable structure in the valid response model. Or in other words, the more distinct the valid response model is from invalid response strategies, the more accurate information we can provide to derive minimally biased estimates $\mathcal{T}_i$ under the valid response model. Latent class analysis is a limited but potentially very useful instrument in order to separate some measurement error associated with a small number of invalid response strategies from the valid response model.

Furthermore, probabilities of group membership derived via latent class analysis can indicate the percentage of SpRPs in the sample. Using this information, we need to set a case-specific risk level of incorrectly identified valid responses. Estimated or theoretical percentiles of $\mathcal{T}_i$ give information for a sensible choice of cut-off values. Further information can be collected by investigating the empirical distribution of $\mathcal{T}_i$. A computationally intensive but conservative approach can be implemented by step-wise estimation of the valid response model and exclusion of most extreme cases of SpRPs.

5.3 Motivation for the Modification with empirical Data

The previous section depicted the theoretical derivation of the new measure for the identification of SpRPs and defined the key quantities as well as major impact factors for $\mathcal{T}_i$. I further discussed methods and different approaches to using the new measure to detect SpRPs. The main purpose of this section is to empirically motivate the proposed modification of the original $\Upsilon_i(\Sigma, S)$ and further evaluate the new measure $\mathcal{T}_i$. However, I will focus on $\Upsilon_i(\mathcal{N}, \Sigma) = \text{constant} + \mathcal{T}_i$ as defined in (5.1), in order to allow for a direct comparison between $\Upsilon_i(\Sigma, S)$ and its modified version.

5.3.1 Distributional Properties

Prior to further evaluational analyses to $\Upsilon_i(\mathcal{N}, \Sigma)$, this section is dedicated to briefly capturing descriptive plots and measures with regards to the new measure.

The two relevant quantities in $\Upsilon_i(\mathcal{N}, \Sigma)$ are $D_i^2(\Sigma)$ and $D_i^2(\mathcal{N})$. Therefore, it is sensible to investigate their distributions in the different evaluation scenarios.

For the following histograms I differentiate between sample points of different groups. Histograms for the experimental study sample depict plausible responders (cells 1 and 4) in black colour and semi/-implausible responders (cells 2, 5, 3, and 6) in grey. The online questionnaire sample does not have experimentally induced different groups of responders. Nonetheless as an additional information source, I used the latent class analysis estimates for this sample to classify responders. The classification is based on the modal probability approach. For every individual I estimate the posterior probability of belonging to the latent valid responders' and invalid responders' class. Response patterns are assigned to the class with largest posterior probability. The invalid response group is depicted in grey colour. Furthermore, vertical lines indicate the mean values of either classes or groups in the histograms, respectively. The colour coding is consistent throughout the section.

D_i^2 is statistically $D_i^2 \sim \chi^2(p)$ where p is the number of observed variables, when the true population parameters (e.g., Σ) are used and also asymptotic with estimated parameters drawing on a large sample. Figure 5.2 shows histograms for $D_i^2(\Sigma)$ for the

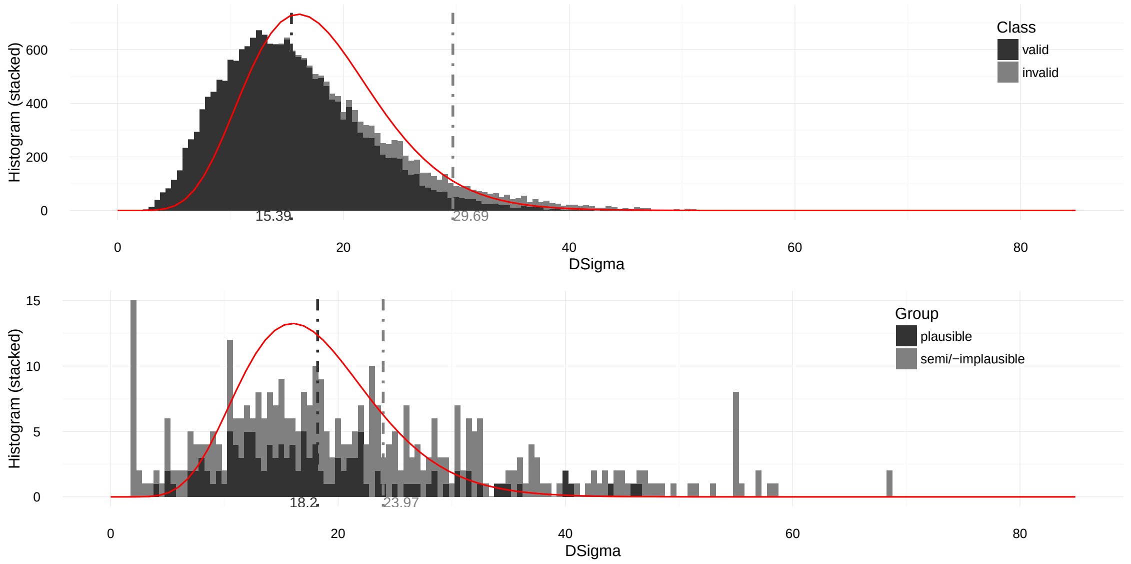


Figure 5.2: Stacked histogram based on class/group membership for $D_i^2(\Sigma)$ for online questionnaire sample (top) and experimental study sample (bottom) with the corresponding χ^2 distribution curve.

experimental and online questionnaire study whereby the model implied covariance matrix was estimated including all sub-groups and assuming no invalid response patterns. For the online questionnaire sample, we can see that the distribution has a slight shift to smaller numbers compared to the theoretical distribution curve. Furthermore, we see that the invalid class members tend to have, in general, larger $D_i^2(\Sigma)$ values with mean at 29.69, where valid class members have a mean of 15.39. The expected mean for a $\chi^2(18)$ distribution is 18. Hence, the response patterns with the largest deviation from the mean in $D_i^2(\Sigma)$ are assigned to the invalid class membership based on the latent class analysis model. With regards to the experimental study sample, we can see that although the shape of the histogram does not necessarily comply with the theoretical distributions of $\chi^2(18)$, there are predominantly members of the experimentally induced semi-/implausible conditions who fall into extreme positions where the theoretical distribution tails approximate to 0.

There are several potential reasons that can help to explain why the empirical data does not follow the theoretical distribution. Observed variables are assumed to be multivariate normally distributed. Univariate distribution plots in Section 3.3.2 revealed that many variables have a left-skewed distribution and data appears right-censored. Non-normality of observed variables would explain that the theoretical distribution is not matched well by the empirical distribution. Furthermore, we expect the parameter estimates for the latent variable models to be biased given the fact that in our traditional analyses setting SpRPs are not accounted for by the model. This is especially a problem for the experimental study sample which consists of predominantly invalid responders. The small sample size of plausible responders in the experimental study setting is in itself a potential cause that the empirical data cannot reproduce the expected theoretical distribution.

The second important component of the new test measure $\Upsilon_i(\mathcal{N}, \Sigma)$ is $D_i^2(\mathcal{N})$. Similar to previously discussed Figure 5.2 the histogram for $D_i^2(\mathcal{N})$ in Figure 5.3 for the experimental study sample reveals predominantly members of the experimentally induced semi-/implausible conditions to be assigned extreme values. However, the separation between members of the plausible and semi/-implausible groups are not as distinct as it is the case in $D_i^2(\Sigma)$. The difference between mean values 17.82 (plausible) and 19.16 (semi/-implausible) is smaller than the corresponding mean

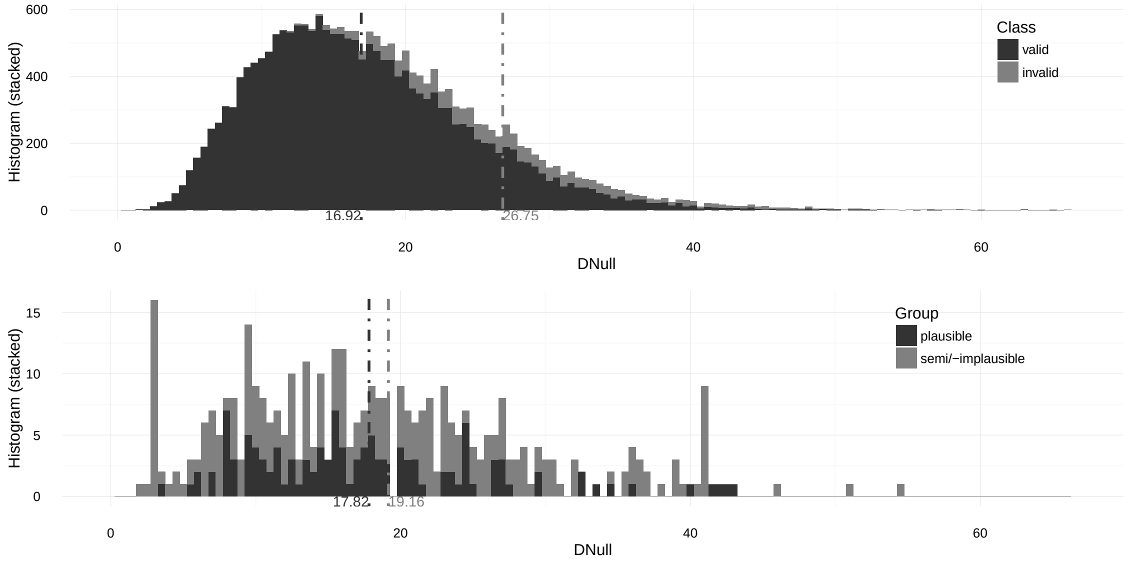


Figure 5.3: Stacked histogram for $D_i^2(\mathcal{N})$ for online questionnaire sample (top) and experimental study sample (bottom).

values in $D_i^2(\Sigma)$. A similar contrast between histograms of $D_i^2(\mathcal{N})$ and $D_i^2(\Sigma)$ is observable for the online questionnaire sample and the two classes valid versus invalid responses. Mean values 16.92 (plausible) and 26.75 (semi/-implausible) still indicate a very good separation between members of the two classes, but we can observe a larger variance for invalid responders.

The clear separation between online questionnaire study sample class members is not as easily obtained for the experimental study sample groups. The online questionnaire sample classification is based on the latent class analysis model. $D_i^2(\Sigma)$ is the essential information source when we minimise the log-likelihood function to obtain the latent class model estimates. Therefore, it is unsurprising that we see a clear separation of class members in the online questionnaire study sample when we investigate the $D_i^2(\Sigma)$ histogram. The achieved separation of experimental study groups is more important. Here, we can see poor discriminatory power of $D_i^2(\Sigma)$ and $D_i^2(\mathcal{N})$ for each of them on their own. This is especially observable when the means between groups are compared.

In Figure 5.4 we can investigate whether a combination of $D_i^2(\Sigma)$ and $D_i^2(\mathcal{N})$, or more accurately their difference, as is apparent in the new test measure $\Upsilon_i(\mathcal{N}, \Sigma)$ helps to discriminate between valid and semi-plausible response patterns. Once

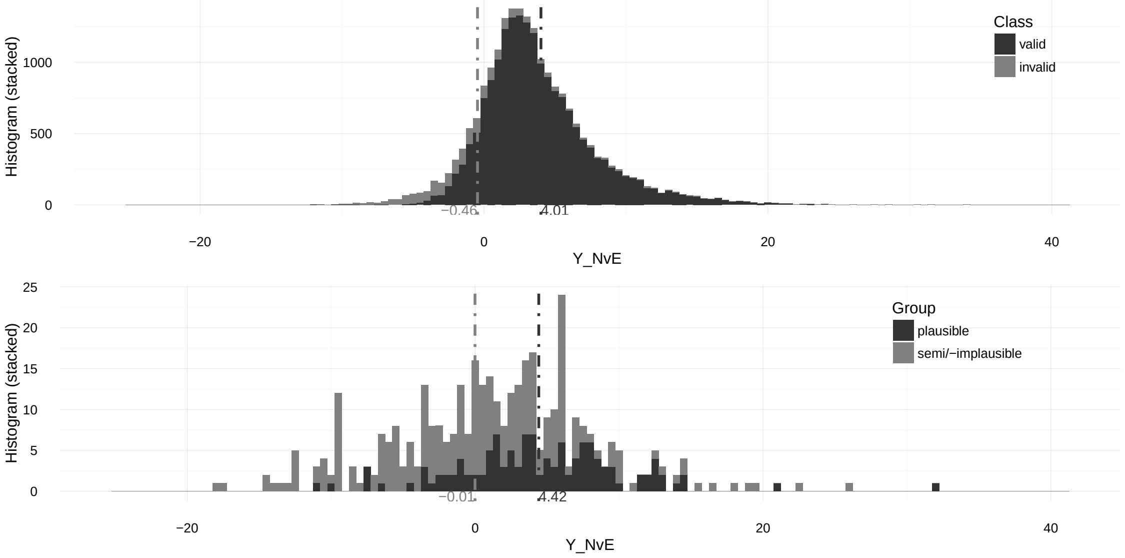


Figure 5.4: Stacked histogram for $\Upsilon_i(\mathcal{N}, \Sigma)$ for online questionnaire sample (top) and experimental study sample (bottom).

again for the online questionnaire study sample, we observe a discrepancy between mean values -0.46 (invalid) and 4.01 (valid) between the member of the two class members. However, it does not seem to improve discriminatory power when compared with $D_i^2(\Sigma)$ values between valid and invalid class members. The histogram for the experimental study sample, on the other hand, suggests a large increase in discriminatory power between members of experimentally induced groups. Semi/-implausible group members are assigned extremely small values by $\Upsilon_i(\mathcal{N}, \Sigma)$ with a mean close to 0, where plausible group members are pre-dominantly distributed around their mean value of 4.42.

These findings empirically substantiate the interpretations of $\mathcal{T}_i$ in the previous Section 5.1. To reiterate, where a two-sided cut-off to identify SpRPs is justified, the alternative interpretation suggested a left-sided cut-off for extreme small values to be the most conservative approach towards detection of SpRPs while keeping the risk of incorrectly as extreme identified plausible response patterns low.

Furthermore, above histograms suggest that components in $\Upsilon_i(\mathcal{N}, \Sigma)$ allow for a standardised interpretation of deviation from the hypothesised valid response model. The different means of the two classes or groups, respectively, are similar in $\Upsilon_i(\mathcal{N}, \Sigma)$ amongst valid and plausible responders, as well as amongst invalid and semi/-implausible responders. In fact, in the following Section 5.4, I will derive the theoretical distribution for the non-constant part $\mathcal{T}_i$ of $\Upsilon_i(\mathcal{N}, \Sigma)$, which derives at the conclusion that expected value always satisfies $\mathbf{E}[\mathcal{T}_i] = 0$ for valid responses under the theoretical valid response model.

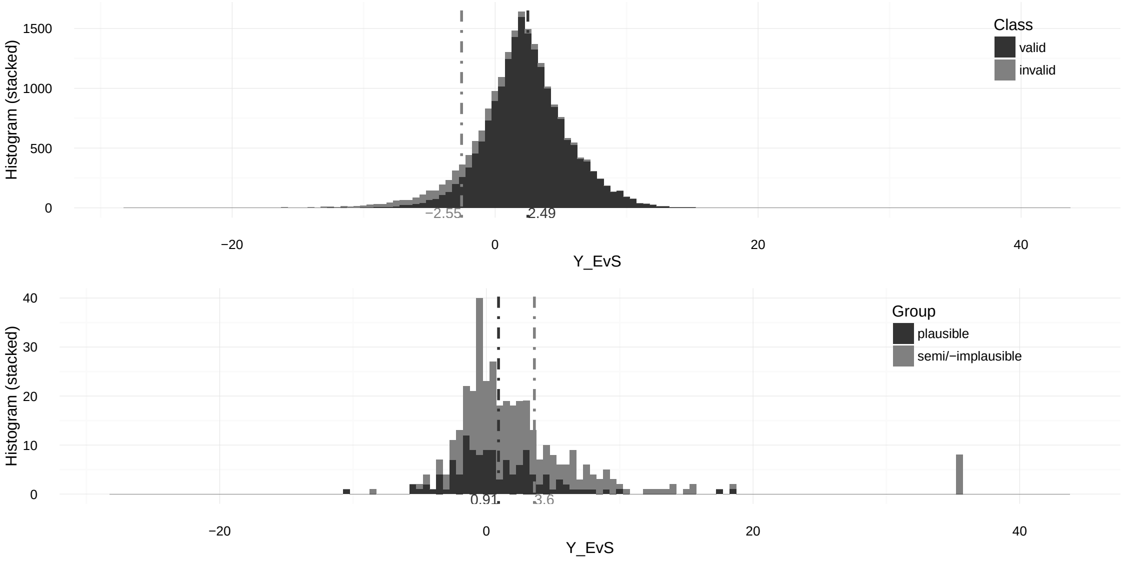


Figure 5.5: Stacked histogram for $\Upsilon_i(\Sigma, S)$ for online questionnaire sample (top) and experimental study sample (bottom).

In contrast to the proposed modification, the original version $\Upsilon_i(\Sigma, S)$ of the identification measure does not provide a large discriminatory power. Corresponding histograms for $\Upsilon_i(\Sigma, S)$ can be found in Figure 5.5. For the online questionnaire study sample, the discriminatory power does not seem to be largely different from what we can see using $\Upsilon_i(\mathcal{N}, \Sigma)$. However, the mean values for the experimental study sample groups have a smaller difference between the plausible, 0.90, and the semi/-implausible, 3.60, groups. Furthermore, an appropriate method of detection for members of the semi/-implausible group members seems to be a right-sided cut-off in $\Upsilon_i(\Sigma, S)$. In contrast to that, the online questionnaire study sample histogram would suggest a left-sided cut-off. Interpretation of $\Upsilon_i(\Sigma, S)$ does not seem to be as clear as it is the case for the modified version $\Upsilon_i(\mathcal{N}, \Sigma)$.

In conclusion, histograms in this section suggest that neither of $\Upsilon_i(\mathcal{N}, \Sigma)$ components, $D_i^2(\Sigma)$ nor $D_i^2(\mathcal{N})$, are effective test measure for the detection of SpRPs.

However, their difference in $\mathcal{T}_i$ indicates to be powerful tool in separating semi/-implausible from plausible responses. Graphically investigating the distribution of $\Upsilon_i(\Sigma, S)$ and the modified version $\Upsilon_i(\mathcal{N}, \Sigma)$ seem to justify the modification with regards to discriminatory power. In the following Section 5.3.2, I will discuss discriminatory power and the choice of cut-off values for the new and original version of the proposed detection measure in more detail.

5.3.2 Comparing Results of original and modified Versions

The following evaluation procedures will be applied on four main scenarios: In the first scenario (JpH), I first estimate parameters for the theoretical model presented in Chapter 3 and Figure 3.1 using the sample of the online questionnaire study. In the next step, the estimated factor loadings ($\hat{\lambda}_{jk}$) from the first step and means (fixed to $\mu_k = 0$), variances (set to $\sigma_{kk} = 1$) and covariances ($\hat{\sigma}_{km}$ for $k \neq m$) of latent variables $\boldsymbol{y}$, are used to calculate the model implied covariance matrix using the sample of the experimental study.

In the second scenario (C14pH), parameters are estimated using only the plausible responding sub-group of the experimental study, namely member of cells 1 and 4. Estimated parameters are then adapted for reproducing the model implied covariance matrix using the entire sample. The third scenario (C14r10%pH) consists of an artificial sub-group including cells 1 and 4 but this time also a randomly chosen sample out of the remaining semi-plausible conditions, cells 2, 3, 5, and 6, such that the latter constitutes 10% of the total sample for the estimation of parameters. Averaged parameters over 1000 repetitions are then adapted for reproducing the model implied covariance matrix using the entire sample. Lastly in the fourth scenario (HpH), the complete experimental study sample is used for estimation and reproducing the model implied covariance matrix.

JpH represents a scenario in which we have access to another study with comparable sample. Optimally, we would like to assume that a second study sample is (mostly) free of invalid responses and produces estimates close to the true population parameter values. In such a case, we can use those estimates in order to decrease the influence of invalid responses on the information source used for the new test measure. This should help to identify semi-/implausible response patterns in the experimental study sample more successfully. However, we might be confronted with a situation in which we do not have access to another study sample, or we cannot assume measurement equivalence between two different samples. The samples might not be random draws from the same population, or represent incomparable groups from a common population but not representable under the same measurement model.

The following three scenarios serve the purpose of evaluating the new test measure in settings with only one sample. HpH is the most conservative scenario where only a theoretical measurement model structure is hypothesised and no further information about true parameter values nor group membership is known. With scenario C14pH, I would like to evaluate whether knowing a small sub-group to be plausible can help to discriminate between valid and invalid responses in the entire study sample. Similarly, in scenario labelled C14r10%pH, I simulate a situation in which the experimental study sample consists of only 10% semi/-implausible response patterns. The goal is to identify the discrimination power of the new test measure in a setting in which the experimental study sample is not dominated by nearly 68% semi/-implausible responses. Although, several estimates based on randomly drawn sub-groups from the semi/-implausible response groups are averaged for this procedure, this is not a perfectly representative simulation because the total sample size is smaller than $n = 380$.

Binary-logistic Regression

To test the hypothesis from the first section of this chapter about the magnitude and direction of the three possible contrast components $C_i(\mathcal{N})$, $C_i(\Sigma)$, and $C_i(S)$ to distinguish between plausible and semi-plausible response patterns, I use a binary-logistic regression model to differentiate between groups from the experimental study sample. I argued for the use of $C_i(\mathcal{N})$ and $C_i(\Sigma)$ instead of $C_i(S)$ as components for the new measure for the detection of SpRPs. Using binary-logistic regression, we can empirically investigate which components have the largest contribution in predicting plausible versus semi/-implausible group membership.

Let G_i be a random variable where $G_i = 1$ indicates membership of the semi-/implausible responding sub-sample (cells 2, 3, 5, and 6) and $G_i = 0$ denotes being a member of the plausible responding sub-sample (cells 1 and 4). The elements of $\boldsymbol{g}$ denote the group membership of each sample point of the experimental study sample which we see as outcomes of G_i. Table 5.1 shows the results of a binary-logistic regression modelled such that

$$\operatorname{logit}[\Pr(G_i = 1 | C_i(\mathcal{N}), C_i(\Sigma), C_i(S))] = \beta_0 + \beta_N C_i(\mathcal{N}) + \beta_T C_i(\Sigma) + \beta_S C_i(S) \quad (5.29)$$

where β_0 is the intercept and other β quantities are the regression coefficients.

Table 5.1: Results of the binary-logistic regression of experimental study sub-sample membership on three contrast components, where parameters for the contrast components are estimated using different samples (evaluation scenarios)

Contrast		Evaluation scenario							
component		JpH		C14pH		C14r10%pH		HpH	
$C_i(\mathcal{N})$	β_N	0.23	***	0.24	***	0.22	***	0.15	**
	SE	(0.05)		(0.05)		(0.05)		(0.05)	
$C_i(\Sigma)$	β_T	−0.34	***	−0.41	***	−0.38	***	−0.21	*
	SE	(0.08)		(0.09)		(0.09)		(0.09)	
$C_i(S)$	β_S	0.08		0.13		0.12		0.01	
	SE	(0.07)		(0.07)		(0.07)		(0.07)	

Sign. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$
Note The intercepts β_0 were omitted. SE denotes the standard error.
JpH Online questionnaire sample.
HpH Experimental study sample estimates.
C14pH Plausible response sub-sample of the experimental study.
C14r10%pH Plausible response sub-sample and a randomly drawn small portion of invalid responses.

We can see that throughout all four scenarios, JpH, C14pH, C14r10%pH, and HpH, the null model and theoretical model components contribute the most (significant β coefficients) for the prediction of $\boldsymbol{g}$. The contrast component under the saturated model does not contribute significantly towards a better prediction of response pattern plausibility. Furthermore, we observe that the regression coefficients for the $C_i(\mathcal{N})$ and $C_i(\Sigma)$ are of opposite signs. Hence, larger values in $C_i(\mathcal{N})$ and smaller values in $C_i(\Sigma)$ indicate a more likely semi/-implausible response group membership.

Consequently, response patterns that show a larger deviation from the null model and smaller deviation from the theoretical model are more plausible. These findings are logically in line with our interpretation and the experimental study sample group membership.

Discriminatory Power

We have seen that the combination of the components within the new test measure $\Upsilon_i(\mathcal{N}, \Sigma)$ significantly contributes to predicting plausibility of response patterns. In order to judge the magnitude of discriminatory power provided by $\Upsilon_i(\mathcal{N}, \Sigma)$ versus $\Upsilon_i(\Sigma, S)$ I will fix a tolerance level for the misclassified valid responses and investigate the success of correctly identified invalid response patterns by either of the identification measures using a one-sided cut-off. The choice of left-/ versus right-sided cut-off is carried out such that the side with the largest discriminatory power is chosen. In this section, the actual cut-off values are chosen based on known group membership and set tolerance level. Histograms for both measure in Section 5.3.1 visually suggested a right-sided cut-off for $\Upsilon_i(\Sigma, S)$ and a left-sided cut-off for $\Upsilon_i(\mathcal{N}, \Sigma)$.

Results in Table 5.2 show the percentage of semi-/implausible response patterns, that are identified as extreme values when a cut-off value is chosen such that we only allow for 10% of plausible response patterns to be incorrectly identified as extreme values. Furthermore, the rows distinguish between results of the evaluation scenarios of this section. This procedure has been implemented for the original person-fit index $\Upsilon_i(\Sigma, S)$ and the new measure $\Upsilon_i(\mathcal{N}, \Sigma)$. Performance comparisons between both measures reveal consistent results throughout all of the four data-model estimation scenarios. For $\Upsilon_i(\Sigma, S)$ chosen cut-off values range between 3.55 and 4.20 such that only 10% of valid response patterns are incorrectly classified as invalid. Classifying response patterns as invalid that were assigned larger values than their respective cut-off only identifies between 9% and 14% invalid response patterns. Significantly better results were achieved with $\Upsilon_i(\mathcal{N}, \Sigma)$. Chosen cut-off values have a range between -0.89 and -0.93 and detect between 30% and 36% of Cell 2, 3, 5, and 6 members. Scenario JpH shows the best performance with 36% correctly classified invalid response patterns but even the worst case scenario HpH does well with 30%

Table 5.2: Percentage of correctly classified semi/-implausible sub-sample members of the experimental study sample, where parameters for identification measure are estimated using different samples (evaluation scenarios)

Evaluation scenario	Identification measure $\Upsilon_i(\Sigma, S)$			$\Upsilon_i(\mathcal{N}, \Sigma)$		
	% SpRP	$>$	Cut-off	% SpRP	$<$	Cut-off
JpH	13		4.09	36		-0.93
C14pH	14		3.84	35		-0.91
C14r10%pH	14		3.55	33		-0.90
HpH	09		4.20	30		-0.89

Note Cut-off values chosen such that maximum 10% of valid responses are incorrectly classified as SpRPs.

correctly identified invalid response patterns.

Table 5.3 provides a more detailed view on which of the cells' members are flagged as invalid with cut-off values depicted in Table 5.2 for the scenarios using the new measure $\Upsilon_i(\mathcal{N}, \Sigma)$. In general, we see similar patterns in percentage of flagged group members throughout all four scenarios. The least plausible conditions referenced as 'implausible' have 32% to 45% flagged members. These results are in line with our hypotheses since those response patterns are the result of clearly invalid response strategies. The most salient feature is whether less obvious response strategies are as successfully identified as such. Semi-plausible conditions have between 25% and 35% of members whose response patterns were flagged as invalid. Unsurprisingly, plausible conditions have no more than 10% members incorrectly flagged as invalid responders since this is the a priori set cut-off criterion. The reader should be reminded that plausible and semi/-implausible response groups do not consist of equal sample sizes. Table in A.2 provides actual numbers of identified responders. We can see that we would incorrectly identify 12 plausible responders but, in the best evaluation scenario (JpH), detect 93 semi/-implausible responders, and in even in the least successful evaluation scenario (HpH) we would detect 77 semi/-implausible responders.

Investigating flag percentages per cell in more detail, we can see that overall cells 4, 5, and 6 as part of first factor manipulation have more members flagged invalid as their respective partner cells 1, 2, and 3. This is surprising given that the first

Table 5.3: Percentage of sub-sample members of the experimental study sample identified as extreme values in $\Upsilon_i(\mathcal{N}, \Sigma)$, where parameters for $\Upsilon_i(\mathcal{N}, \Sigma)$ are estimated based on different samples (evaluation scenarios)

Sub-sample	Cell(s)	Evaluation scenario			
		JpH	C14pH	C14r10%pH	HpH
Plausible	1	.13	.13	.13	.15
	4	.08	.08	.08	.07
Semi-plausible	2	.30	.35	.30	.26
	5	.28	.30	.27	.25
Implausible	3	.45	.44	.44	.38
	6	.41	.33	.35	.32
Plausible	1,4	.10	.10	.10	.10
Semi-plausible	2,5	.29	.32	.28	.26
Implausible	3,6	.43	.38	.38	.35
Semi-/implausible	2,3,5,6	.36	.35	.33	.30

Note Percentages of response patterns' value smaller than cut-off value in respective scenario.

three cells were part of the experimentally manipulated factor one warning conditions (see Section 3.2.2). One explanation for this is the artificially careful responding behaviour that is not in line with the theoretical valid response model which we would expect in the online questionnaire study. This hypothesis is substantiated by the fact that cell 4 is least subject to flagging when the online questionnaire study sample is used to estimate the model parameters. For the remaining three evaluation scenarios this can be explained by a possible correlation between $\Upsilon_i(\mathcal{N}, \Sigma)$ and sub-group size ($n = 39$ cell 1 versus $n = 84$ cell 4), especially because discrepancies of flagged cell members between cell 1 and 4 are largest in scenario HpH, 15% versus 7%, than in the other three scenarios, e.g. JpH, 13% versus 8%. Overall the warning condition comprises only 40% of the experimental study sample, rendering the normal instruction condition more effective in the model estimation process.

In conclusion, we have seen that the components $C_i(\mathcal{N})$ and $C_i(\Sigma)$ in the modified identification measure $\Upsilon_i(\mathcal{N}, \Sigma)$ are with regards to the empirical data superior to the linear combination of $C_i(\Sigma)$ and $C_i(S)$ in the original version of the test measure

$\Upsilon_i(\Sigma, S)$. With $\Upsilon_i(\mathcal{N}, \Sigma)$ we successfully identified nearly three times as many members of the semi/-implausible group than with the original index. Cut-off values for the identification of extreme values were set based on an arbitrary tolerance level with regards to plausible responders incorrectly identified as extreme. Hence as a next step, it is sensible to derive the theoretical distribution for the identification measure, such that we are able to estimate cut-off values while controlling for the risk of incorrectly detected valid responses, when the actual group membership is unknown.

5.4 Deriving the theoretical Distribution

To derive the theoretical distribution of $\mathcal{T}_i$ under the hypothesis that the valid response model holds, I will briefly revise the development of the identification measure from previous chapters, linking it to a log-likelihood ratio test. Ultimately, through linear transformation into its quadratic form, I will proof the new measures' distribution to be a linear combination of centralised χ^2 variables.

Log-Likelihood Ratio Test One of the most common procedures when deciding whether an alternative model provides a better fit to the data than a comparable/ nested null model is the log-likelihood ratio test:

$$\text{LRT} = 2\ln(\frac{\text{Likelihood alternative model}}{\text{Likelihood null model}}) \tag{5.30}$$

If the alternative model significantly improves the likelihood of the data at hand, we usually reject the null model and decide in favour for the alternative model.

In a multivariate normal variables setting, when the null model is the independence model with covariance matrix $\mathcal{N}$ and the alternative model is defined by the model implied covariance matrix Σ, (5.30) becomes

$$\begin{aligned}
& 2\ln\left[\prod_{i=1}^{n}\frac{1}{\sqrt{(2\pi)^p\cdot|\Sigma|}}\times\exp^{-\frac{D_i^2(\Sigma)}{2}}\right]- \qquad 2\ln\left[\prod_{i=1}^{n}\frac{1}{\sqrt{(2\pi)^p\cdot|\mathcal{N}|}}\times\exp^{-\frac{D_i^2(\mathcal{N})}{2}}\right]\\
= {} & 2\left[\sum_{i=1}^{n}-\frac{1}{2}\left(p\ln(2\pi)+\ln|\Sigma|+D_i^2(\Sigma)\right)\right]-2\left[\sum_{i=1}^{n}-\frac{1}{2}\left(p\ln(2\pi)+\ln|\mathcal{N}|+D_i^2(\mathcal{N})\right)\right]\\
= {} & \qquad\left[\sum_{i=1}^{n}p\ln(2\pi)+\ln|\mathcal{N}|+D_i^2(\mathcal{N})\right]- \qquad \left[\sum_{i=1}^{n}p\ln(2\pi)+\ln|\Sigma|+D_i^2(\Sigma)\right]
\end{aligned}$$

where $D(\Sigma)$ is the Mahalanobis distance as defined in (2.1) based on the covariance matrix implied by the alternative model and $D(\mathcal{N})$ is based on the covariance matrix implied by the null model. Furthermore, we have

$$\text{LRT} \;\sim\; \chi^2(df_{\text{Alt. M.}} - df_{\text{Null M.}}), \tag{5.31}$$

under the null model, theoretically following a χ^2 distribution with degrees of freedom set by the difference of parameters in the null model and the alternative model.

Individual Log-Likelihood Contrast A different setting often practised consists of setting the alternative model as saturated model and the null model as a hypothesised (more parsimonious) model. Hence, a significant test result of (5.30) would be interpreted such that the hypothesised model significantly decreases the likelihood of the data under the hypothesised model, or in other words, does not fit the data well.

In Section 2.1.3, I introduced Reise and Widaman (1999) idea of using the individual log-likelihood contribution, the i^{th} addend in (2.31), as a measure of an individual data point's contribution towards the overall model misfit. However, in Section 5.1, I further concluded that analysing a hypothesised model assuming there are no invalid responders present will lead to spurious correlations between variables. This, in turn, will produce results for the hypothesised model that assimilates the saturated model. Ultimately, the power to detect data points that do not follow the hypothesised model will be vanishingly small.

Instead, I proposed to contrast the hypothesised model with the independence model (no correlation between variables) and by further simplifying (5.30) to

$$\text{LRT} = \sum_{i=1}^{n} \left[\ln |\mathcal{N}| - \ln |\Sigma| + D_i^2(\mathcal{N}) - D_i^2(\Sigma) \right], \tag{5.32}$$

I defined

$$\begin{aligned} \Upsilon_i(\mathcal{N}, \Sigma) = &\quad \ln |\mathcal{N}| - \ln |\Sigma| + D_i^2(\mathcal{N}) - D_i^2(\Sigma) \\ = &\quad \text{constant} + D_i^2(\mathcal{N}) - D_i^2(\Sigma). \end{aligned} \tag{5.33}$$

Thus, $\Upsilon_i(\mathcal{N}, \Sigma)$ is an individual's contribution to the LRT statistic in (5.32).

The Test Measure I have extensively reviewed the new test measure in Section 5.1 and theoretically interpreted its components. For the derivation of its theoretical distribution, I will briefly recall the central results. As the test measure for the identification of invalid response patterns I focused on the components of $\Upsilon_i(\mathcal{N}, \Sigma)$ that are i dependent and defined the new test measure

$$\begin{aligned} \mathcal{T}_i(\boldsymbol{x_i}) = &\quad D_i^2(\mathcal{N}) - D_i^2(\Sigma) \\ = &\quad (\boldsymbol{x_i} - \boldsymbol{\mu})^T \mathcal{N}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}) - (\boldsymbol{x_i} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}) \\ = &\quad (\boldsymbol{x_i} - \boldsymbol{\mu})^T (\mathcal{N}^{-1} - \Sigma^{-1}) (\boldsymbol{x_i} - \boldsymbol{\mu}). \end{aligned} \tag{5.34}$$

It was previously concluded, that extreme values in $\mathcal{T}_i$ can occur in two ways: Extreme negative values are the result of $D_i^2(\mathcal{N}) < D_i^2(\Sigma)$, whereas a extreme positive value for i results from $D_i^2(\mathcal{N}) > D_i^2(\Sigma)$. For instance, if an outlier in the standardised multivariate distribution that is defined by the covariance matrix Σ is not an outlier with regards to $\mathcal{N}$, it would be assigned a negative value.

I have shown in (5.15) that $\mathcal{T}_i$ can under a factor analyses model with simple factor structure be written as

$$T_i = \sum_{k=1}^{q} y_{ik}^{*2} + \sum_{k=1}^{q} w_{y,k}\, y_{ik}^{*2} - \sum_{j=1}^{p} w_{x,j}\, (x_{ij} - \mu_j)^2, \tag{5.35}$$

where

$$w_{y,k} = \sum_{j=1}^{p} \frac{\lambda_{jk}^2}{\psi_j}, \qquad\qquad w_{x,j} = \frac{\sigma_{jj}}{\psi_j} - 1. \tag{5.36}$$

Hence, i is given an unweighted and weighted, $w_{y,k}$, penalty on extreme values assigned by the model, y_{ik}^{*2}, and a weighted, $w_{x,j}$, penalty on univariate variation in centralised observed scores, $(x_{ij} - \mu_j)^2$. However, those two kinds of penalties are of opposite sign. Weights are influenced by some ratio of explained variance, $\sigma_{jj} - \psi_j$ as total variance minus error variance, or factor loadings λ_{jk} versus error variance ψ_j.

Ultimately, I concluded in Section 5.1.2 that $\mathcal{T}_i$ leads to the emergence of very beneficial properties for the identification of invalid response patterns when used in a latent variables framework. Extreme negative values in $\mathcal{T}_i$ indicate in sum large univariate outliers in observed variables that do not result from or are not justified by, respectively extreme latent variable scores. Vice versa, extreme positive values in $\mathcal{T}_i$ assigned by the model can be the result of extreme latent variable scores that do not match the corresponding univariate deviance in the response pattern. Hence, extreme latent variable scores do not automatically lead to a total penalty for i. The most important caveat of other identification measures is the fact that valid but extreme factor scores lead to extreme person-fit values (see conclusions drawn from the review in Section 2.1). $\mathcal{T}_i$ has properties that help to eliminate this problem. In $\mathcal{T}_i$, extreme latent variable scores can counterbalance a deviant response pattern and reduce the penalty. Therefore, we would anticipate an expected value of 0 for valid responses where an aberrance from the model equals the aberrance of the corresponding response pattern. Furthermore, penalties for a response pattern leading to either extremely small or extremely large values in $\mathcal{T}_i$ are moderated by how reliable the measurement model is. Hence, penalties are only large (providing more certainty) when we have an at least minimally accurate hypothesised model that is able to capture the theoretical model that underlies valid responses.

Estimated Components We do not in practice know $\mathcal{N}$, Σ, and $\boldsymbol{\mu}$. They are estimated from the observed data, under the assumption that the model is specified correctly with only valid responses in the sample. Hence in case, we have a contaminated sample with invalid responses present and do not account for in the model, the estimation will be biased. The estimates are defined under the latent variable model framework in Section 2.1.3. $\mathcal{T}_i$ under a factor analysis model is discussed in Section 5.1 in detail and briefly revised above. Furthermore, depending on the actual Σ and the number of invalid responses in the sample, estimates might be more or less accurate in different situations. In Chapter 7 I will draw conclusion on optimal situations based on simulation results.

The Quadratic Form In order to test for extreme values, it is important to derive the theoretical distribution of this new identification measure. By definition, we do not know the numerous invalid response strategies involved that lead to invalid responses. However, we can derive the theoretical distribution of a test measure for valid responses given known (not estimated) values of parameters. Ultimately, this will give us an indication of what the distribution should look like if we had only valid responses.

For the components of $\mathcal{T}_i$ we know $D_i^2(\Sigma) \sim \chi^2(p)$ and $D_i^2(\mathcal{N}) \not\perp\!\!\!\perp D_i^2(\Sigma)$. Because we do not know the theoretical distribution of $D_i^2(\mathcal{N})$ and both components are correlated, deriving the theoretical distribution of $\mathcal{T}_i$ is not straightforward. However, we can transform $\mathcal{T}_i$ such that we have a quadratic form of centralised normal random variables $\boldsymbol{\delta_i} = \boldsymbol{x_i} - \boldsymbol{\mu}$:

$$\mathcal{T}_i = \boldsymbol{\delta_i}^T(\mathcal{N}^{-1} - \Sigma^{-1})\boldsymbol{\delta_i} \tag{5.37}$$

We can further transform (5.34) such that we have quadratic form of multivariate standard normal random variables $\boldsymbol{z_i} = \Sigma^{-\frac{1}{2}}\boldsymbol{\delta_i}$ and write

$$\begin{aligned} \mathcal{T}_i = &\quad \boldsymbol{\delta_i}^T\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}(\mathcal{N}^{-1} - \Sigma^{-1})\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\boldsymbol{\delta_i} \\ \mathcal{T}_i = &\qquad\qquad \boldsymbol{z_i}^T\Sigma^{\frac{1}{2}}(\mathcal{N}^{-1} - \Sigma^{-1})\Sigma^{\frac{1}{2}}\boldsymbol{z_i} = \quad \boldsymbol{z_i}^T\mathcal{A}\boldsymbol{z_i}, \end{aligned} \tag{5.38}$$

where $\mathcal{A} = \Sigma^{\frac{1}{2}}(\mathcal{N}^{-1} - \Sigma^{-1})\Sigma^{\frac{1}{2}}$. $\mathcal{T}_i(\boldsymbol{x_i})$ may then be expressed as a quadratic form of p random variables $\boldsymbol{z_i}$ defined by $\mathcal{A}$. We can write

$$\mathcal{T}_i = \boldsymbol{z_i}^T\,\mathcal{A}\,\boldsymbol{z_i} = \boldsymbol{z_i}^T\,\mathcal{W}\Gamma\mathcal{W}^T\,\boldsymbol{z_i} \tag{5.39}$$

using the spectral decomposition of $\mathcal{A}$, where Γ is the diagonal matrix of eigenvalues with elements γ_j and $\mathcal{W}$ is the orthogonal matrix with the eigenvectors as its columns. Ultimately, we can define random variables $\boldsymbol{u_i} = \mathcal{W}^T \boldsymbol{z_i}$ which are mutually independent standard normal variables, with identity covariance matrix and expectation vector 0. This follows because

$$\mathbf{Var}[\boldsymbol{u_i}] = \mathcal{W}\ \mathbf{Var}[\boldsymbol{z_i}]\ \mathcal{W}^T = \mathcal{W}\mathcal{W}^T = \mathrm{I}. \tag{5.40}$$

Therefore, we can further simplify $\mathcal{T}_i$ to

$$\mathcal{T}_i = \sum_{j=1}^{p} \gamma_j u_{i,j}^2, \tag{5.41}$$

which is a linear combination of independent squared standard normal variables, $u_i \sim N(0,1)$, or respectively, independent χ^2 variables, $u_i^2 \sim \chi^2(1)$, with one degree of freedom.

The Distribution $\mathcal{T}_i$ follows the distribution of a linear combination of independent $\chi^2(1)$ variables with only parameters $\boldsymbol{\gamma}$ and non-centrality parameters all 0. Given that

$$u_i^2 \sim \chi^2(1) \sim Gamma(\frac{1}{2}, 2)$$

and

$$\gamma u_i^2 \sim Gamma(\frac{1}{2}, 2\gamma),$$

for this scalar random variable, I define the characteristic function as

$$\varphi(t) = \prod_{j=1}^{p} \varphi_{\gamma_j u_{i,j}^2}(t) = \prod_{j=1}^{p} (1 - \beta_j it)^{-\alpha} = \prod_{j=1}^{p} (1 - 2\gamma_j it)^{-\frac{1}{2}}, \tag{5.42}$$

where i is the imaginary unit, and $t \in \mathbb{R}$ is the argument of the characteristic function $\varphi(t)$. Having derived the characteristic function, the behaviour and properties of $\mathcal{T}_i$'s probability distribution is represented based on a one-to-one correspondence. For a random scalar variable we can simply use the inversion theorem and, correspondingly, define

$$f(u) = F'(u) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itu}\varphi(t)dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itu}\prod_{j=1}^{p}(1-2\gamma_j it)^{-\frac{1}{2}}dt, \quad (5.43)$$

as the probability density of $\mathcal{T}_i$. Box (1954) and Imhof (1961) were able to obtain the probability density by integration of the inversion formula for a special case: a linear combination of centralised χ^2 variables with even degrees of freedom. However more importantly, with the characteristic function uniquely defining the cumulative distribution function, we can also directly define it by, for instance, using the inversion theorem of Gil-Pelaez (1951):

$$F(u) = \frac{1}{2} + \frac{1}{2\pi}\int_{0}^{+\infty}\frac{e^{itu}\varphi(-t) - e^{-itu}\varphi(t)}{it}dt, \quad (5.44)$$

Imhof (1961) rewrites (5.44), such that we can numerically integrate over a finite range of $0 \leq t \leq T$, where the upper bound can be chosen based on the tolerance of approximation error (cf. Davies, 1973). This is just one of several methods for these kinds of numerical inversion of the characteristic function (cf. Bohman, 1975; Waller, Turnbull, and Hardin, 1995).

Based on the cumulant generating functions

$$\mathcal{K}_j(t) = \ln(1-2\gamma_j t)^{-\frac{1}{2}} = -\frac{1}{2}\ln(1-2\gamma_j t) \quad (5.45)$$

and, hence,

$$\mathcal{K}(t) = \sum_{j=1}^{p}\mathcal{K}_j(t) = -\frac{1}{2}\sum_{j=1}^{p}\ln(1-2\gamma_j t), \quad (5.46)$$

we can define the s^{th} cumulant of $\mathcal{T}_i$ as

$$\kappa_s = 2^{s-1}(s-1)!\sum_{j=1}^{p}\gamma_j^s. \quad (5.47)$$

Using the first four cumulants, we can derive the expected value, variance, skewness, and kurtosis of $\mathcal{T}_i$. For instance, we have:

$$\mathbf{E}[\mathcal{T}_i] = \mathbf{E}[\sum_{j=1}^{p} \gamma_j u_{i,j}^2] = \sum_{j=1}^{p} \gamma_j \mathbf{E}[u_{i,j}^2] = \sum_{j=1}^{p} \gamma_j \tag{5.48}$$

$$\mathbf{Var}[\mathcal{T}_i] = \mathbf{Var}[\sum_{j=1}^{p} \gamma_j u_{i,j}^2] = \sum_{j=1}^{p} \gamma_j^2 \mathbf{Var}[u_{i,j}^2] = 2\sum_{j=1}^{p} \gamma_j^2. \tag{5.49}$$

Based on the result in (5.48), we can prove that $\mathbf{E}[\mathcal{T}_i] = 0$. Using the fact that

$$\sum_{j=1}^{p} \gamma_j = \text{trace}[\mathcal{A}] \quad \text{and} \quad \text{trace}[\Sigma\mathcal{N}^{-1}] = p,$$

we then write

$$\begin{aligned} \mathbf{E}[\mathcal{T}_i] &= \text{trace}[\mathcal{A}] \\ &= \text{trace}[\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}(\mathcal{N}^{-1} - \Sigma^{-1})] \\ &= \text{trace}[\Sigma\mathcal{N}^{-1}] - \text{trace}[\Sigma\Sigma^{-1}] \\ &= p - p = 0. \end{aligned}$$

Computational Implementation In general, there are several computational implementations available for this distribution. One of the first more practical methods was developed by Box (1954) who sought to numerically invert the characteristic function of a quadratic form of similar kind. However, amongst other restrictions, this method is only applicable to a linear combination of central χ^2 variables with an even number of degrees of freedom. Imhof (1961) gives exact and approximate methods for computing the distribution of quadratic forms in normal variables. One of those methods includes numerically inverting the characteristic function, as is also proposed by Davies (1973) (for algorithm, see Davies, 1980). Imhof (1961) finds numerical inversion to perform better than Pearson's three-moment central χ^2 approximation in these situations. Sheil and O'Muircheartaigh (1977) and Farebrother (1984) take advantage of the fact that the distribution can be written as an infinite sum of central chi-squared variables. This approach is based on findings in Ruben (1962). Farebrother (1990) proposed a method which expresses a quadratic form in an alternative form, using the so-called tridiagonal form. Kuonen (1999) utilises saddlepoint approximations. Liu, Tang, and Zhang (2009) approximate the distribution using a noncentral χ^2 distribution where the degrees of freedom and the non-centrality parameter are calculated using the first four cumulants.

However, almost all of the above approaches, with the exception of Imhof (1961) and Davies (1973), produce differing results or are only applicable to non-negative linear combinations (e.g., in our case $\gamma_j > 0$ for all j). Duchesne and Lafaye De Micheaux (2010) empirically compared the performance of several approaches and provides an R statistics software package implementing some of the previously mentioned methods (see package *CompQuadForm*, Duchesne and Lafaye De Micheaux, 2010).

In our case, we always have a linear combination of central χ^2 variables with the same degrees of freedom, $df = 1$. Computing the cumulative probabilities in our simpler case is computationally not intensive. Hence, I will use the software implementation of Imhof's exact method, which allows us to bind the approximation error such that we could make it arbitrarily small.

Chapter 6

Evaluation of the new Measure

So far, two samples from an experimental and an online questionnaire study were analysed with a latent variable model assuming no invalid responses in the sample. I further analysed both samples, with a latent class model in which an invalid class was defined to accommodate one type of invalid response strategy into the model. Therefore, we have two different valid response models: one response model derived via the latent variable model and another derived using latent class analysis. Furthermore, the latent class model provided us with an invalid response model that is based on the item wording response strategy. Using these three different response models, I will further evaluate the new detection measure for SpRPs.

This chapter seeks to provide a numerical exercise and an empirical evaluation of $\mathcal{T}_i$ for valid and invalid responses based on information sources that are derived in different ways. I will empirically investigate $\mathcal{T}_i$ for valid and invalid responses based on randomly generated data under the latent class response models as true population distributions. The focus lies on comprehending the valid versus invalid response behaviour in key quantities of $\mathcal{T}_i$ (analytically discussed in Section 5.1.2). I will provide summary statistics and visually investigate $\mathcal{T}_i$. Doing so, we will be able to spot distributional changes in $\mathcal{T}_i$ for valid responses when estimated components are used as information source instead. In this context, it seems sensible to further experiment with different information sources for $\mathcal{T}_i$ as identification instrument for SpRPs in the experimental study sample. I will employ the combined approach towards detection as discussed in Section 5.2. Instead of only relying on the valid response model estimated using latent variable analysis, I will feed $\mathcal{T}_i$ with the valid response model estimates based on results of the latent class analysis. In doing so, it shall be determined if the more accurate estimates from the valid response class model form a better information source for $\mathcal{T}_i$ as detection measure for semi/-implausible group members of the experimental study sample.

6.1 Numerical Example using Latent Class Response Models

After having derived the theoretical distribution, I will discuss and visualise a numerical example of $\mathcal{T}_i$ and compare the empirical results, i.e. distribution, mean, and variance, with the expected theoretical results. I will use the Big Five latent variable model introduced in Section 3.3. Estimated models, namely latent class model parameters for valid and invalid responses and latent variable model parameters assuming no invalid responses present, discussed in Chapter 3, will serve as examples for this numerical exercise. Parameters based on the online questionnaire data of the Johnson (2005) study have been chosen for this purpose (see Table in Appendix A.3). Valid responses and invalid responses are randomly drawn from their respective multivariate normal distributions defined by the corresponding latent class model parameters with sample size $n = 100,000$. The information source for the computation of $\mathcal{T}_i$ will, in the first scenario, be the theoretical Σ, $\mathcal{N}$, and $\boldsymbol{\mu}$ and, in the second scenario, their corresponding estimates $\hat{\Sigma}$, $\hat{\mathcal{N}}$, and $\hat{\boldsymbol{\mu}}$ obtained from a sample also consisting of, in the model unaccounted, invalid responses.

6.1.1 Valid Responses under the valid Response Model

First I will focus on valid responses, where $\mathcal{T}_i$ is estimated drawing on the theoretical parameters that were also used to define the multivariate distribution from which they were randomly drawn. Within this context, I will also investigate empirical results for the components of $\mathcal{T}_i$.

Results for $\mathcal{T}_i$

The empirical density function of $\mathcal{T}_i$ for the sample data is graphed in Figure 6.1. The curve is based on the kernel density estimates.

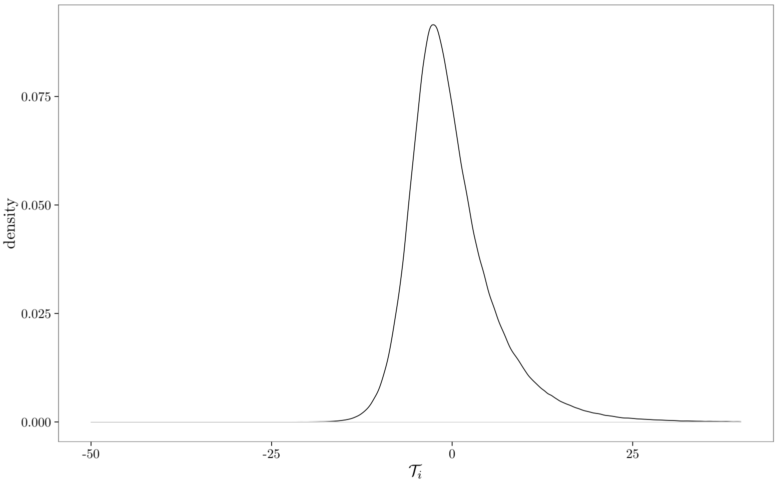


Figure 6.1: Kernel density estimates of $\mathcal{T}_i$ for valid responses based on the randomly drawn valid response model sample.

Furthermore, Figure 6.2 shows the empirical cumulative distribution curve compared to the theoretical cumulative distribution function of $\mathcal{T}_i$. The theoretical cumulative distribution function with parameters $\boldsymbol{\gamma}$ was estimated via the CompQuadForm computational implementation of Imhof's exact method, using pre-defined standard accuracy parameters (see Section 5.4). The distribution parameters are $\boldsymbol{\gamma}$ = (3.31, 1.83, 1.24, -0.14, -0.23, -0.29, -0.34, -0.35, -0.37, -0.38, -0.42, -0.42, -0.46, -0.48, -0.51, -0.58, -0.67, -0.75). Both curves lie on top of each other such that they are visually almost inseparable, suggesting exact fit.

The overall goal of deriving a theoretical distribution was to provide cut-off values for extreme values. Therefore, Table 6.1 shows empirical and theoretical 5^{th} and 95^{th} percentiles. In order to derive the percentiles from the theoretical distribution, a standard univariate optimisation procedure was employed. The algorithm was set to search for values of $\mathcal{T}_i$ that, when given as input to the computational implementation of Imhof's exact method, produce the cumulative probabilities 0.05 and 0.95. The first row of Table 6.1 shows the theoretical cumulative probability for optimised values, which one can find in the third row. The last row gives the difference

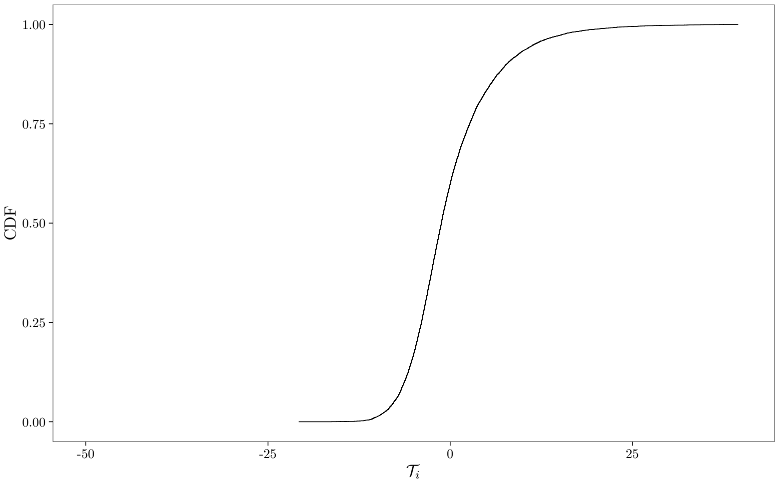


Figure 6.2: Cumulative distribution of $\mathcal{T}_i$ based on the randomly drawn valid response model sample and theoretical cumulative function based on parameters $\boldsymbol{\gamma}$.

between the theoretical and empirical results, which are reported in the second row. On can see that the optimisation provided us with the theoretical 5^{th} percentile

Table 6.1: Empirical and theoretical percentiles of $\mathcal{T}_i$ for valid responses under the theoretical valid response model

	5^{th} Percentile	95^{th} Percentile
Optimisation Accuracy	0.050	0.950
Empirical Cut-Off	-7.686	11.564
Theoretical Cut-Off	-7.682	11.512
Cut-Off Difference	-0.004	0.052

and 95^{th} percentile, which in turn gives us .05 and .95 as cumulative probability when estimated with Imhof's methods. The differences between the empirical and theoretical percentiles are negligible suggesting a successful simulation of valid responses and validating the theoretical results.

Furthermore in Table 6.2, we find comparison of the empirical mean and variance with their corresponding theoretical values using the theoretically derived results in (5.48) and (5.49), respectively. The differences between empirical mean value and theoretical expected value as well as between variance values are negligible. These

Table 6.2: Empirical and theoretical expected value and variance of $\mathcal{T}_i$ for valid responses under the theoretical valid response model

	Empirical	Theoretical
Expected value	0.01	0.00
Variance	37.88	37.78

results are inevitable, once we have established (Figure 6.2) that the two methods of estimating the cumulative distribution function produce essentially identical results.

In summary, I was able to confirm that the theoretical properties of $\mathcal{T}_i$'s distribution could be confirmed with the empirical results and are graphically accessible. Furthermore, I was able to verify the functionality of he computational implementation of Imhof's cumulative distribution function. We have seen that via an optimisation procedure we can derive theoretical percentiles using this implementation, as well.

Results for Components of $\mathcal{T}_i$

Other important information sources are the four components of $\mathcal{T}_i$ as grouped in (5.15): sum of unweighted factor scores squared (*raw factor penalty*, rFP), sum of weighted factor scores squared (*weighted factor penalty*, wFP), sum of standardised observed-variable scores squared (*raw z-score penalty*, rZP), and sum of weighted standardised observed-variable scores squared (*weighted z-score penalty*, wZP):

$$\begin{aligned} \mathcal{T}_i = {} & \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^* + \sum_{k=1}^{q} y_{ik}^{*2} \sum_{j=1}^{p} \frac{\lambda_{jk}^2}{\psi_j} - \sum_{j=1}^{p} z_{ij}^2 \frac{\sigma_{jj}}{\psi_j} + \sum_{j=1}^{p} z_{ij}^2 \\ = {} & \quad \text{rFP} \quad + \quad \text{wFP} \quad + \quad \text{wZP} \quad + \quad \text{rZP} \; , \end{aligned}$$

and tFP = (rFP + wFP) and tZP = (wZP + rZP). By further summarising penalty components, the sum of raw and weighted factor penalties can find interpretation as *total factor penalty* (tFP) and, respectively, we can contrast this to the *total z-score penalty* (tZP).

Additionally, I will take a brief look at the behaviour of components as discussed in Section 5.1.2. The equation below reiterates this alternative interpretation in (5.20). In this form, the raw factor and z-score penalties are separate components, as well. The term in the middle was introduced such that it can be interpreted as weighted differences squared residuals (*weighted residual penalty*, wRP):

$$\begin{aligned}\mathcal{T}_i = &\quad \boldsymbol{y}_i^{*T}\Phi^{-1}\boldsymbol{y}_i^{*} - \sum_{j=1}^{p}\frac{\delta_{ij}^2-\delta_{ij}^{*2}}{\psi_j} + \sum_{j=1}^{p} z_{ij}^2 \\ = &\quad \text{rFP} \quad + \quad \text{wRP} \quad + \quad \text{rZP}\,.\end{aligned}$$

Figure 6.3 shows kernel density estimates for all components of $\mathcal{T}_i$, rZP, rFP, wZP, and wFP, separately. The components are calculated for the generated valid response sample under the parameters of the theoretical valid response model. We can easily identify the two density curves for the raw factor and z-score penalties, where both follow a χ^2-distribution with, $k = 3$ and $p = 18$ degrees of freedom, respectively. On the positive line, we additional have the density curve for the weighted factor scores. The only component with negative values is the weighted z-score penalty.

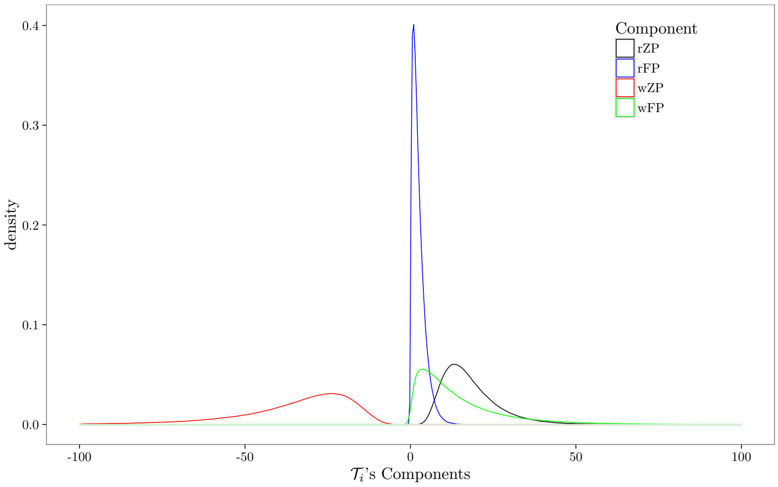


Figure 6.3: Kernel density estimates for all components of $\mathcal{T}_i$ for valid responses under the theoretical valid response model defined by the valid class of the latent class analysis estimated for the online questionnaire study sample.

The graphical interpretation is more intuitive when penalties are summarised based on observed variables, tZP, and penalties based on latent variables, tFP. In the next section I will compare density curves for the valid and invalid response groups for the summarised components. We can preview Figure 6.5 and focus only on the dashed curves, which shows the equivalent of previous Figure 6.3 but for the summarised components. Here, we can see that the penalty for aberrant response patterns on the negative line counteracts the penalty for aberrant factor scores on the positive line.

6.1.2 Valid and invalid Responses under the valid Response Model

In the previous section, I only looked at the valid response group given theoretical valid response model parameters are known. In this section, I will stay in the same scenario but compare the valid response group to results of the invalid response groups.

Results for $\mathcal{T}_i$

Once again, we will take a look at $\mathcal{T}_i$ in order to compare valid response group results to those of invalid response groups. This includes the response strategies previously labelled as item wording and long string. The response strategy item wording was introduced in Chapter 4 and the response model estimates for the (latent) invalid class will serve as theoretical distribution for this numerical exercise. Similar to the previous section, I will additionally compare results of the components of $\mathcal{T}_i$ for valid and item wording response groups.

Figure 6.4 shows the same densities for valid responses in Figure 6.1. Here, those are indicated with dashed/dotted curves. Additionally, we can see the distribution for the two invalid response groups. A simple long string response strategy was simulated were each i had a consistent answer option (1, 2, 3, 4, or 5) throughout all observed variables. $\mathcal{T}_i$ for all response patterns, including those derived from invalid response strategies, were estimated with the valid response model parameters. We can see that in this scenario, there is a clear difference of density curves for the valid response group and the two invalid response groups.

Table 6.3 summarises the percentages of invalid responders that we would identify if we used the theoretical 5^{th} and 95^{th} percentiles of $\mathcal{T}_i$'s theoretical distribution for valid responses. Cut-off values are also visually indicated by the two vertical dashed/dotted lines in Figure 6.4. With these cut-off points, we would be able to

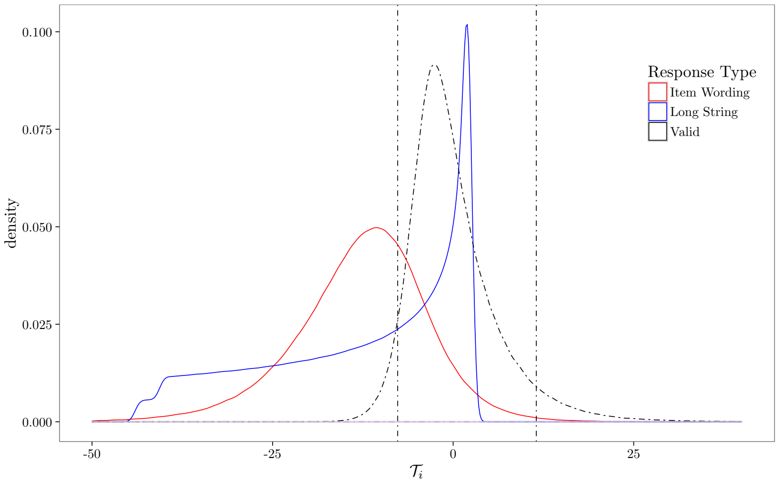


Figure 6.4: Kernel density estimates of $\mathcal{T}_i$ for valid responses (dashed/dotted curve) and two types of invalid responses (item wording and long string) when theoretical model for valid responders is known.

successfully identify 71% and 53% of invalid responses in the two invalid groups if we had knowledge of the theoretical parameters for the valid response model.

Table 6.3: Percentage of responders identified as extreme values using the theoretical percentiles as cut-off values for each response group separately

	Invalid responses		Valid responses
	Item wording	Long string	
5^{th} percentile	.71	.53	.05
95^{th} percentile	.00	.00	.05
Two-sided test	.71	.53	.10

In conclusion, the majority of these types of invalid responses have highly aberrant response patterns. These outliers are not matched by, from the valid response model assigned, extreme latent variable scores, leading to predominantly negative values.

Results for Components of $\mathcal{T}_i$

When comparing the results for valid responses with those for invalid responses, it is helpful to identify the driving components responsible for the overall differences in $\mathcal{T}_i$. First, we will take a look at differences in $\boldsymbol{y}^*$, which plays the central role in factor penalty components of $\mathcal{T}_i$. I will focus on the item wording response strategy in the following detailed discussion. We are particularly interested in differences between valid and invalid response groups when the invalid response strategy produces response patterns that are of a more subtle (semi-plausible) nature and as such harder to detect (as opposed to a more easily identifiable long string response pattern). As summarised in Table 6.4, where we have expected zero means for factor scores in the valid response group, we find non-zero means for the invalid response group under the valid response model. Estimated factor scores for the invalid responses result in slightly smaller variances when estimation is based on the valid response class model.

Table 6.4: Empirical mean and variance for estimated factor scores based on the theoretical valid response model for valid and invalid (item wording) response groups

Factor	Mean		Variance	
score	valid	invalid	valid	invalid
y_1^*	0.00	0.09	0.82	0.40
y_2^*	0.00	0.18	0.91	0.38
y_3^*	0.00	-1.06	0.74	0.76

Table 6.5: Empirical mean and variances for $\mathcal{T}_i$ and its components for valid and invalid (item wording) response groups

Penalty			Mean		Variance	
			valid	invalid	valid	invalid
raw	z-score	rZP	18.00	33.61	73.83	166.86
weighted	z-score	wZP	-35.32	-60.48	371.92	484.86
raw	Factor score	rFP	2.45	2.86	4.03	5.88
weighted	Factor score	wFP	14.88	11.31	210.95	74.14
total	z-score	tZP	-17.33	-26.86	128.98	112.29
total	Factor score	tFP	17.32	14.17	266.38	116.98
weighted	Residual	wRP	-20.45	-49.17	49.82	312.69
	$\mathcal{T}_i$		-0.00	-12.69	37.72	85.97

From Table 6.5 we can deduce that the aberrant mean values for the invalid response group do not lead to large differences in mean values of neither raw (rFP) nor weighted factor score penalties (wFP). Where the variance for the raw factor score penalty is slightly increased, the variance for the weighted factor score penalty is much smaller in comparison to the corresponding results for the valid response group. The latter is due to the assigned weights, which depend on the measurement model reliability of the valid response model. Overall this effect does not seem very large. On the other side, we can see clear differences between z-score penalties of valid and invalid responses. Absolute mean values of the z-score penalties are much larger for the invalid response group, as is the case for their variances. Lastly, on the alternative interpretation of $\mathcal{T}_i$, we can see that absolute weighted residual penalty is much larger for invalid response group. This applies even more so for its variance, which experiences a dramatic increase for the invalid response group in comparison to the valid response group.

These results can be observed in Figure 6.5, where the density curves for total z-score, located on the negative line, and total factor score penalties for the invalid response group are drawn. The dashed/dotted curves indicate the corresponding density functions for the valid response group. These two have their modes closer to 0 than is the case for the invalid response group. Where the curves on the positive line are visually not very different there is an apparent shift to the left of the opposite curve for invalid responses. Hence, we see an incongruence of extreme values in the observed variables not justified by extreme positions on the latent dimensions for invalid response patterns.

A similar conclusion can be drawn when we investigate correlation patterns between components of $\mathcal{T}_i$, as summarised in Table 6.6, and for the alternative representation of components, reported in Table 6.7. We can roughly summarise that the correlation between the components, and therefore between components and $\mathcal{T}_i$, are in general stronger in the valid response group than they are in the invalid response group. As is to be expected, correlations are strong between raw and weighted penalties of the same kind in both response groups. Raw and weighted penalties are a linear combination of the same values. If the item (j) specific weights (here, interpretable as types of reliability coefficients) do not vary between items, we could approximate the raw penalty as a linear function of the weighted penalties. This seems to be particularly the case for raw and weighted z-score penalties. Even in

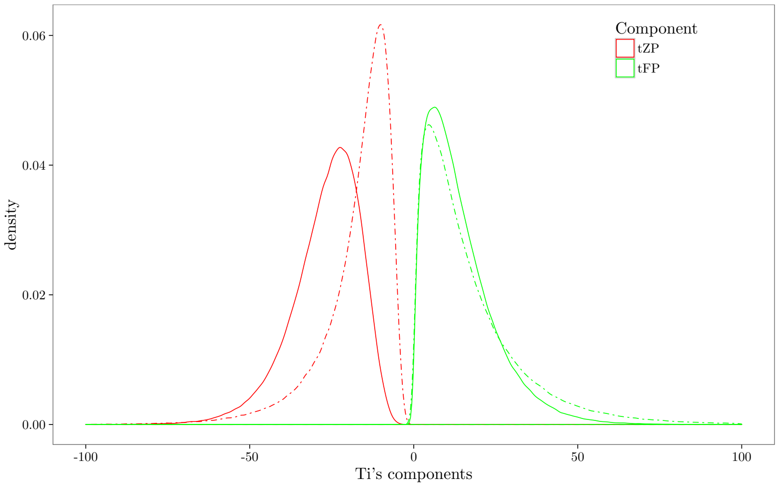


Figure 6.5: Kernel density estimates for components of $\mathcal{T}_i$ under the alternative representation for invalid responses under the theoretical valid response model defined by the valid class of the latent class analysis estimated for the online questionnaire study sample.

the invalid response group, we can observe a large correlation. Hence, it seems that the information that we can draw from observed scores is not significantly modified by taking the model into account. However, as we have seen in Figure 6.5 and based on the variance estimates in Table 6.5, the differences in actual magnitudes (sum of components of the same penalty source) as captured in the total z-score (tZP) and factor score penalties (tFP) do provide essential information. It seems that contrasting penalties of the same penalty source extracts information that would otherwise be obscured by the large amount of shared information (strong correlations).

I shall now focus on correlations between factor score and z-score penalties as these differ most between valid and invalid response groups. For instance, where raw factor score penalties are closely related to z-score penalties in the valid response group, with a correlation of .90, we experience a decrease for said correlation in the invalid response group, with correlation coefficient around .66. The same conclusion applies to the absolute correlation between total factor score and total z-score penalties as reported in Table 6.7. The weighted residual penalty includes fitted and observed variable penalties and, as such, does not give as clear of a picture. However,

Table 6.6: Correlations between components of $\mathcal{T}_i$ within valid response and invalid response groups

	Groups									
	valid					invalid				
	rZP	wZP	rFP	wFP	$\mathcal{T}_i$	rZP	wZP	rFP	wFP	$\mathcal{T}_i$
rZP	1					1				
wZP	-.96	1				-.95	1			
rFP	.90	-.89	1			.66	-.64	1		
wFP	.86	-.95	.88	1		.59	-.65	.88	1	
$\mathcal{T}_i$	.72	-.74	.86	.86	1	-.14	.28	.48	.43	1

Table 6.7: Correlations between components of $\mathcal{T}_i$ within valid response and invalid response groups

			Groups					
			valid			invalid		
	Penalty		tZP	tFP	wRP	tZP	tFP	wRP
total	z-Score	tZP	1			1		
total	Factor Score	tFP	-.96	1		-.63	1	
weighted	Residual	wRP	-.85	-.63	1	-.90	-.37	1

the correlation between the weighted residual and the total factor score penalties decreases in comparison to the valid response group.

The decrease in correlation is further observable in a scatter plot as given in Figure 6.6. The total z-score penalty values are plotted against the total factor score penalty values. Furthermore, a colour scheme has been applied to the data points such that the more dense areas in the scatter plot are indicated via red data points. This is following a heat map colour scheme, where blue areas indicate a lower density of points around this area. The scatter plot on the left side shows the results for the valid response groups. The strong negative relationship is easily observable, and a regression line could be drawn with about a 60 degrees angle. A 45 degrees angle with the same scale on both axes would have indicated a more symmetric distribution of $\mathcal{T}_i$, as it is simply the sum of both penalties. On the right-hand side of Figure 6.6, we have the equivalent scatter plot for the invalid response group, which would not allow for a clear regression line. More negative values in total z-score penalty are associated with in magnitude less positive total factor score penalty values. This is in line with previous interpretation of mean and variances of $\mathcal{T}_i$'s penalty components.

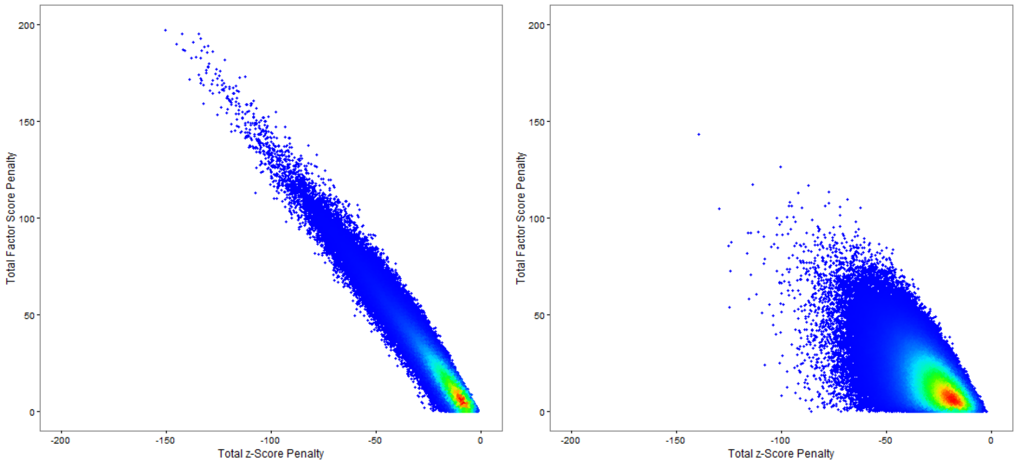


Figure 6.6: Scatterplot of total z-score and factor score penalties for valid (left) and invalid (right) response groups with coded point density following a heat map colour scheme.

We conclude that the penalties for z-score and factor score deviation are unmatched for invalid response patterns. This was shown by an overall decreasing relationship between components of $\mathcal{T}_i$. In particular, we saw that the z-score penalties were most affected by aberrant response patterns of the item wording kind.

6.1.3 Valid and invalid Responses under the biased valid Response Model

So far, I have investigated $\mathcal{T}_i$'s behaviour under the theoretical valid response model, contrasting the valid response to the invalid response groups. These revealed high discriminatory potential of $\mathcal{T}_i$ when the valid response model is known. In this section, I seek to investigate the discriminatory power of $\mathcal{T}_i$ when a biased estimate of valid response model is used instead. For this purpose, I use the estimates for the valid response model based on the latent variable analysis model for the Johnson (2005) study data without accounting for invalid response strategies. These are biased because the invalid responses are also included in the estimation.

Figure 6.7 shows the empirical densities of $\mathcal{T}_i$ again for valid responses and invalid responses of type item wording. However this time, I use the biased estimates for computation of $\mathcal{T}_i$. As a reference, the dashed/dotted curve represents $\mathcal{T}_i$ for valid responses when the valid response model is known, as was shown in the previous graphs. The reference curve was included to show how little the two curves differ although one curve is based on biased estimates. We can see a slight shift of the density curve for valid responses when $\mathcal{T}_i$ is estimated with biased estimates for the valid response model. There is a much larger difference to be seen when the respective density curves for invalid responses are compared. The difference between the density curve in Figure 6.4 for invalid responses and the corresponding curve in Figure 6.7 is very apparent. The density curve for invalid responses has a sharper peak and $\mathcal{T}_i$ values are closer to those from valid responses when the biased estimates are used as information source. However, the density curve for invalid responses is still distinct enough given the estimates for the valid response model are not extremely biased by the (estimated) 10% of invalid responses in the analysis sample.

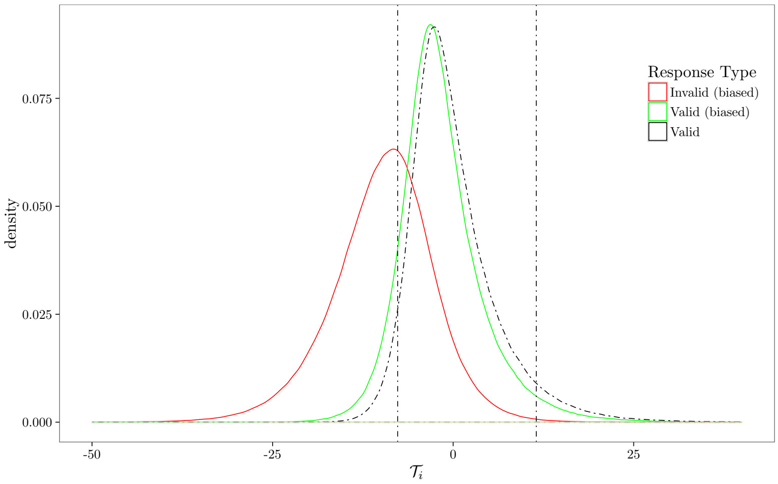


Figure 6.7: Kernel density estimates of $\mathcal{T}_i$ for valid responses and invalid responses (of type item wording) when we have biased estimates of the theoretical model for valid responses and a reference curve for valid responses when the theoretical model is known (dashed/dotted curve).

Table 6.8 summarises the percentages of invalid and valid responders that we would identify when we use the theoretical 5^{th} and 95^{th} percentiles of the theoretical distribution for valid responses. These are the cut-off values as calculated in Section 6.1.2, i.e. if the true covariance matrix was known or consistently estimated. However, $\mathcal{T}_i$ values are estimated based on biased parameter estimates of the valid response model. This is to give us an indication of estimation bias for the valid response model when we have invalid responses present that are not accounted for by the model. Where I still successfully identify about 76% of invalid responses, I incorrectly classify around 10% of valid responses as extreme values, resulting in a total of 20% incorrectly flagged valid responses. Where we would expect 5% flagged valid responses if the estimated model for valid responses was not biased, we see 15% and 5% as empirical values, suggesting that the contaminated sample slightly biased the parameter estimates, leading to a shift of the valid response distribution. It seems that $\mathcal{T}_i$ values estimated for valid responses using the biased estimates lead to an increased variance and a slight shift to the left of $\mathcal{T}_i$ values for valid responses.

Table 6.8: Percentage of responders identified as extreme values using the theoretical percentiles as cut-off values for the two response groups separately

	Groups	
	invalid	valid
5^{th} percentile	.76	.15
95^{th} percentile	.00	.05
Two-sided test	.76	.20

Lastly, Table 6.9 shows flagged valid and invalid responses based on cut-off values that are estimated when we use the biased estimates of the valid response model. This is in line with a real world scenario, in which we do not have full knowledge of the theoretical valid response model. Instead, parameters for this model are estimated based on a sample with invalid responses included but not taken into account by the model. Instead of previously defined tolerance rule of 10% valid responses classified as extreme, we would incorrectly classify about 14% of valid responses when we do not have full knowledge of the theoretical valid response model. However, we would still correctly identify about 61% of invalid responses in a real world scenario.

Ultimately, we loose discriminatory power when we estimate valid response model parameters when invalid response patterns are present in the sample. Furthermore,

Table 6.9: Percentage of responders identified as extreme values using estimated percentiles as cut-off values for the two response groups separately

	Groups	
	invalid	valid
5^{th} percentile	.60	.11
95^{th} percentile	.00	.03
Two-sided test	.61	.14

we increase the risk of incorrectly classifying valid response patterns as extreme values. It seems reasonable to weigh cost and benefit when we use $\mathcal{T}_i$, based on the individual setting. In case the valid response model has low measurement accuracy and/or we expect severe bias of valid response model parameter estimates, it is sensible to limit the identification procedure to a left-sided detection rule, i.e. only flag $\mathcal{T}_i$ values smaller than the cut-off value on the left side. Especially, we can approach the detection such that we use a more conservative, e.g. 1^{st} percentile, cut-off criterion to minimise the risk of incorrect positive classifications.

6.2 LCA Parameters as Information Source for the new Measure

In the previous sections, we saw that the discriminatory power of $\mathcal{T}_i$ heavily depends on the magnitude of estimation bias and valid response model accuracy. In chapter Chapter 4, I used latent class analysis to improve the accuracy of parameter estimates for the valid class model. Furthermore, I was able to decrease the error variances within the valid class model by incorporating an invalid response strategy into the model. In this section, I seek to combine the benefits a latent class analysis can provide in acquiring more accurate valid response model parameter estimates and the discriminatory potential of using $\mathcal{T}_i$ as detection measure for invalid response patterns, when it is estimated by drawing on the valid class model parameter (as more accurate estimates of the valid response model).

The combined approach towards detection will be evaluated on the experimental Huang et al. (2012) study sample. However, the valid response model parameters required for the computation of $\mathcal{T}_i$ will differ between evaluation scenarios. Percentiles of $\mathcal{T}_i$ will be estimated from several different estimates of the valid response model. Those will serve as cut-off values for the response patterns in the experimental study sample. Based on experimental sub-sample membership, I will investigate the success of correctly and incorrectly classified extreme values of $\mathcal{T}_i$.

First, I will compare the discriminatory power of $\mathcal{T}_i$ when estimated with latent variable analysis parameter estimates (1^{st} evaluation scenario, *Experimental Study LVA*) and when estimated from valid class parameter estimates based on latent class analysis (2^{nd} evaluation scenario, *Experimental Study LCA*). Furthermore, I will also use the corresponding parameter estimates obtained from the online questionnaire sample from the Johnson (2005) study (3^{rd} evaluation scenario, *Online Questionnaire LVA*) and then apply it as information source for $\mathcal{T}_i$. Both studies include the same personality assessment questions. Hence, if we are willing to assume measurement invariance between valid response groups of both study samples, these results might represent more accurate estimates for the valid response model because I suspect a smaller percentage of SpRPs in the online sample. The online questionnaire sample was also analysed with the latent class analysis design and represents the last evaluation scenario (4^{th} evaluation scenario, *Online Questionnaire LCA*).

In Table 6.10, the percentage of sub-sample members identified as extreme negative, left-sided cut-off (L), and extreme positive, right-sided cut-off (R), values in $\mathcal{T}_i$ are summarised throughout all evaluation scenarios. The total percentage of extreme negative and extreme positive values as a two-sided test (T) are reported as well. Using the respective parameters in each of the four evaluation scenarios, the 5^{th} and 95^{th} percentiles for valid responses are estimated and serve as cut-off values. First, we take a look at the evaluation scenario where experimental data analysis parameters from the latent variable model assuming no invalid responses serve as the information source for the computation of $\mathcal{T}_i$. In previous Chapter 5, we looked at the discriminatory power of $\mathcal{T}_i$ and set cut-off values from a perspective of a privileged information scenario in which experimental sub-sample membership is known. We can see that even without sub-sample membership information and based on estimated percentiles of $\mathcal{T}_i$ using the biased parameter estimates, we can achieve medium discriminatory power between invalid and valid responses. A two-sided detection rule correctly detects 17% semi-plausible responders from experimental

Table 6.10: Percentages of respondents in experimental study sub-samples flagged as extreme values with cut-off defined by estimated 5^th and 95^th percentiles of $\mathcal{T}_i$ based on different information sources as estimates for the valid response model

Sub-sample	Cell(s)	Source of parameter estimates											
		Experimental study						Online questionnaire					
		LVA			LCA			LVA			LCA		
		L	R	T	L	R	T	L	R	T	L	R	T
Plausible	1	.08	.08	.15	.18	.03	.21	.13	.03	.15	.21	.03	.23
	4	.02	.08	.11	.18	.10	.27	.11	.04	.14	.11	.02	.13
Semi-plausible	2	.09	.05	.14	.32	.12	.44	.26	.04	.30	.39	.02	.40
	5	.12	.08	.20	.23	.22	.45	.23	.08	.31	.27	.06	.33
Implausible	3	.25	.04	.29	.45	.22	.67	.42	.02	.44	.49	.02	.51
	6	.16	.04	.20	.28	.21	.49	.38	.00	.38	.43	.00	.43
Plausible	1,4	.04	.08	.12	.18	.07	.25	.11	.03	.15	.14	.02	.16
Semi-plausible	2,5	.11	.07	.17	.27	.17	.45	.25	.06	.31	.32	.04	.36
Implausible	3,6	.20	.04	.24	.35	.21	.57	.40	.01	.40	.46	.01	.46
Semi-/implaus.	2,3,5,6	.16	.05	.21	.32	.19	.51	.33	.03	.36	.39	.02	.42

^LVA, LCA Results for the latent variable analysis (LVA) and the latent class analysis (LCA).
^L,R,T L = Cut on left side, R = right side, T = two-sided test.

cells 2 and 5, and 24% implausible responders from cells 3 and 6. The risk of identifying plausible responders from cells 1 and 4 is kept just slightly above the tolerance level of 10% percent at 12%. Secondly, we can compare these results with the flagged percentages that result when we use the valid response latent class model as the information source for the estimation of $\mathcal{T}_i$. It is apparent that there is a significant increase in detection success of semi-plausible responders with a 45% and implausible responders with a 57% detection rate. However, the risk of incorrectly identifying valid responders reaches intolerable levels with 25% of them being categorised as extreme values in $\mathcal{T}_i$. Thirdly, I shall investigate the results when the online questionnaire analysis estimates for the valid response model assuming no invalid responders are used as the information source for $\mathcal{T}_i$. Although in this scenario I detected 40% of implausible responders and 31% of semi-plausible, I incorrectly classified 15% valid responders as extreme. Using the parameter estimates based on another sample is only a sensible approach if we can assume measurement equivalence between the valid response models of both samples. Lastly, I contrasted the former results of the online questionnaire parameter estimates with the outcomes that result when we draw on the valid latent class model analysis parameters which were estimated with the online questionnaire sample, as well. Within this contrast the combined approach leads to an improve of discriminatory power. Here, I slightly increased the risk of flagged valid responders to 16%, but gain in detection rate of semi-plausible responders, which is 36%, and of implausible responders, which is 46%, in comparison to the corresponding online questionnaire evaluation scenario.

Theses results show that a combined approach is a promising method to detect SpRPs when parameter estimates are not analysed with a sample of predominantly invalid responders. The experimental study sample consists of 68% semi-/implausible responders. However, the risk of incorrectly identifying valid responders is worrisome. Therefore, it seems sensible to investigate a more conservative approach to the selection of cut-off points.

Table 6.11: Percentages of respondents in experimental study sub-samples flagged as extreme values with cut-off defined by estimated 1^{st} and 99^{th} percentiles of $\mathcal{T}_i$ based on different information sources as estimates for the valid response model

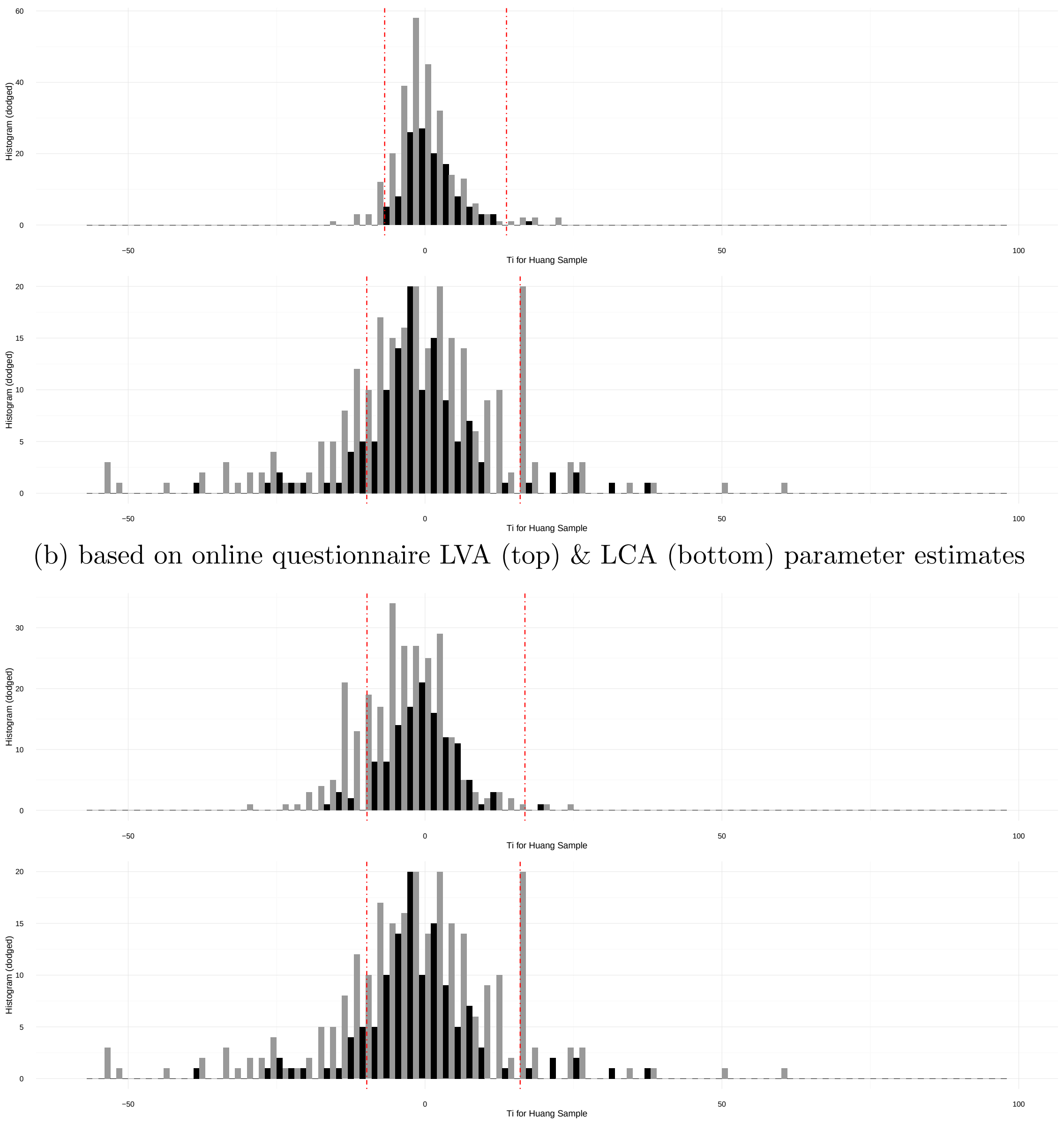
Sub-sample	Cell(s)	Source of parameter estimates											
		Experimental study						Online questionnaire					
		LVA			LCA			LVA			LCA		
		L	R	T	L	R	T	L	R	T	L	R	T
Plausible	1	.00	.00	.00	.15	.03	.18	.03	.00	.03	.05	.00	.05
	4	.00	.01	.01	.11	.07	.18	.06	.01	.07	.10	.01	.11
Semi-plausible	2	.02	.00	.02	.14	.05	.19	.19	.00	.19	.26	.00	.26
	5	.05	.05	.09	.19	.12	.31	.14	.03	.17	.14	.02	.16
Implausible	3	.16	.04	.20	.33	.11	.44	.31	.02	.33	.35	.02	.36
	6	.11	.01	.12	.25	.19	.43	.28	.00	.28	.33	.00	.33
Plausible	1,4	.00	.01	.01	.12	.06	.18	.05	.01	.06	.08	.01	.09
Semi-plausible	2,5	.03	.02	.06	.17	.09	.26	.17	.02	.18	.20	.01	.21
Implausible	3,6	.13	.02	.15	.28	.15	.43	.29	.01	.30	.34	.01	.35
Semi-/implaus.	2,3,5,6	.09	.02	.11	.23	.12	.35	.23	.01	.25	.27	.01	.28

LVA, LCA Results for the latent variable analysis (LVA) and the latent class analysis (LCA).
L,R,T L = Cut on left side, R = right side, T = two-sided test.

In Table 6.11 we see the equivalent results of the previous table, but now with estimated 1^{st} and 99^{th} percentiles of each corresponding evaluation scenario serving as cut-off points for the identification of extreme $\mathcal{T}_i$ values. First, we look at the at the outcomes when experimental data analysis parameters from the latent variable model assuming no invalid responses serve as the information source for the estimation of $\mathcal{T}_i$. The table shows that the more conservative cut-off criteria decreased the overall amount of responders whose $\mathcal{T}_i$ values were identified as extreme throughout all experimental sub-groups. However, although we would only identify 6% of semi-plausible and 15% of implausible responders, I also decreased the risk of valid responders flagged as extreme to 1%. Secondly, I compare these results to the outcome of the corresponding combined approach towards detection. Once again, the combined approach in this evaluation scenario detects more semi-/implausible responders, however, with an intolerable risk of flagging valid responders. This risk can be reduced to a just about tolerable level of 12% flagged valid responders when only a left-sided cut-off is employed. In doing so, I largely increased the percentage of identified semi- and implausible responders by 24% to 35% as well, in comparison to the previous evaluation scenario. Thirdly, I use the analysis results of the online questionnaire sample as the information source for $\mathcal{T}_i$. This seems to be very successful in contrast to the corresponding evaluation scenario with liberal cut-off criteria. Here, results are much better than the liberal cut-off criteria in the original evaluation scenario Experimental Study LVA. The risk for valid responders to be flagged is at 6% where 18% of semi-plausible and 30% of implausible responders were identified as such. Lastly, we can compare these results with those acquired when the combined approach is applied. With a tolerable risk of 9% incorrectly identified valid responders, I successfully detect 21% of semi-plausible and 35% implausible responders. Hence, this evaluation scenario is the most successful in detecting SpRPs while keeping the risk for flagged valid responders below the 10% tolerance level.

The conservative approach with the 1^{st} percentile as left-sided only cut-off has proven to be the method of choice when analysed sample consists of primarily invalid response patterns and the valid response model incorporates high levels of measurement error. The combined approach increased the discriminatory power successfully when the online questionnaire sample analysis parameter estimates were used as the information source. Where liberal cut-off criteria did not work for the combined approach, the conservative cut-off criteria has shown to be the most successful in discriminating between valid and invalid responses.

Figure 6.8 provides a visual summary of above results in the form of stacked


(a) based on experimental study LVA (top) & LCA (bottom) parameter estimates

(b) based on online questionnaire LVA (top) & LCA (bottom) parameter estimates

Figure 6.8: Histograms for values of $\mathcal{T}_i$ presented in dodged form for the experimental study sub-samples semi-/implausible (in grey) and valid (in black) with corresponding estimates of 1^{st} and 99^{th} percentiles indicated as dashed/dotted vertical lines for different evaluation scenarios.

histograms of the $\mathcal{T}_i$ values for the experimental study sample. Each histogram differentiates between semi-/implausible responders (in grey) and plausible responders (in black). Additionally, the cut-off points based on the corresponding 1^{st} and 99^{th} percentiles are drawn as dotted red vertical lines. The first two histograms represent the evaluation scenarios in which the experimental study sample was used to estimate valid model parameters. The last two histograms are for evaluation scenarios in which the online questionnaire sample analysis estimates for the parameters served as the information source. The second and last histograms represent $\mathcal{T}_i$ based on the combined approach towards detection for the respective study samples. The combined approaches seem to increase the variance of $\mathcal{T}_i$, assigning more extreme values to both valid and invalid responders. In comparison to the theoretical shape of $\mathcal{T}_i$'s distribution for valid responses, we see that more valid responders were assigned more extreme values than would be expected. One explanation might be that accommodating only one form of invalid response strategy into the latent class model still biases the valid response model parameter estimation. Another invalid response strategy might have a larger influence on the estimation of the valid response model class parameter estimates. Lastly, it is easily observable that more liberal cut-off criteria, i.e. 5^{th} and 95^{th} percentiles, dramatically increase the risk of incorrectly identifying plausible responders.

In conclusion, when we would like to use a simple approach towards detection without employing latent class models a 5% cut-off criterion is effective even if the sample consists of predominantly SpRPs. A combined approach does not seem to improve discriminatory power in such a case because the risk of identifying valid responses increases dramatically. If measurement equivalence is given, then it seems sensible to use an uncontaminated sample for the estimation of the valid response model as the information source for the estimation of $\mathcal{T}_i$. In Section 5.2, I argued that this assumption might not be met in our case. However, using the online questionnaire parameters instead did lead to largely increased discriminatory power with a slightly enlarged risk of flagging invalid responses. In general, when we use parameters estimates from another sample for the identification of SpRPs, we should be cautious and use a conservative cut-off criterion, i.e. 1^{st} percentile only, to keep the risk of incorrectly identified valid responders small. I conclude that a combined approach does improve discriminatory power when the parameters are not estimated with a sample that consists of too many SpRPs.

Chapter 7

A Simulation Study

The previous chapters dealt with relevant literature for the development of an identification measure for SpRPs. Furthermore, two empirical datasets were discussed to identify properties of SpRPs under latent variable models. Lastly, a new identification measure was developed and its discriminatory power in identifying invalid responders for the empirical data analysed. In this chapter, I seek to further examine the statistical properties of the new measure and evaluate its discriminatory power using simulated valid and invalid responses.

In the following sections, I will define several valid response models and two invalid response strategies with the corresponding invalid response models. Responses will be simulated for each of these conditions. We will be able to define the percentile values for extreme values that can serve as cut-off values for the identification of invalid responses. The same procedure will be implemented simulating a real world scenario where the theoretical valid response population model is unknown. Hence, the valid response model will be estimated from samples that include invalid responses. The estimated valid response model then serves as bases for the estimation of cut-off values. Ultimately, results will reveal situations of large and small discriminatory power and implications for estimation bias.

7.1 Simulation Conditions

In this section, theoretical population models for the simulation of valid responses and the simulation of invalid responses will be defined. Furthermore, different sample scenarios for the evaluation of the extent of estimation bias and the discriminatory power of $\mathcal{T}_i$ will be defined.

7.1.1 Theoretical Models

The valid responders' models are all factor analysis models as defined in Chapter 3 for the empirical data. We have multivariate normal observed variables $\boldsymbol{x}$ following a factor analysis model, $\boldsymbol{x} = \boldsymbol{\mu} + \Lambda\boldsymbol{y} + \boldsymbol{\epsilon}$, with q factors $\boldsymbol{y}$ and Λ as factor loading matrix of factor loadings $\lambda_{j,k}$ with rank q. Λ is constraint according to a simple factor structure, such that $\Lambda^T\Psi^{-1}\Lambda$ is a diagonal matrix. Notationally, I use λ_j without the subscript k, if I refer to the single non-zero element in row j of the Λ matrix. I define latent variables such that $\boldsymbol{y} \sim N(0, \Phi)$. Φ is the factor covariance matrix with factor variances, $\phi_{k,k}$, as diagonal elements and factor covariances, $\phi_{k,m}$ with $k \neq m$, as non-diagonal elements. $\Sigma = \Lambda\Phi\Lambda^T + \Psi$ defines the observed variable covariance matrix, where Ψ is a diagonal matrix containing the error variances ψ_j of independent error terms $\boldsymbol{\epsilon}$.

These latent variable models are defined by 10 settings to which I will refer to as simulation factors in the following. We can see a summary of these valid model defining information in Table 7.1. For instance, we can find the number of specified latent variables (first row) and what values I assigned to their means and variances (rows four and five) in Table 7.1. 6 of these 10 simulation factors are alternated between different valid response model specifications. Hence, we have 6 experimental simulation factors with either 2 or 3 levels (alternated settings). Based on these experimental simulation factor levels, we have 324 different combinations of valid response latent variable population models. Correspondingly, I will define two separate invalid response strategies, where the first follows the example set by the invalid responders' class in Chapter 4 and the second is set to produce long string response patterns mentioned in Section 2.1.1. The invalid response model is implied by aspects of the respective valid response model simulation condition.

Valid Response Models

Table 7.1 shows the valid model specifications with some of the simulation factors experimentally manipulated. Hence, I restrict the generalisability of this study to models with only first-order latent variables and a simple factor loading structure (each of the observed variables only load on a single latent variable). Furthermore, I generate observed variables such that they are standard normal and drawn from a multivariate normal distribution as implied by the respective population covariance matrix for valid responses. First, I vary the number of latent variables in the model.

Table 7.1: Specifications of the valid response population models

Setting	Levels			Notation
Number of factors	4		8	defines q
Percentage of neg. correlated LV	0	.25	.50	implies no. of $\phi_{k,m} \leq 0$
Absolute inter-LV correlation	0	.25	.50	defines $\lvert\phi_{k,m}\rvert\ \forall j \neq m$
Factor mean		0		defines $\nu_k = 0\ \forall k$
Factor variance		1		defines $\phi_{k,k} = 1\ \forall k$
Number of indicators per LV	4		8	implies p
Percentage of rev. ind. per LV	0	.25	.50	implies no. of $\lambda_j < 0$
Indicator variances		1		defines $\sigma_{j,j} = 1\ \forall j$
Indicator mean		0		defines $\mu_j = 0\ \forall j$
Percentage error var. of ind.	.75	.50	.25	defines ψ_j and implies $\lambda_j\ \forall j$

$^{\text{Indicator}}$ Observed variables.
$^{\text{LV, Factor}}$ Latent variable defined by set of indicators.

I define two levels with either $q = 4$ or $q = 8$ latent variables. The minimum amount of 4 latent variables are chosen to allow for variations of further aspects. For instance, we have three levels for the simulation factor percentage of negatively correlated latent variables. In general, I set latent variables to be correlated positively. However, we can induce different correlation patterns among latent variables. This is similar to the valid response models in previous chapters, where the latent variable Emotional Stability (N) was negatively correlated with the other Big Five personality factors. In this case, we had $\frac{1}{3}$ of the latent variables of opposite directional dependence to the other latent variables. In a similar manner, I alternated the percentage of latent variables that are of opposite directional to the remaining majority of latent variables.

For instance, in the case of 4 latent variables, we either have 0 (0%), one (25%), or two (50%) latent variables that are of negative kind. Based on the number of latent variables that we could refer to as of positive and negative kind, we then have 0, 3, or 4 negative correlations out of 6 correlations. This is because the latent variables of equal directional kind are positively correlated with each other (except in the simulation settings where I define the latent variables to be independent). A further experimentally manipulated simulation factor is the absolute correlation between latent variables. Latent variables are either independent, weakly, or moderately high correlated. The maximum absolute correlation of $|\phi_{k,m}| = .5$ was set to provide an at least minimally distinct latent variable structure. The simulation factor number of indicators per latent variable has two levels. The minimum amount of 4 indicators per latent variable was chosen to allow for parsimonious variations in the percentage of reversed indicators per latent variable. Hence, in the case of 4 indicators per latent variable, we have 0, 1, or 2 indicators with negative factor loadings. The number of indicators per latent variable also defines the total number p of observed variables because we have a simple factor loading structure. Lastly, I set two levels for the percentage of residual variance versus explained variance of observed variables. Based on a meta-analysis we can expect about in average 30% error variance when seeking to measure constructs like attitudes, personality, or job performance, in disciplines like marketing, psychology, sociology, or education (Cote and Buckley, 1987). Originally, a maximum of $\psi_j = .50 \times \sigma_{j,j}$ error variance for all j in the valid response model was chosen based on the notion that highly unreliable measurement will not serve as information for the identification of invalid responders. However, a large error variance condition of 75% was, yet, added to the simulation as an extreme setting in order to investigate possible trends in the behaviour of the new identification measure under widely varying error variances. Given the percentage of error variance, standard normal observed variables, and a simple factor structure, the absolute factor loadings are defined $|\lambda_{j,k}| = \sqrt{\sigma_{j,j} - \psi_j} = \sqrt{1 - \psi_j}$ for all j and for each j one corresponding latent variable k. The remaining factor loadings are set to 0 following a simple factor structure.

For easier access of the general structure for the valid response model, Figure 7.1 shows the path diagram with parameter values based on one example condition (on the left). Some parts in the path diagram are printed in red to accentuate

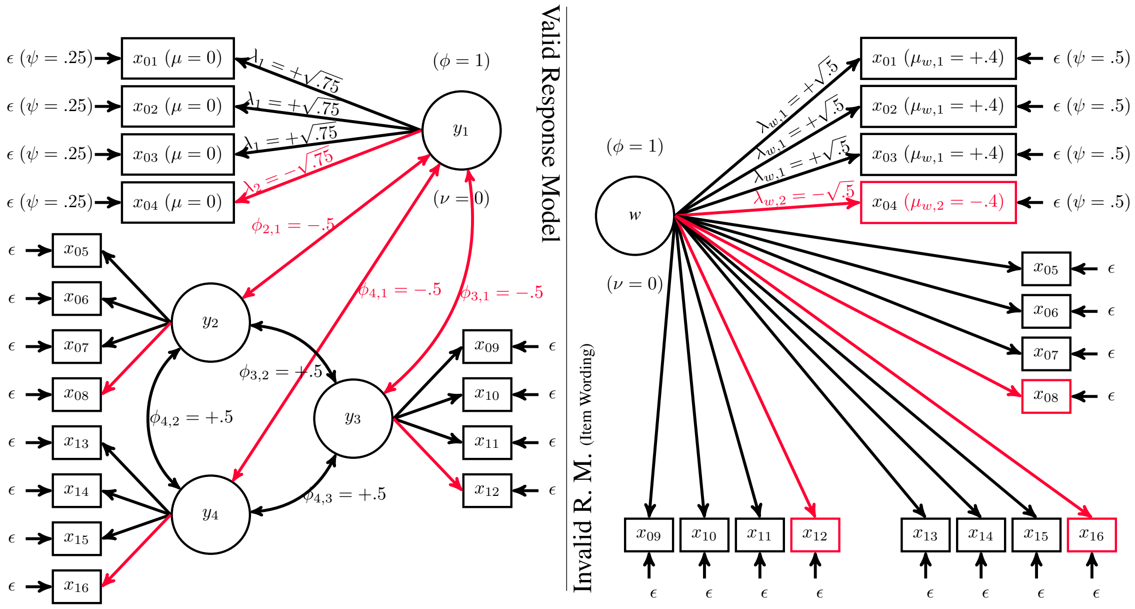


Figure 7.1: Path diagram for the valid response model with 4 latent variables, 25% negatively correlated latent variables (= 1 latent variable), an inter-factor correlation of .5, 4 indicators per latent variable, 25% reversed indicators per latent variable, and on average 25% residual variance.

differences for the otherwise repeating patterns of model definitions. For instance in this example, we have 25% of reversed items and, hence, one factor loading (λ_2) of opposite sign for each of the four measurement models. Furthermore, we have negative covariances between the first factor and the others factors ($\phi_{2,1}$, $\phi_{3,1}$, and $\phi_{4,1}$).

Invalid Responses Model: Item Wording

The first invalid responses model is in line with the semi-plausible response strategy introduced in Section 4.2; namely, the tendency to favour an idiosyncratic positive or negative answer category based on question wording. I define extra latent variables that are not part of the constructs of interest in the valid response model. This latent variable describes the individual tendency of an invalid responder to favour a specific range of answer options independent from actual item content (e.g., question intend in a survey). Therefore, I assume that invalid responses are not a function of the actual constructs of interest and no relation between constructs of interest and the individual tendency. All observed variables serve as indicators for a single latent variable. This latent variable is to capture the individual tendency to prefer a (range of) answer option(s). Based on negative versus positive question wording, the factor loadings switch sign. Observed variables that are defined as reversed items and as such have a negative factor loading in the valid responders' model also negatively load on the single latent variable in the invalid responders' model. When, with respect to the simulation condition, the percentage of reversed observed variables in the valid model is 0, I define only positive factor loadings. Reiterating the item wording response strategy, respondents choose an answer option that is independent of item content. However, instead of answering questions based on their inherent (self-assessment) position on the intended item scale, they chose an answer option which follows a positive response with respect to item wording. Thus, in this context, I refer to a positive answer option as a function of item wording rather than a function of the direction in which the question is asked. That is, the response is not caused by whether the answer options are given such that they start or end with 'strongly agree' versus 'strongly disagree'. Consequently, we have to reverse the sign of factor loadings where are positive response means to 'disagree' with the question statement. The results of the observed variable intercepts for the invalid class has shown that the means of observed variables are shifted by approximately .4 absolute standard deviation. Hence, I define the mean value for observed variables that are positively related to the respective latent variable as $\mu_{w,1} = .4$ and for reversed indicators $\mu_{w,2} = -.4$. The residual variance has been set to 50% of the total observed variable variance. This was chosen based on analysis results of the invalid responders' class' average percentage of residual variance. The number of observed variables for the invalid responders' model is defined by the number of observed variables in the respective simulation condition's valid response model.

For easier access to the general structure for the valid response model, Figure 7.1 shows on the right the path diagram with parameter values based on the respective example simulation condition presented on the left. The symbols printed in red accentuate that reversed items have factor loadings and mean values of opposite sign.

Invalid Responses Model: Long String

The second invalid response strategy is chosen to be a less sophisticated version of item wording invalid response strategy introduced previously; namely, the tendency to answer in a 'long string' pattern, consistently choosing a preferred range of answer options without regards to actual question content.

The model for this invalid response strategy has a similar set up as is the case for the first invalid response strategy. However, the latent variable defining the preferred answer option has a uniform distribution, such that $w \sim U(-3, 3)$. This is similar to a long string response pattern if we had chosen observed variables with discrete answer options. We can think of this scenario where a participant randomly draws a number from the uniform distribution of w before choosing from (observed) answer options. The observed answers are then a consistent function of the (latent) choice and some random variation (error). The range from -3 to 3 for latent variable was chosen as reference to valid responses: for the observed variables in the valid response model with $x_j \sim N(0, 1)$ and a sample size of $n = 1000$ approximately one valid respondent (based on x_j's distribution in the valid response model) would respond with a value more extreme than ±3. Correspondingly, I define $\mu_j = 0$, and error variance $\psi_j = .5$ and consequently $\lambda_j = .41$, for all j.

7.1.2 Samples

For each simulation condition I set a total sample size of $n = 1000$ with each 100 replications. To investigate the discriminatory power of $\mathcal{T}_i$, I draw a certain amount of response patterns from the valid response model and the remaining number of response patterns from the respective invalid response model. Hence, the percentage of valid responders is a further simulation condition factor that has 3 levels. A mixed valid and invalid responders sample can consist of either 90%, 70%, and 50% valid responders. Hence, we have a global count of simulation conditions of 972. In order to ensure the applicability in usual research conditions, the valid response model is estimated including the entire sample assuming that there are no invalid responses present.

7.2 Evaluation Scenarios

To evaluate the effectiveness of $\mathcal{T}_i$ in discriminating invalid responses and to understand the results, three different evaluation scenarios will be implemented.

Theoretical $\mathcal{T}_i$ Parameters and theoretical Percentiles In the first evaluation scenario, I will use the theoretical 5^{th} and 95^{th} percentiles of $\mathcal{T}_i$ under a valid responses only assumption. These will then be used as cut-off values for $\mathcal{T}_i$ estimated using the theoretical model parameters for valid responses. This will indicate whether $\mathcal{T}_i$ is potentially effective in assigning extreme values to invalid responses based on the respective invalid response strategies. This technique is sensible if the valid response model is known and used to identify invalid responses in another/replication study sample.

Estimated $\mathcal{T}_i$ Parameters and theoretical Percentiles In the second evaluation scenario, I will, once again, use the theoretical 5^{th} and 95^{th} percentiles of $\mathcal{T}_i$ under a valid response only assumption. These will than be used as cut-off values for $\mathcal{T}_i$ estimated using biased estimates from a sample in which invalid responders are present. In other words, I use the true population covariance matrix Σ (and $\boldsymbol{\mu}$) for the valid response model to estimated cut-off values for $\mathcal{T}_i$ but these percentiles are used on $\mathcal{T}_i$ values for all participants that are estimated based on the estimated (biased) valid response model covariance matrix $\hat{\Sigma}$. The second evaluation scenario is primarily implemented to assess how biased estimates for the valid response model are, based on different simulation condition. Furthermore, we will be able to empirically judge to which extent discrimination performance depends on the extent of parameter estimation bias.

Estimated $\mathcal{T}_i$ Parameters and estimated Percentiles In the last evaluation scenario, I will apply a real world scenario where neither theoretical quantiles for $\mathcal{T}_i$ of valid responses nor the theoretical valid response population model are known. Percentiles for $\mathcal{T}_i$ of valid responses will me estimated based on the biased valid response model parameters. In the results section, we will see if biased/estimated valid response models provide enough information to derive cut-off values in order to identify invalid responses. Furthermore, we will be able to define conditions of high and low discrimination power.

7.3 Classification Results

In this section, I will discuss cumulative probabilities of $\mathcal{T}_i$ for the mixed sample simulation conditions. Based on the three different evaluation scenarios, different cut-off values will be chosen.

In the following tables, I differentiate between simulation conditions defined by the combination of the two following experimental factors: percentage of valid versus invalid responses in the sample and measurement accuracy of the valid model (amount of noise in valid responses). Within these combinations, I average percentage of extreme responses (defined through the respective cut-off values) throughout all other simulation conditions. These, in turn, are averaged throughout all 100 replications for each simulation condition. Lastly, the last rows in each condition cell (labelled test) will give flagged responses similar to a two-sided test where extreme values on both sides are added up.

7.3.1 Theoretical $\mathcal{T}_i$ Parameters and theoretical Percentiles

Table 7.2 classifies extreme values of $\mathcal{T}_i$ based on simulated valid and invalid responses. The parameters used to estimate $\mathcal{T}_i$ are the theoretical valid response model parameters. The theoretical 5^{th} and 95^{th} percentiles for valid responses are used as cut-off values for all responses. Table 7.2 shows the results for the first invalid response type, item wording (IW), and for the second invalid response type, long string (LS). A two-sided cut-off for valid responses confirms 10% of simulated valid responses (on each side 5%) are flagged as extreme values.

In order to evaluate the discriminatory potential of $\mathcal{T}_i$ to differentiate between valid and invalid responses, I will focus on the percentage of invalid responses that are flagged as extreme under the most influential experimentally alternated simulation condition, namely, the percentage of error variance in the theoretical valid response model.

Throughout all simulation conditions with 75% of error variance we would on

Table 7.2: Percentage of simulated responses in the groups valid, item wording, and long string identified as extreme values averaged throughout all simulation conditions and replications within conditions, separately presented for different conditions of a priori defined average percentage of error variance in the valid response population model (simulation study evaluation scenario: theoretical parameters and theoretical percentiles)

	Percentage error variance								
	75%			50%			25%		
	Invalid		Valid	Invalid		Valid	Invalid		Valid
	IW	LS		IW	LS		IW	LS	
5th percentile	.00	.28	.05	.12	.55	.05	.64	.87	.05
95th percentile	.28	.09	.05	.14	.03	.05	.04	.00	.05
Two-sided test	**.28**	**.37**	.10	**.27**	**.58**	.10	**.68**	**.87**	.10

IW, LS Invalid response strategy item wording (IW) and long string (LS).

average identify 28% of item wording and 37% of long string responses, where all of item wording responses are assigned positive extreme values and most of long string responses have negative extreme values. In the 50% error variance conditions we see little change in discriminatory potential for item wording responses. However, we have about an equal amount of positive and negative extreme values. On the other side, long string responses are more effective discriminated in 50% error variance conditions with about 58% of them identified as extreme values (mostly negative extreme values). We experience the largest discriminatory potential in simulation conditions with a less severe ratio of measurement error versus explained variance (25% error variance). Hence when the theoretical valid response model is known, we would correctly identify 68% and 87% of invalid responses of type item wording and long string, respectively.

We can conclude, that the discriminatory potential is better for invalid response of type long string and, in general, very successful in discriminating both invalid response types when the measurement model for valid responses does not lack accuracy.

7.3.2 Estimated $\mathcal{T}_i$ Parameters and theoretical Percentiles

Similar to previous Table 7.2, Table 7.3 also classifies extreme values of $\mathcal{T}_i$ based on simulated valid and invalid responses but disaggregates results to show another set of simulation conditions in more detail. Furthermore, the parameters used to estimate $\mathcal{T}_i$ here are the (biased) estimates that arise when we seek to analyse the valid response model based on a sample that includes invalid responses. The ratio of valid versus invalid responses in the sample varies between simulation conditions. However as was the case for previous subsection's results, the theoretical 5^{th} and 95^{th} percentiles for valid responses are used as cut-off values for all responses.

These results will give an indication of the extent of bias between simulation conditions when valid response model parameters are estimated. As a scalar indicator of estimation bias, I will focus on the cumulative probabilities of $\mathcal{T}_i$ for valid responses or, more accurately, percentages of extreme valid responses. Results deviating from expected 5% extreme negative and 5% extreme positive values will indicate the estimation bias. Furthermore, we can compare the percentages of invalid responses classified as extreme values with the respective theoretical results in Table 7.2 from the previous subsection. This will give an indication of the extent of bias that worked in favour of invalid responses, no longer being assigned extreme values in $\mathcal{T}_i$.

We can see strong bias for valid response model parameter estimates in simulation conditions with large amount of noise in the valid response model. This results in a shift of $\mathcal{T}_i$ values to the left (more negative values) for valid responses when invalid responses are present. Furthermore, the bias in 75% measurement error conditions becomes increasingly severe the more invalid responses are in the sample. This is partially the case for both invalid response study types, item wording, with .17, .41, and .65 and, long string, with .10, .24, and .73, with increasing presence of invalid responses, where the long string simulation conditions with 90% valid responses is less affected by this trend. The 50% error variance conditions do not indicate a strong bias when we only look at extreme values in $\mathcal{T}_i$ of valid responses. There is little change to be observed except in simulation conditions with 50% invalid responses in the sample. Bias in simulation conditions with low noise in the valid response model, show a small bias based on valid response percentages flagged as extreme. In fact, the bias seems to result in a decrease of variance in $\mathcal{T}_i$ for valid responses, such that

Table 7.3: Percentage of simulated responses in the groups valid, item wording, and long string identified as extreme values averaged throughout all simulation conditions and replications within conditions, separately presented for different conditions of a priori defined average percentage of error variance in the valid response population model and alternated percentage of valid responders in the sample (simulation study evaluation scenario: estimated parameters and theoretical percentiles)

Percentage error variance		Percentage valid responders						Invalid study type
		90%		70%		50%		
		I	V	I	V	I	V	
75%	5th	.02	**.13**	.10	**.39**	.23	**.63**	Item wording
	95th	.34	**.04**	.32	**.03**	.28	**.02**	
	Test	.35	.17	.42	.41	.51	.65	
50%	5th	.13	**.06**	.15	**.08**	.17	**.11**	
	95th	.16	**.05**	.16	**.04**	.15	**.04**	
	Test	.29	.10	.31	.12	.32	.14	
25%	5th	.55	**.02**	.41	**.00**	.27	**.00**	
	95th	.06	**.05**	.07	**.06**	.07	**.06**	
	Test	.61	.08	.48	.06	.34	.06	
75%	5th	.16	**.06**	.02	**.23**	.08	**.72**	Long string
	95th	.13	**.05**	.37	**.01**	.38	**.00**	
	Test	.28	.10	.39	.24	.46	.73	
50%	5th	.41	**.03**	.11	**.03**	.04	**.09**	
	95th	.04	**.05**	.05	**.04**	.16	**.01**	
	Test	.44	.08	.16	.07	.20	.10	
25%	5th	.79	**.01**	.41	**.00**	.09	**.00**	
	95th	.00	**.06**	.00	**.07**	.01	**.04**	
	Test	.79	.08	.42	.07	.10	.04	

?th P. Left/ right-sided cut-off based on 5th/ 95th percentile.

Test Combined/ two-sided test.

I,V Invalid/ valid sample.

we have very few to no extreme negative values based on theoretical cut-off values.

Another source of information regarding estimation bias can be drawn from investigating changes in the pattern of flagged invalid responses. A general trend throughout all simulation conditions is that the bias seems to work in favour of invalid responses where fewer are assigned extreme values. Especially, the invalid response type long string seems to influence the estimation of valid response model parameters. Another observable trend regarding invalid responses of type item wording is the larger dispersion of their $\mathcal{T}_i$ scores, in comparison to $\mathcal{T}_i$ values when estimated based on the theoretical parameters. Interestingly, there is almost a complete shift of the $\mathcal{T}_i$ for invalid responses, to the right for response of type long string, and to the left for responses of type item wording when $\mathcal{T}_i$ is based on estimated valid response model parameters.

I conclude that there is an increase of bias for the valid response model parameters when there are more invalid responses in the sample. Valid response values in $\mathcal{T}_i$ seem to experience a negative shift in the distribution when estimation bias is severe. Lastly, we lose discriminatory power with increasing severity of estimation bias and increasing ratio of invalid versus valid responses in the sample. This effect occurs stronger for the invalid study type long string.

7.3.3 Estimated $\mathcal{T}_i$ Parameters and estimated Percentiles

Table 7.4 classifies extreme values of $\mathcal{T}_i$ based on simulated valid and invalid responses. The parameters used to estimate $\mathcal{T}_i$ are the (biased) estimates that arise when we seek to analyse the valid response model based on a sample while not taking into account the invalid responses in the sample. The ratio of valid versus invalid responses in the sample varies between simulation conditions. This is a real world scenario where we do not have full knowledge of the theoretical valid response model. Correspondingly, 5^{th} and 95^{th} percentiles for valid responses are estimated based on valid model analysis results and used as cut-off values for all responses.

Ultimately, these results will allow us to evaluate the discriminatory power of $\mathcal{T}_i$ in a realistic study scenario. For this purpose, I will focus on the success of identifying invalid responses as extreme values. Furthermore, I set this success in relation to how many valid responses has been identified as extreme values. Optimally, we would like

Table 7.4: Percentage of simulated responses in the groups valid, item wording, and long string identified as extreme values averaged throughout all simulation conditions and replications within conditions, separately presented for different conditions of a priori defined average percentage of error variance in the valid response population model and alternated percentage of valid responders in the sample (simulation study evaluation scenario: estimated parameters and estimated percentiles)

Perc.		Percentage valid responders									Invalid
error		90%			70%			50%			study
var.											type
		T	I	V	T	I	V	T	I	V	
	5th	.06	.01	.07	.08	.01	.11	.09	.01	.18	Item wording
75%	95th	.05	.30	.03	.07	.20	.01	.07	.13	.00	
	Test	.12	**.30**	**.10**	.15	**.21**	**.12**	.16	**.14**	**.18**	
	5th	.06	.12	.05	.06	.10	.05	.06	.08	.04	
50%	95th	.05	.16	.04	.06	.13	.02	.06	.10	.01	
	Test	.11	**.28**	**.09**	.12	**.23**	**.07**	.12	**.18**	**.06**	
	5th	.08	.56	.03	.14	.44	.01	.15	.30	.00	
25%	95th	.06	.06	.06	.06	.06	.06	.06	.06	.05	
	Test	.14	**.62**	**.08**	.19	**.50**	**.06**	.21	**.37**	**.05**	
	5th	.06	.18	.05	.11	.00	.15	.17	.00	.33	Long string
75%	95th	.06	.12	.05	.09	.28	.00	.09	.17	.00	
	Test	.12	**.29**	**.10**	.20	**.29**	**.16**	.25	**.18**	**.33**	
	5th	.08	.45	.03	.08	.20	.03	.09	.04	.15	
50%	95th	.06	.04	.06	.06	.04	.07	.06	.12	.01	
	Test	.14	**.49**	**.10**	.14	**.24**	**.10**	.16	**.16**	**.16**	
	5th	.09	.81	.01	.18	.58	.00	.11	.22	.00	
25%	95th	.07	.00	.08	.09	.00	.12	.04	.01	.07	
	Test	.16	**.81**	**.09**	.26	**.59**	**.12**	.16	**.24**	**.07**	

$^{\text{T,I,V}}$ T = total sample, I = invalid group, I = valid group.

$^{\text{?th P.}}$ Left/ right-sided cut-off based on 5th/ 95th percentile.

$^{\text{Test}}$ Combined/ two-sided test.

to maximise extreme invalid response values and minimise extreme valid responses.

Simulation conditions with only 10% invalid responses in the sample do not exceed more than 10% as extreme identified valid responses. In contrast, I was able to detect between 28% and 62%, for invalid responses of type item wording, and 81% of invalid responses of type long string. We detect more invalid responses the more accurate the measurement models for valid responses are. The latter trend applies to all simulation conditions. The risk of flagging valid responses can only be found larger than the expected 10% for the high bias conditions (cells in the top-right corner of the table) identified in the previous subsection's results. For invalid simulation studies of type item wording, we experience 12% and 18% of valid responses flagged in 75% error variance conditions, with 10% invalid responses in the sample. For invalid simulation studies of type long string, we can see a similar trend but to a more severe extent, such that we have 16% and even 33% of flagged valid responses. Surprisingly, simulation scenarios with half the sample consisting of invalid responses and where the valid response model provides a good level of measurement accuracy, $\mathcal{T}_i$ proves to possess very large discriminatory power: We detect 37% (item wording) and 24% (long string) of invalid responses, where we would only flag around 6% of valid responses. With regards to the total sample size, this means that we would successfully detect 185 and 120 of invalid responders, where only around 30 valid responders would be incorrectly identified as invalid. In the Table in Appendix A.4 the percentage values in Table 7.4 are translated to actual numbers of responders based on the sample size of 1000, used throughout all simulation conditions. In general, we can see that simulation conditions with good levels of measurement accuracy (25% error variance) are very successful in discriminating between valid and invalid responses. In fact, if we were to use only a left-sided cut-off, we would detect 56% to 30% of item wording, and 56% to 30% of long string responders, with almost no loss of valid responders (ranging from 0 to 3% flagged valid responses).

Concluding from the results in this subsection, there is a tendency of increasing discriminatory power from large error variance to good levels of error variance conditions. We saw that provided with a valid response model that is not affected by large amounts of measurement error, we can successfully detect invalid responders. This is the case even though invalid responses (up to 50%) in the sample increase the measurement error for analysis purposes, when not taken into account. Furthermore, with increasing amount of invalid responders in the sample, the bias in valid response model parameter estimates increases. In simulation scenarios with severe bias where we have large amounts of error variance in the valid response model and large ratio of invalid versus valid responses, the application of $\mathcal{T}_i$ as an identification measure was not successful. Lastly, we saw that the risk of incorrectly identifying valid responses can be reduced with a more sophisticated use of left-sided or right-sided versus two-sided application of cut-off values.

7.4 Further Results and Implications

Results in the previous section revealed that the extent of successful identification of invalid responses is strongly linked to properties of the valid response model. Furthermore, we saw that with bias in valid response model parameter estimates we increase the risk of excluding valid responses while trying to detect invalid responses. Hence, in this section, I will identify situations of high discriminatory potential and of low risk of identifying valid responses.

As a side note, I would like to mention that another invalid study type of long string was implemented in order to identify if estimation bias is the strongest factor in determining discriminatory potential and discriminatory power of $\mathcal{T}_i$. In this scenario, the long string responses were drawn from a uniform discrete distribution of equal intervals in the range spanned by ±3. Similar to a 5-point Likert scale, I allowed for answer options -3, -1.8, -0.6, 0.6 , 1.8, and 3. Furthermore, no noise was added, such that long string responses have correlation matrix with all entries equal to 1 (perfect correlation, linear dependence). In this scenario, the discriminatory potential was very strong. However, the model estimation was too strongly driven by long string responses, such that in conditions with more than 10% invalid responses the risk of excluding invalid responses reached intolerable levels. At this point, it is important to remember that the application of $\mathcal{T}_i$ as identification measure is based on the assumption that invalid responses in a sample are not derived from only single common invalid response model, as is the case for valid responses. However, all simulation studies in this chapter were implemented as such, in order to investigate discriminatory potential and power of $\mathcal{T}_i$ under most conservative settings.

7.4.1 Discriminatory Potential

To identify valid response model properties that increase the discriminatory potential of $\mathcal{T}_i$, two multiple regression results are shown in Table 7.5. The dependent variable is percentage of detected invalid responses when tested on theoretical 5^{th} and 95^{th} percentiles of valid responses as cut-off values for invalid responses. We saw that the measurement accuracy of valid response models have a differential effect on the results. Since $\mathcal{T}_i$ adjusts penalties based on measurement accuracy the discriminatory power is mostly affected by the simulated amount of noise. Hence, I will focus on simulation conditions with only 25% error variance to evaluate which other simulation factors also predict discriminatory potential. This is to identify other properties of valid response models that increase or decrease the potential of assigning extreme $\mathcal{T}_i$ to invalid responses. This multiple regression is applied for the invalid study types item wording and long string separately.

Table 7.5: Multiple regression coefficients - sub-sample: simulation conditions with 25% error variance - dependent variable: percentage of extreme $\mathcal{T}_i$ values of invalid responders - predictors: simulation condition factors

	Invalid study type	
Independent variables	Item wording	Long string
(Intercept)	0.35 (0.01)**	0.52 (0.02)**
Number of LVs	0.03 (0.00)**	0.02 (0.00)**
Percentage of negatively correlated LV	0.05 (0.01)**	0.04 (0.02)
Correlation between LVs	−0.10 (0.01)**	−0.02 (0.02)
Number of indicators per LV	0.03 (0.00)**	0.01 (0.00)**
Percentage of reversed OVs	−0.00 (0.01)	0.66 (0.02)**
Multiple R-squared	.94	.74

OV, LV Observed variable, latent variable.

According to these results, the percentage of extreme $\mathcal{T}_i$ values assigned to invalid responses of type item wording increases with increasing number of latent variables, increasing number of observed variables used as indicators for those latent variables, increasingly orthogonal factor structure, and increasing percentage of latent variables that are negatively correlated with the other latent variables. The order of mentioning is based on decreasing effect sizes evaluated drawing on the

t-values associated with the corresponding predictors. The percentage of observed variables that have a reversed meaning for the interpretation of latent variables (negative factor loading) does not have a significant contribution towards explaining the discriminatory potential ($t = -0.13$). All together this multiple regression model explains about 94% of variation in the dependent variable.

Results for the discriminatory potential with regards to invalid responses of type long string show a slightly different picture. In this scenario, the percentage of observed variables with negative factor loading is by far the strongest predictor ($t = 28.44$). Furthermore, with increasing numbers of latent variables ($t = 6.48$) and observed variables ($t = 6.16$) we increase the percentage of invalid responses that have extreme $\mathcal{T}_i$ values. The remaining two predictors do not significantly improve the prediction of discriminatory potential in contrast to what we have seen in the invalid response type scenario item wording. All predictors together help to explain 74% of the variation in discriminatory potential.

It seems that differences between valid response model and invalid response model are the most important factors when it comes to maximising discriminatory potential. The invalid response model item wording does include information about observed variables with reversed meaning with regards to latent variables. Hence, this explains why increasing items with negative factor loadings did not have a significant effect on discriminatory potential. In contrast, when we look at the long string invalid response study, the same predictor is the largest contributor. The only other significant predictors help to either increase the information in the data by increasing number of observed variables or provide a multi-factor structure as opposed to the single latent variable structure in the invalid response model. To reiterate, $\mathcal{T}_i$ is defined such that it only allows for large penalties when the measurement model for valid responses is of good quality, e.g. low residual error variance and high factor loadings.

7.4.2 Risk of extreme valid Response Values

In order to identify valid response model properties that increase the risk of identifying valid responses as invalid, two multiple regression results are shown in Table 7.6. The dependent variable is percentage of valid responses that are assigned extreme values in $\mathcal{T}_i$ when tested on estimated 5^{th} and 95^{th} percentiles of valid responses. This multiple regression is applied for the invalid study types item wording and long string, separately.

Table 7.6: Multiple regression coefficients - dependent variable: percentage of extreme $\mathcal{T}_i$ values of valid responders - predictors: simulation condition factors

	Invalid study type	
Independent variables	Item wording	Long string
(Intercept)	0.02 (0.01)*	0.07 (0.02)**
Percentage of valid responders	−0.01 (0.01)*	−0.22 (0.02)**
Percentage of error variance	0.13 (0.00)**	0.20 (0.01)**
Number of LVs	−0.00 (0.00)	0.01 (0.00)**
Percentage of negatively correlated LV	0.02 (0.00)**	0.02 (0.01)
Correlation between LVs	0.04 (0.00)**	−0.02 (0.01)
Number of indicators per LV	0.00 (0.00)	0.01 (0.00)**
Percentage of reversed OVs	0.00 (0.00)	0.19 (0.01)**
Multiple R-squared	.52	.44

OV, LV Observed variable, latent variable.

The multiple regression results for the invalid study type item wording reveal four of the in total seven predictors to significantly predict the risk of extreme $\mathcal{T}_i$ values for valid responders. By far the strongest predictor is the percentage of error variance in the valid response model, where with decreasing measurement accuracy we increase the risk of flagging valid responses ($t = 30.22$). Further risk factors are correlation between latent variables ($t = 9.49$), percentage of latent variables that are negatively correlated with the majority of the remaining latent variables ($t = 3.74$), and percentage of valid responders in the sample ($t = -2.56$). Overall, all predictors together explain around 52% of the variation within the dependent variable risk of flagging valid responders.

In contrast to that, the results for the invalid study type long string shows three equally strong effects of the predictors percentage of valid responders ($t = -14.33$), percentage of error variance ($t = 16.14$), and percentage of observed variables with reversed measurement of latent variables ($t = 15.47$). Furthermore similar to the results of discriminatory power, we gain two more significant predictors: number of latent variables and observed variables ($t = 4.47/4.74$). In total, all predictors help to explain around 44% of variation in risk between different simulation conditions.

In conclusion, the percentage of error variance versus explained variance of observed variables for the valid response model is a consistent criterion when we seek to assess the risk of flagging valid responders. Furthermore, the more aspects of valid response model are distinct from the invalid response model, such as distinct factor structure, the less likely are we to risk flagging valid responders. Lastly, the more (distinct) information the data can provide (e.g., increasing number of observed variables and latent variables) the easier it is to avoid the risk of flagging valid responders.

Chapter 8

Discussion

This thesis provides an extensive review on identification measures for undesired response patterns in the sample. The review concludes with the problem that most detection instruments are developed for categorical data. Furthermore, those statistics are frequently correlated with participants' response patterns who have extreme but valid latent trait scores. A second review focuses on latent class analysis as another method for dealing with semi-plausible response patterns by accommodating invalid response strategies into the model. Based on findings of introduced studies using LCA in similar contexts, I conclude that LCA in combination with identification measures provides a powerful tool for dealing with SpRPs. However, there are several disadvantages associated with this method, i.e. the accessibility for non-expert audiences, the requirement for case-specific implementations, and computational difficulties. Furthermore, in order to define an appropriate model, we require knowledge about the nature of employed invalid response strategies and may only account for a small number of such strategies (subject to model identification).

An experimental study and an online questionnaire study provide data for the analysis of the valid response model. Structural differences in the estimated parameters between experimentally induced valid and invalid response settings as well as a generalisable setting (online questionnaire study) helped to derive a possible invalid response strategy (item wording). In a second step, I successfully accommodate this invalid response strategy into the model using a factor mixture model. Model fit indices were significantly improved using LCA and findings validated the nature of the derived item wording invalid response strategy.

The main focus of the thesis lies on the development of a new identification measure ($\mathcal{T}_i$) to efficiently detect SpRPs in the sample for continuous latent variable models. $\mathcal{T}_i$ is theoretically derived and interpreted as well as evaluated in empirical and simulated scenarios. The new measure was conceptualised such that it adjusts for extreme but valid factor scores, model accuracy (versus measurement error), and is easily implementable as well as universally applicable in a wide range of continuous data scenarios. Its unique nature allows for the identification of all kinds of invalid response strategies without the need of prior knowledge about their characteristics. In fact, in a second step, we can use $\mathcal{T}_i$ to derive characteristics of invalid response strategies if we so wish to do. $\mathcal{T}_i$ proves to be successful in identifying participants of experimentally induced semi-/implausible groups, at the same time allowing for adjusting the risk of incorrectly flagging valid responses (subject to case-specific needs). Furthermore, factors that are important for maximising the discrimination power of $\mathcal{T}_i$ were identified in a simulation study. I also show that combining LCA with $\mathcal{T}_i$ can increase the detection rate of invalid responses even further. Ultimately, the thesis introduces a fairly new problem and provides solutions including a new detection measure that addresses issues that are not covered by extant literature or other existing methods.

Findings of this thesis are as usual subject to limitations. Where the LCA approach in this thesis is used in a unique combined manner with $\mathcal{T}_i$, it focuses on the detection of SpRPs rather than the appropriate accommodation of invalid response strategies. Priorities in this thesis were intentionally set as justified extensively in the first four chapters to ensure that the methodology can be used in a wide range of settings and achieve a higher degree of generalisability through an appropriate balance of parsimony and complexity. Nonetheless, in order to fully take advantage of the unique experimental design of data used in this thesis, it would be interesting to see if a hybrid approach for the present categorical ordered data in a factor mixture model can be used to improve model fit. Following the example of previous studies in this field, the model can be further enriched through the use of covariates for the latent class variable (e.g., demographical data or identification measures). We could further allow the Big Five personality factor Conscientiousness to affect class membership and test whether semi-plausible data, in fact, does not provide any information to the constructs of interest.

Furthermore, I focused on measures that can be taken to deal with SpRPs after the data collection process and, consequently within this scope, I did not reach the topic of prevention, i.e. study design. Many aspects in the study design have an effect on participants' motivation to give accurate answers. Amongst those most commonly mentioned is to keep the survey length (number of questions) to a minimum in order to prevent so-called tiring-out effects. Furthermore, I recommend to appeal to participants' intrinsic motivation by giving non-monetary incentives such as a questionnaire feedback on their performance or on other measures that can be derived from the constructs of interest (e.g., summary on personality data). Hence, incentives are directly linked to and require the participant to care about their answers. It might be helpful to provide 'don't know' or neutral answer options without the penalty of incentive reduction for filling out the survey, such that invalid response strategies are easily filtered out. However, there are many disadvantages and quite a bit of discussion around these topics. For the interested reader I recommend textbooks on data quality or more general topics on survey methodology (e.g., Leeuw et al., 2008). Since there is no well established guidelines on data collected via the use of micro-jobbers, I recommend experimenting on the collection of data for the mere assessment of data quality first, such as setting different monetary incentive sizes (e.g., amount of money) or different micro-jobbing platforms.

The extracted information from the review of identification measures was used for the development of a measure that addresses previous issues with existent methods but it remains to be seen how $\mathcal{T}_i$ compares to other identification measures. The comparison was judged secondary to other forms of evaluation of $\mathcal{T}_i$ in light of already existing other studies that impressively compare the discrimination power of many identification measures (e.g., Meijer and Sijtsma, 2001; Karabatsos, 2003; Meade and Craig, 2012). Furthermore, although the data and simulated scenarios were chosen to allow drawing generalisable conclusions to a certain extent for the discrimination power of $\mathcal{T}_i$, it remains to be seen whether this is true. $\mathcal{T}_i$ was tested on personality self-assessment data and simulated settings with a limited number of structural complexity in the latent variables. Findings suggests that adding complexity to the model does improve discrimination power. I assume this applies as well when additional covariates are part of the valid response model. We saw that even when assumptions of the model are not met (i.e. multivariate normality of the data) and the sample consists of a majority of invalid responses, $\mathcal{T}_i$ performs well in discriminating between valid and invalid responders. Hence, $\mathcal{T}_i$ is evaluated based on data that meets the assumptions (simulated data) and on data that provides asymptomatic normality only (Likert-type).

Another limitation is that $\mathcal{T}_i$'s performance was established based on very specific evaluation criteria, such as a 10% tolerance rate for incorrectly flagging valid responses. I acknowledge that in some research areas different thresholds bare a varying magnitude of risk. However, users are advised to adjust cut-off thresholds to their own needs and adjust risk levels based on actual (estimated or hypothesised) number of valid versus invalid responses in the sample. Another option that has not been discussed in this thesis but has potential in decreasing the risk of flagging valid responses is a step-by-step detection of SpRPs. In cases where we are particularly worried about the incorrect classification risk, we could refit the model with a sub-sample that does not include the individual with the most aberrant response pattern, identified via identification measures or classified using posterior probabilities based on LCA. Algorithms of this kind can become computationally very extensive depending on several aspects of the study setting, such as sample size, estimated proportion of SpRPs present in the sample, and the complexity of the valid response model. In the case of $\mathcal{T}_i$, this method will reduce estimation bias and measurement error in the valid response model caused by SpRPs at each step. In doing so, we would gradually receive more accurate sources of information for $\mathcal{T}_i$. A better information source for $\mathcal{T}_i$ (e.g., more accurate estimates of the valid response model) will not only improve discriminatory power, but also help to accurately set the desired risk threshold of flagging valid responses. Similar work exists using forward search algorithms in identifying outliers, where several measures such as goodness-of-fit statistics or residuals are utilised and summarised within plots to support informed decisions (e.g., Mavridis and Moustaki, 2008, 2009).

The simulation study helped to identify scenarios in which $\mathcal{T}_i$ performs better or worse. So far with the simulation study, I evaluated $\mathcal{T}_i$'s discrimination power based on two distinct kinds of invalid response strategies, covering a basic and a more complex version of semi-plausible responding. However, it remains to be seen how well $\mathcal{T}_i$ performs when participants employ other types of invalid response strategies, such as partly invalid responses, where respondents switch from a valid to an invalid response strategy throughout the questionnaire. The simulation study focused on the percentage of incorrectly identified valid responses and correctly identified invalid responses. Due to the ratio of valid versus invalid responses, where invalid responses optimally represent the minority in the sample, a discussion is necessary when comparing actual numbers of flagged valid versus invalid responses. For example, in a real world simulation scenario $\mathcal{T}_i$ flagged 8% and 9% valid, and 81% and 62% invalid responses for the item wording and long string invalid response strategies, respectively (see Table 7.4). However for a sample size of $n = 1000$ and 10% invalid responses in the sample, these percentages translate to flagged 72 and 81 valid, and 62 and 81 invalid responses (see Table in Appendix A.4). These are almost equivalent numbers of valid and invalid participants that when excluded from the analysis sample can reduce bias but also induce another form of bias for the estimation of the valid response model, i.e. a bias resulting from the extraction of 5 to 7 percent valid responses from the sample. However, we also saw that if we were to employ only a one-sided, or better left-sided, cut-off we flag 81% (81) invalid versus 1% (9) valid response in the item wording simulation scenario and 56% (56) invalid versus 3% (27) valid responses in the long string simulation scenario. Hence, we dramatically decrease the risk of flagging valid responses. Even more so based on the simulation study results, I assume that the ratio of flagged responders' groups will be more optimal when the presence of SpRPs in the sample more severely affects the estimation of the valid response model. Hence, sacrificing some number of valid responses will be worth the exclusion of invalid responders who have a strong influence on the estimation of the valid response model. However, this assumption needs to be tested. Future research should focus on identifying a balance of benefits of excluding invalid responses and the risk that is associated with incorrectly excluding valid responses from the sample. One solution for this dilemma could be the use of very conservative cut-off threshold(s) such that the risk of flagging invalid responses is minimised, even if that means that only the most severe of SpRPs are detected. To reiterate, when we are concerned about the risk of flagging valid responses, it would be advisable to use only a left-sided cut-off criteria as was theoretically justified (see Section 5.1) and empirically shown (see Section 5.3) to be a more conservative criterion.

Ultimately, we saw that $\mathcal{T}_i$ is a very promising instrument for the detection of all kinds of undesired responses. As with any other approach it has advantages and disadvantages. Some of the disadvantages can be accounted for when used in combination with LCA. Although LCA is usually used to accommodate invalid response strategies into model, it proves useful in detecting SpRPs in combination with identification measures. In this context, I hope I was able to contribute to the research topic not only by providing new instruments, but also that the narrative helped to gain a better understanding of invalid responses, a set of elaborate guidelines, and a more sophisticated stand towards invalid responses and their consequences. Furthermore, the original methods developed, used, and evaluated in this thesis certainly not only have an application in the introduced setting where the increasing use of micro-jobbers in social sciences is a problem but can also be extended to the use in other applications. The essential logic behind the development of $\mathcal{T}_i$ lies in contrasting non-model specific information (e.g., null model) to model specific information (e.g., hypothesised/restricted and estimated model) to filter out unique information for model abberrant responses. Measures to detect model aberrant responses are numerous, but those do not control for information that is not model specific. Therefore, $\mathcal{T}_i$ can complement those existing measures. Even more so, the essential logic behind $\mathcal{T}_i$ can be extended to wide range of other model, e.g. models for categorical data with latent variables or time series models. Theoretically and empirically we have seen that newly developed measure $\mathcal{T}_i$ acts unique in the manner it approaches outliers. For instance, it can be used in industrial settings for the detection of atypical mechanisms or fraud. Another example is its potential application in information technology such as cyber security where the recent focus primarily lies on machine learning algorithms but could certainly be enriched with more sophisticated statistical methodology.

Bibliography

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, *144*(4), 419. doi:10.2307/2981826

Allison, P. D. (2009). Missing data. In R. E. Millsap, and A. Maydeu-Olivares (Eds.), *The sage handbook of quantitative methods in psychology* (pp. 72–89). Thousand Oaks, CA: Sage Publications.

Asparouhov, T., and Muthen, B. (2014). *Variable-specific entropy contribution.* Technical appendices related to new features in version 7. Retrieved from http://www.statmodel.com/download/UnivariateEntropy.pdf

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading* (pp. 233–251). Oxford: Basil Blackwell.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). Wiley series in probability and statistics. Hoboken, N.J.: Wiley.

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

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *Eqs structural equations program manual.* Multivariate Software. Encino, CA.

Bergstrom, B., Gershon, R., and Lunz, M. (1994). Computerized adaptive testing exploring examinee response time using hierarchical linear modeling. Paper presented at the annual meeting. National Council on Measurement in Education. New Orleans, Louisiana.

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

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., and Monroe, K. (1992). Mmpi-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340–345. doi:10.1037/1040-3590.4.3.340

Birenbaum, M. (1985). Comparing the effectiveness of several irt based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement*, *45*(3), 523–534. doi:10.1177/001316448504500309

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement*, *10*(2), 167–174. doi:10.1177/014662168601000208

Bohman, H. (1975). Numerical inversions of characteristic functions. *Scandinavian Actuarial Journal*, *1975*(2), 121–124. doi:10.1080/03461238.1975.10405087

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*(2), 290–302. Retrieved from http://www.jstor.org/stable/2236731

Brown, A., and Croudace, T. (2015). Scoring and estimating score precision using multidimensional irt. In S. P. Reise, and D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment (a volume in the multivariate applications series)*. New York: Routledge/Taylor and Francis Group.

Browne, M. W., and Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, and J. S. Long (Eds.), *Testing structural equation models*. Newbury Park: Sage Publications.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Byrd, R. H., Nocedal, J., and Schnabel, R. B. (1994). Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, *63*(1), 129–156. doi:10.1007/BF01582063

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., and Chin, T.-Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *Bulletin de Méthodologie Sociologique*, *103*(1), 5–25. doi:10.1177/075910630910300103

Cannell, C. F., Miller, P. V., and Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389–437). CA, San Francisco: Jossey-Bass.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. doi:10.1207/s15327906mbr0102_10

Celeux, G., and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*(2), 195–212. doi:10.1007/BF01246098

Chen, F., Bollen, K., Paxton, P., Curran, P. J., and Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, *29*(4), 468–508. doi:10.1177/0049124101029004003

Cialdini, R. B. (1993). *Influence: Science and practice* (3rd). New York: Harper Collins.

Conijn, J. M., Emons, W. H. M., van Assen, M. A. L. M., and Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research*, *46*(2), 365–388. doi:10.1080/00273171.2010.546733

Costa, P. T., and McCrae, R. R. (1995). Solid ground in the wetlands of personality: A reply to block. *Psychological Bulletin*, *117*(2), 216–220. doi:10.1037/0033-2909.117.2.216

Cote, J. A., and Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research*, *24*(3), 315–318. doi:10.2307/3151642

Couch, A., and Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*(2), 151–174. doi:10.1037/h0040372

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. doi:10.1177/001316444600600405

Curran, P. G., Kotrba, L., and Denison, D. (2010). Careless responding in surveys: Applying traditional techniques to organizational settings. Paper presented at the 25th annual conference. Society for Industrial/ Organizational Psychology. Atlanta, GA.

Davies, R. B. (1973). Numerical inversion of a characteristic function. *Biometrika*, *60*(2), 415–417. doi:10.1093/biomet/60.2.415

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of chi-square random variables. *Applied Statistics*, *29*(3), 323–333. doi:10.2307/2346911

Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology*, *73*(6), 1246–1256. doi:10.1037/0022-3514.73.6.1246

Donlon, T. F., and Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*(1), 105–113. doi:10.1177/001316446802800110

Drasgow, F., Levine, M. V., and McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*(1), 59–79. doi:10.1177/014662168701100105

Drasgow, F., Levine, M. V., and McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, *15*(2), 171–191. doi:10.1177/014662169101500207

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x

Duchesne, P., and Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics and Data Analysis*, *54*(4), 858–862. doi:10.1016/j.csda.2009.11.025

Eriksson, K., and Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*(3), 159–163. Retrieved from https://search.proquest.com/docview/1011287632

Farebrother, R. W. (1984). Algorithm as 204: The distribution of a positive linear combination of chi-square random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *33*(3), 332–339. Retrieved from http://www.jstor.org/stable/2347721

Farebrother, R. W. (1990). Algorithm as 256: The distribution of a quadratic form in normal variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *39*(2), 294–309. Retrieved from http://www.jstor.org/stable/2347778

Fowler, H. M. (1954). An application of the ferguson method of computing item conformity and person conformity. *The Journal of Experimental Education*, *22*(3), 237–246. doi:10.1080/00220973.1954.11010480

Fraley, R. C. (2004). *How to conduct behavioral research over the internet: A beginner's guide to html and cgi/perl*. New York: The Guilford Press.

Friedman, H. S., Tucker, J. S., Schwartz, J. E., Martin, L. R., Tomlinson-Keasey, C., Wingard, D. L., and Criqui, M. H. (1995). Childhood conscientiousness and longevity: Health behaviors and cause of death. *Journal of Personality and Social Psychology*, *68*(4), 696–703. doi:10.1037/0022-3514.68.4.696

Frith, C., and Frith, U. (2005). Theory of mind. *Current Biology*, *15*(17), R644–R645. doi:10.1016/j.cub.2005.08.041

Gil-Pelaez, J. (1951). Note on the inversion theorem. *Biometrika*, *38*(3/4), 481. doi:10.2307/2332598

Glaser, R. (1949). A methodological analysis of the inconsistency of response to test items. *Educational and Psychological Measurement*, *9*, 727–739. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/001316444900900408

Glaser, R. (1950). Multiple operation measurement. *Psychological Review*, *57*(4), 241–253. doi:10.1037/h0057126

Glaser, R. (1951). The application of the concepts of multiple-operation measurement to the response patterns on psychological tests. *Educational and Psychological Measurement*, *11*(3), 372–382. doi:10.1177/001316445101100307

Glaser, R. (1952). The reliability of inconsistency. *Educational and Psychological Measurement*, *12*(1), 60–64. doi:10.1177/001316445201200106

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26–34. doi:10.1037/0003-066X.48.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several fivefactor models. In I. Mervielde, I. Deary, F. D. Fruyt, and F. Ostendorf (Eds.), *Personality psychology in europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. (2000, July 20). Personal Communication cited in Johnson (2005).

Goldberg, L. R., and Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*, *48*(1), 82–98. doi:10.1037/0022-3514.48.1.82

Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93–104. doi:10.1037/0003-066X.59.2.93

Greszki, R., Meyer, M., and Schoen, H. (2014). The impact of speeding on data quality in nonprobabilityy and freshly recruited probability-based online panels. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Goritz, J. A. Krosnick, and P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective*. Chichester, UK: John Wiley and Sons.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150. doi:10.2307/2086306

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Claussen (Eds.), *Measurement and prediction* (Vol. 4, pp. 66–90). Studies in social psychology in world war II. Princeton: Princeton University Press.

Halkitis, P. N. (1996). Estimating testing time: The effects of item characteristics on response latency. Paper presented at the annual meeting. American Educational Research Association. New York, New York. Retrieved from http://files.eric.ed.gov/fulltext/ED397119.pdf

Harnisch, D. L., and Linn, R. L. (1981). Analysis of item response patterns. questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*(3), 133–146. doi:10.1111/j.1745-3984.1981.tb00848.x

Harnisch, D. L., and Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver: Kluwer.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, *21*(3), 360–373. doi:10.1177/0894439303253985

Hodge, V. J., and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126. doi:10.1007/s10462-004-4304-y

Hu, L., and Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling. concepts, issues, and applications* (pp. 76–99). London: Sage Publications.

Hu, L., and Bentler, P. M. (1998). Fit indices in covariance structure analysis: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453. doi:10.1037/1082-989X.3.4.424

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., and DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. doi:10.1007/s10869-011-9231-8

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, *48*(3/4), 419. doi:10.2307/2332763

Ipeirotis, P. G. (2010). Demographics of mechanical turk [(tech. rep. no. ceder-10-01)]. New York University. New York. Retrieved from http://hdl.handle.net/2451/29585

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Proceedings of the Association for Research in Personality*, *39*(1), 103–129. doi:10.1016/j.jrp.2004.09.009

Jöreskog, K. G., and Sörbom, D. (1993). *Structural equation modeling with the simplis command language*. Chicago: Scientific Software.

Judge, T. A., Higgins, C. A., Thoresen, C. J., and Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*(3), 621–652. doi:10.1111/j.1744-6570.1999.tb00174.x

Kaczmirek, L. (2009). *Human-survey interaction: Usability and nonresponse in online surveys*. Cologne: Herbert von Halem Verlag.

Kahn, R. L., and Cannell, C. F. (1957). *The dynamics of interviewing*. New York: John Wiley and Sons.

Kaminska, O., McCutcheon, A. L., and Billiet, J. (2011). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, *74*(5), 956–984. doi:10.1093/poq/nfq062

Kane, M. T., and Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, *4*(1), 105–126. doi:10.1177/014662168000400111

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage Publications.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298. doi:10.1207/S15324818AME1604_2

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the rasch model. *Psychometrika*, *56*, 535–547. doi:10.1007/BF02294459

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer, and I. W. Molenaar (Eds.), *Rasch models. foundations, recent developments, and applications* (pp. 97–110). New York: Springer Verlag.

Klauer, K. C., and Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 193–206. doi:10.1111/j.2044-8317.1990.tb00935.x

Kogut, J. (1986). Review of irt-based indices for detecting and diagnosing aberrant response patterns. Research Report 86-4. University of Twente, Department of Education. Enschede.

Kolmogrov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell Istituto Italiano degli Attuari*, (4), 83–91.

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

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*, 537–567. doi:10.1146/annurev.psych.50.1.537

Krosnick, J. A., and Fabrigar, L. R. (2001). *Designing good questionnaires: Insights from psychology*. New York: Oxford University Press.

Krosnick, J. A., Narayan, S., and Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29–44. doi:10.1002/ev.1033

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, *86*(4), 929–935. doi:10.1093/biomet/86.4.929

Kurtz, J. E., and Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the neo-pi-r. *Journal of Personality Assessment*, *76*(2), 315–332. doi:10.1207/S15327752JPA7602_12

Lange, K., Westlake, J., and Spence, M. A. (1976). Extensions to pedigree analysis iii. variance components by the scoring method. *Annals of Human Genetics*, *39*(4), 485–491. doi:10.1111/j.1469-1809.1976.tb00156.x

Leeuw, E. D. d., Hox, J. J., and Dillman, D. A. (Eds.). (2008). *International handbook of survey methodology*. EAM book series. New York: Lawrence Erlbaum Associates.

Levine, M. V., and Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, *35*(1), 42–56. doi:10.1111/j.2044-8317.1982.tb00640.x

Levine, M. V., and Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109–131). New York: Academic Press.

Levine, M. V., and Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, *53*(2), 161–176. doi:10.1007/BF02294130

Levine, M. V., and Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics*, *4*(4), 269–290. doi:10.3102/10769986004004269

Li, M.-n. F., and Olejnik, S. (1997). The power of rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*(3), 215–231. doi:10.1177/01466216970213002

Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402. doi:10.1080/01621459.1967.10482916

Liu, H., Tang, Y., and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis*, *53*(4), 853–856. doi:10.1016/j.csda.2008.11.025

Lord, F. M., and Novick, M. R. (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubke, G. H., and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. doi:10.1037/1082-989X.10.1.21

MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. doi:10.1037/1082-989X.1.2.130

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, *2*, 49–55.

Malhotra, N. (2009). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, *72*(5), 914–934. doi:10.1093/poq/nfn050

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530. doi:10.1093/biomet/57.3.519

Mason, W. A., and Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, *44*(1), 1–23. doi:10.3758/s13428-011-0124-6

Mason, W. A., and Watts, D. J. (2009). Financial incentives and the 'performance of crowds'. *Association for Computing Machinery Explorations Newsletter*, *11*(2), 100–108.

Mavridis, D., and Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate behavioral research*, *43*(3), 453–475. doi:10.1080/00273170802285909

Mavridis, D., and Moustaki, I. (2009). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, *18*(4), 1016–1034. doi:10.1198/jcgs.2009.08060

McGrath, R. E., Mitchell, M., Kim, B. H., and Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. doi:10.1037/a0019216

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

Medsker, G. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management*, *20*(2), 439–464. doi:10.1016/0149-2063(94)90022-1

Meijer, R. R. (1994). The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311–314. doi:10.1177/014662169401800402

Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology*, *71*(2), 147–160. doi:10.1111/j.2044-8325.1998.tb00668.x

Meijer, R. R., Muijtjens, A. M. M., and van der Vlueten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues an empirical example. *Applied Measurement in Education*, *9*(1), 77–89. doi:10.1207/s15324818ame0901_7

Meijer, R. R., and Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*(3), 261–272. doi:10.1207/s15324818ame0803_5

Meijer, R. R., and Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135. doi:10.1177/01466210122031957

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300–307. doi:10.1037/0033-2909.115.2.300

Miller, J. (2006). Research reveals alarming incidence of 'undesirable' online panelists. Research Conference Report. RFL Communications. Skokie, IL. Retrieved June 19, 2014, from http://burke.com/library/articles/jeff%20miller%20rcr%20pdf.pdf

Miller, J., Officer, C. O., and Baker-Prewitt, J. (2009). Beyond 'trapping' the undesirable panelist: The use of red herrings to reduce satisficing. PUB. NO. CP51. CASRO Panel Quality Conference. Retrieved June 19, 2014, from http://www.survey4.burke.com/Library/Conference/Beyond%20Trapping%20the%20Undesirable%20Panelist_FINAL.pdf

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. Berlin: Walter de Gruyter.

Molenaar, I. W., and Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75–106. doi:10.1007/BF02294745

Molenaar, I. W., and Hoijtink, H. (1996). Person-fit and the rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, *9*(1), 27–45. doi:10.1207/s15324818ame0901_4

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach: Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality and Quantity*, *37*(3), 277–302. doi:10.1023/A:1024472110002

Morren, M., Gelissen, J. P., and Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. *Sociological Methodology*, *41*(1), 13–47. doi:10.1111/j.1467-9531.2011.01238.x

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, *47*(4), 355–366. doi:10.1037/h0059934

Moustaki, I., and Knott, M. (2014). Latent variable models that account for atypical responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(2), 343–360. doi:10.1111/rssc.12032

Muthén, L. K., and Muthén, B. O. (1998–2012). *Mplus user's guide*. 7th ed. Muthén and Muthén. Los Angeles, CA.

Nash, J. C. (2014). On best practice optimization methods in r. *Journal of Statistical Software*, *60*(2), 1–14. doi:10.18637/jss.v060.i02

Nash, J. C., and Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: Optimx for r. *Journal of Statistical Software*, *43*(9), 1–14. doi:10.18637/jss.v043.i09

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, *19*(2), 121–129. doi:10.1177/014662169501900201

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, *21*(2), 115–127. doi:10.1177/01466216970212002

Nering, M. L., and Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, *22*(1), 53–69. doi:10.1177/01466216980221004

Neuringer, A. (1986). Can people behave 'randomly?': The role of feedback. *Journal of Experimental Psychology: General*, *115*(1), 62–75. doi:10.1037/0096-3445.115.1.62

Nichols, D. S., Greene, R. L., and Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the mmpi: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2), 239–250. doi:10.1002/1097-4679(198903)45:2⟨239::AID-JCLP2270450210⟩3.0.CO;2-1

Noonan, B. W., Boss, M. W., and Gessaroli, M. E. (1992). The effect of test length and irt model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement*, *16*(4), 345–352. doi:10.1177/014662169201600405

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. Retrieved from http://ssrn.com/abstract=1626226

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, and D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum.

Pontin, J. (2007, March 25). Artificial intelligence: With help from the humans. *The New York Times*. Retrieved from http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen and Lydiche.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*(3), 213–229. doi:10.1177/014662169501900301

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in irt models. *Multivariate Behavioral Research*, *35*(4), 543–568. doi:10.1207/S15327906MBR3504_06

Reise, S. P., and Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods*, *4*(1), 3–21. doi:10.1037/1082-989X.4.1.3

Reiser, M. (1996). Analysis of residuals for the multionmial item response model. *Psychometrika*, *61*(3), 509–528. doi:10.1007/BF02294552

Reiser, M., and VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, *47*(1), 85–107. doi:10.1111/j.2044-8317.1994.tb01026.x

Rogers, H. J., and Hattie, J. A. (1987). A monte carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement*, *11*(1), 47–57. doi:10.1177/014662168701100103

Rossmann, J. (2010). Data quality in web surveys of the german longitudinal election study 2009. Paper presented at the 3rd Graduate Conference. ECPR. Dublin.

Ruben, H. (1962). Probability content of regions under spherical normal distributions, iv: The distribution of homogeneous and non-homogeneous quadratic functions of normal variables. *The Annals of Mathematical Statistics*, *33*(2), 542–570. Retrieved from http://www.jstor.org/stable/2237533

Rudas, T., Clogg, C. C., and Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(4), 623–639. Retrieved from http://www.jstor.org/stable/2346187

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, *20*(3), 207–219. doi:10.1111/j.1745-3984.1983.tb00200.x

Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. doi:10.1006/ceps.1999.1020

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *35*(1), 139. doi:10.1007/BF02290599

Sato, T. (1975). *The construction and interpretation of s-p tables*. Tokyo: Meiji Tokyo.

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpr-online/issue20/art2/mpr130_13.pdf

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., and Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*(1), 41–53. doi:10.1177/01466219922031176

Schuman, H., and Presser, S. (1996). *Question and answers in attitude surveys*. Thousand Oaks, CA: Sage Publications.

Schwarz, N., and Strack, F. (1985). Cognitive and affective processes in judgments of subjective well-being: A preliminary model. In H. Brandstatter, and E. Kirchler (Eds.), *Economic psychology* (pp. 439–447). Linz, Austria: R. Tauner.

Sheil, J., and O'Muircheartaigh, I. (1977). Algorithm as 106: The distribution of non-negative quadratic forms in normal variables. *Applied Statistics*, *26*(1), 92. doi:10.2307/2346884

Sherif, M., and Cantril, H. (1945). The psychology of 'attitudes': Part i. *Psychological Review*, *52*(6), 295–319. doi:10.1037/h0062252

Sherif, M., and Cantril, H. (1946). The psychology of 'attitudes': Part ii. *Psychological Review*, *53*(1), 1–24. doi:10.1037/h0058561

Shook, C. L., Ketchen, D. J., Hult, G. T. M., and Kacmar, K. M. (2004). An assessment of the use of structural equation modeling in strategic management research. *Strategic Management Journal*, *25*(4), 397–404. doi:10.1002/smj.385

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, *7*(22), 131–145. Retrieved from https://pure.uvt.nl/portal/files/1030745/COEFFICI.PDF

Sijtsma, K., and Meijer, R. R. (1992). A method for investigating the intersection of item response functions in mokken's nonparametric irt model. *Applied Psychological Measurement*, *16*(2), 149–157. doi:10.1177/014662169201600204

Simon, H. A. (1957). *Models of man.* New York: Wiley.

Skinner, C. J. (2014, December 4). *Mahalanobis distance under a factor analysis model.* Email communication.

Smith, R., and Brown, H. H. (2005). *Assessing the quality of data from online panels: Moving forward with confidence.* White paper. Rochester, NY: Harris Interactive.

Smith, R. M. (1985). A comparison of rasch person analysis and robust estimators. *Educational and Psychological Measurement*, *45*(3), 433–444. doi:10.1177/001316448504500301

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342. doi:10.1007/BF02294437

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x

Steiger, J. H. (1990). Structural model evaluation and modication: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180. doi:10.1207/s15327906mbr2502_4

Sundre, D. L. (1999, April 1). Does examinee motivation moderate the relationship between test consequences and test performance? Paper presented at the annual meeting. American Educational Research Association. Montreal.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*(1), 95–110. doi:10.1007/BF02294208

Tatsuoka, K. K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education*, *9*(1), 65–75. doi:10.1207/s15324818ame0901_6

Tatsuoka, K. K., and Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, *20*(3), 221–230. doi:10.1111/j.1745-3984.1983.tb00201.x

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. doi:10.1037/h0070288

Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, and R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 149–162). Washington, DC: National Academic Press.

Tourangeau, R. (1987). Attitude measurement: A cognitive perspective. In H. J. Hippler, N. Schwarz, and S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 149–162). New York: Springer Verlag.

Tourangeau, R., Couper, M. P., and Conrad, F. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Tourangeau, R., Couper, M. P., and Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *The Public Opinion Quarterly*, *68*(3), p 368–393. Retrieved from http://www.jstor.org/stable/3521676

Tourangeau, R., and Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299–314. doi:10.1037/0033-2909.103.3.299

Trabin, T. E., and Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York: Academic Press.

Tucker, L. R., and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. doi:10.1007/BF02291170

Tune, G. S. (1964a). A brief survey of variables that influence random-generation. *Perceptual and Motor Skills*, *18*(3), 705–710. doi:10.2466/pms.1964.18.3.705

Tune, G. S. (1964b). Response preferences: A review of some relevant literature. *Psychological Bulletin*, *61*(4), 286–302. doi:10.1037/h0048618

Uebersax, J. (2000, August 10). A brief study of local maximum solutions in latent class analysis. Retrieved February 5, 2016, from http://www.john-uebersax.com/stat/local.htm

van der Linden, D., te Nijenhuis, J., and Bakker, A. B. (2010). The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327. doi:10.1016/j.jrp.2010.03.003

van der Linden, W., and Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer Verlag.

van Der Flier, H. (1977). Environmental factors and deviant response patterns. In Y. P. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 30–35). Amsterdam: Swets and Zeitlinger.

van Der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties*. [Comparability of individual test performance]. Lisse: Swets and Zeitlinger.

van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*(3), 267–298. doi:10.1177/0022002182013003001

van Krimpen-Stoop, E. M. L. A., and Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*(4), 327–345. doi:10.1177/01466219922031446

Vermunt, J. K., and Magidson, J. (2013). *Latent gold 5.0 upgrade manual*. 7th ed. Statistical Innovations Inc. Belmont, MA.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*(1), 65–72. doi:10.1037/h0032060

Wall, M. M., Park, J. Y., and Moustaki, I. (2015). Irt modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement*, *39*(8), 583–597. doi:10.1177/0146621615588184

Waller, L. A., Turnbull, B. W., and Hardin, J. M. (1995). Obtaining distribution functions by numerical inversion of characteristic functions with applications. *The American Statistician*, *49*(4), 346–350. doi:10.1080/00031305.1995.10476180

Warwick, D. P., and Lininger, C. A. (1975). *The sample survey: Theory and practice*. New York: McGraw-Hill.

Weiss, D. J. (1973). The stratified adaptive computerized ability test. Research Report No. 73-3. University of Minnesota, Department of Psychology. Minneapolis, MN.

Wise, S. L., and Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wolf, L. F., and Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242. doi:10.1207/s15324818ame0803_3

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. doi:10.1007/s10862-005-9004-7

Woods, C. M. (2008). Monte carlo evaluation of two-level logistic regression for assessing person fit. *Multivariate Behavioral Research*, *43*(1), 50–76. doi:10.1080/00273170701836679

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., and Stone, M. H. (1979). *Best test design. rasch measurement.* Chicago: MESA Press.

Zuckerman, M., Kuhlman, D. M., Joireman, J., Teta, P., and Kraft, M. (1993). A comparison of three structural models for personality: The big three, the big five, and the alternative five. *Journal of Personality and Social Psychology*, *65*(4), 757–768. doi:10.1037/0022-3514.65.4.757

Appendix A

Tables

A.1 Reviewed Person-Fit Indices

Table A.1: Categories of popular person-fit indices and their feasibility under non-parametric (descriptive) and binary-logistic model approaches (under IRT terminology: Rasch model, two-parameter-logistic, and three-parameter-logistic model)

Label	Author
Non-parametric (descriptive)	
G	Guttman (1944, 1950)
$G^*/U1$	van Der Flier (1977)
r_{pbis}, r_{bis}	Donlon and Fischer (1968)
C	Sato (1975)
$U3$	van Der Flier (1980), Meijer (1994)
A_i, D_i, E_i	Kane and Brennan (1980)
MCI_I	Harnisch and Linn (1981)
$ZU3$	van Der Flier (1982)
NCI_i, ICI_i	K. K. Tatsuoka and Tatsuoka (1983)
$H1_i^T$	Sijtsma (1986), Sijtsma and Meijer (1992)
Rasch model	
U	Wright and Stone (1979)
W	Wright and Masters (1982)
UB, UW	R. M. Smith (1985)
M	Molenaar and Hoijtink (1990)
χ^2_{SC}	Klauer and Rettig (1990)
$T(X)$	Klauer (1991, 1995)
2PLM and 3PLM	
l_0	Levine and Rubin (1979)
D	Weiss (1973), Trabin and Weiss (1983)
ECI statistics	K. K. Tatsuoka (1984)
l_z	Drasgow, Levine, and Williams (1985)
$JK, O/E$	Drasgow, Levine, and McLaughlin (1987)
l_{zm}	Drasgow, Levine, and McLaughlin (1991)
c	Levine and Drasgow (1988)
GRM	
l_{poly}	Drasgow, Levine, and Williams (1985)

A.2 Empirical Results with translated Percentages

Table A.2: Number of sub-sample members of the experimental study sample identified as extreme values in $\Upsilon_i(\mathcal{N}, \Sigma)$, where parameters for $\Upsilon_i(\mathcal{N}, \Sigma)$ are estimated based on different samples (see corresponding Table 5.3)

Sub-sample	Cell(s)	Evaluation scenario			
		JpH	C14pH	C14r10%pH	HpH
Plausible	1	5	5	5	6
	4	7	7	7	6
Semi-plausible	2	17	20	17	15
	5	18	19	17	16
Implausible	3	25	24	24	21
	6	33	27	28	26
Plausible	1,4	12	12	12	12
Semi-plausible	2,5	35	39	34	31
Implausible	3,6	58	52	52	48
Semi-/implausible	2,3,5,6	93	90	85	77

[Note] Number of response patterns' value smaller than cut-off value in respective scenario.

[JpH] Online questionnaire sample.

[HpH] Experimental study sample.

[C14pH] Plausible response sub-sample of the experimental study.

[C14r10%pH] Plausible response sub-sample and a randomly drawn small portion of invalid responses.

A.3 Comparing Parameter Estimates of several Models

Table A.3: Parameters estimates and standard errors for different samples based on different models

Estimate	Sub-sample (cells)							Estimate
	1 & 4	1 & 4 & 2 & 5 & 3 & 6			Online questionnaire data			
	Latent classes							
	$c = 1$	$c = 1$	$c = 2$		$c = 1$	$c = 2$		
	$z = 0$	$z = 0$	$z = 1$	$z = 0$	$z = 0$	$z = 0$	$z = 1$	
				$3.11\ (0.03)^{**}$			3.09^{**}	$\hat{\mu}_{w,1}$
				$3.13\ (0.03)^{**}$			3.18^{**}	$\hat{\mu}_{w,2}$
$\hat{\lambda}_{1,1}$	$1.01\ (0.11)^{**}$	$0.80\ (0.07)^{**}$	$1.07\ (0.10)^{**}$		0.87^{**}	0.88^{**}		
.	$0.72\ (0.11)^{**}$	$0.74\ (0.07)^{**}$	$0.94\ (0.10)^{**}$	$-0.21\ (0.04)^{**}$	0.90^{**}	0.93^{**}	-0.31^{**}	$\hat{\lambda}_{w,1}$
.	$0.69\ (0.13)^{**}$	$0.22\ (0.08)^{**}$	$0.51\ (0.11)^{**}$		0.72^{**}	0.74^{**}		
.	$0.80\ (0.11)^{**}$	$0.57\ (0.07)^{**}$	$1.03\ (0.11)^{**}$		0.70^{**}	0.72^{**}		
.	$0.68\ (0.12)^{**}$	$0.75\ (0.07)^{**}$	$0.74\ (0.08)^{**}$	$-0.17\ (0.04)^{**}$	0.79^{**}	0.81^{**}	-0.32^{**}	$\hat{\lambda}_{w,2}$
$\hat{\lambda}_{1,6}$	$0.78\ (0.10)^{**}$	$0.67\ (0.07)^{**}$	$0.96\ (0.09)^{**}$		0.67^{**}	0.69^{**}		
$\hat{\lambda}_{2,7}$	$0.64\ (0.13)^{**}$	$0.43\ (0.09)^{**}$	$1.19\ (0.11)^{**}$		0.87^{**}	0.89^{**}		
.	$1.04\ (0.11)^{**}$	$0.64\ (0.07)^{**}$	$0.34\ (0.12)^{**}$	$-0.21\ (0.04)^{**}$	1.10^{**}	1.15^{**}	-0.31^{**}	$\hat{\lambda}_{w,1}$
.	$0.39\ (0.12)^{**}$	$0.37\ (0.07)^{**}$	$0.23\ (0.12)^{*}$		0.61^{**}	0.62^{**}		
.	$0.76\ (0.13)^{**}$	$0.44\ (0.09)^{**}$	$1.11\ (0.10)^{**}$		1.08^{**}	1.11^{**}		
.	$0.78\ (0.12)^{**}$	$0.41\ (0.08)^{**}$	$0.88\ (0.11)^{**}$	$-0.17\ (0.04)^{**}$	1.01^{**}	1.02^{**}	-0.32^{**}	$\hat{\lambda}_{w,2}$
$\hat{\lambda}_{2,12}$	$1.01\ (0.11)^{**}$	$0.95\ (0.08)^{**}$	$0.37\ (0.11)^{**}$		1.03^{**}	1.08^{**}		
$\hat{\lambda}_{3,13}$	$0.85\ (0.11)^{**}$	$0.64\ (0.07)^{**}$	$0.38\ (0.09)^{**}$		0.65^{**}	0.63^{**}		
.	$0.18\ (0.08)^{*}$	$0.28\ (0.07)^{**}$	$0.11\ (0.11)$	$-0.21\ (0.04)^{**}$	0.38^{**}	0.37^{**}	-0.31^{**}	$\hat{\lambda}_{w,1}$
.	$0.42\ (0.09)^{**}$	$0.70\ (0.06)^{**}$	$0.35\ (0.06)^{**}$		0.39^{**}	0.36^{**}		
.	$0.88\ (0.10)^{**}$	$0.80\ (0.06)^{**}$	$0.50\ (0.07)^{**}$		0.77^{**}	0.75^{**}		
.	$0.63\ (0.09)^{**}$	$0.68\ (0.06)^{**}$	$0.45\ (0.06)^{**}$	$-0.17\ (0.04)^{**}$	0.62^{**}	0.58^{**}	-0.32^{**}	$\hat{\lambda}_{w,2}$
$\hat{\lambda}_{3,18}$	$0.79\ (0.10)^{**}$	$0.71\ (0.06)^{**}$	$0.38\ (0.06)^{**}$		0.62^{**}	0.61^{**}		
$\hat{\phi}_{2,1}$	$-0.15\ (0.11)^{**}$	$-0.47\ (0.06)^{**}$	$-0.07\ (0.10)$		-0.24^{**}	-0.26^{**}		
$\hat{\phi}_{3,1}$	$-0.16\ (0.11)^{**}$	$-0.42\ (0.07)^{**}$	$-0.31\ (0.09)^{**}$		-0.17^{**}	-0.19^{**}		
$\hat{\phi}_{3,2}$	$0.12\ (0.11)^{**}$	$0.44\ (0.06)^{**}$	$0.00\ (0.11)$		0.02^{*}	0.03^{**}		
				$0.60\ (0.03)^{**}$			0.10^{**}	$1 - \hat{\eta}_0$

▨ Indicating affiliation with recoded observed variables $^{()}$ Unreported standard errors are ≤ 0.2

A.4 Simulation Results with translated Percentages

Table A.4: Numbers of simulated responses identified as extreme values translated from percentages reported in Table 7.4 based on corresponding sub-sample sizes of respective simulation conditions (simulation study evaluation scenario: estimated parameters and estimated percentiles)

Perc. error var.		Percentage valid responders									Invalid study type
		90%			70%			50%			
		T	I	V	T	I	V	T	I	V	
75%	5th	60	1	63	80	3	77	90	5	90	Item wording
	95th	50	30	27	70	60	7	70	65	0	
	Test	120	**30**	**90**	150	**63**	**84**	160	**70**	**90**	
50%	5th	60	12	45	60	30	35	60	40	20	
	95th	50	16	36	60	39	14	60	50	10	
	Test	110	**28**	**81**	120	**69**	**49**	120	**90**	**30**	
25%	5th	80	56	27	140	132	7	150	150	0	
	95th	60	6	54	60	18	42	60	30	25	
	Test	140	**62**	**72**	190	**150**	**42**	210	**185**	**25**	
75%	5th	60	18	45	110	0	105	170	0	165	Long string
	95th	60	12	45	90	84	0	90	85	0	
	Test	120	**29**	**90**	200	**87**	**112**	250	**90**	**165**	
50%	5th	80	45	27	80	60	21	90	20	75	
	95th	60	4	54	60	12	49	60	60	5	
	Test	140	**49**	**90**	140	**72**	**70**	160	**80**	**80**	
25%	5th	90	81	9	180	174	0	110	110	0	
	95th	70	0	72	90	0	84	40	5	35	
	Test	160	**81**	**81**	260	**177**	**84**	160	**120**	**35**	

[T,I,V] T = total sample, I = invalid group, I = valid group.

[?th P.] Left/ right-sided cut-off based on 5th/ 95th percentile.

[Test] Combined/ two-sided test.

A.5 Notations

Table A.5: Globally used symbols directory table

Symbol	Elements	Description
Matrices		
S	$\{s^2_{j,j}\}$	Sample covariances
Σ	$\{\sigma_{j,j}\}$	Model implied covariances
Λ	$\{\lambda_{j,k}\}$	Factor loadings
Φ	$\{\phi_{m,k}\}$	Factor covariances
Ψ	$\{\psi_{j,j}\}$	Error variances
Vectors		
$\boldsymbol{y_i}$	$(y_{i,1},\ldots,y_{i,k},\ldots,y_{i,q})$	Latent variables
$\boldsymbol{\nu}$	$(\nu_1,\ldots,\nu_k,\ldots,\nu_q)$	Factor means
$\boldsymbol{x_i}$	$(x_{i,1},\ldots,x_{i,j},\ldots,x_{i,p})$	Observed responses
$\bar{\boldsymbol{x}}$	$(\bar{x}_1,\ldots,\bar{x}_j,\ldots,\bar{x}_p)$	Manifest variable means
$\boldsymbol{\mu}$	$(\mu_1,\ldots,\mu_j,\ldots,\mu_p)$	Expectations for manifest variables
$\boldsymbol{\epsilon}$	$(\epsilon_1,\ldots,\epsilon_j,\ldots,\epsilon_p)$	Error terms
$\boldsymbol{\delta_i}$	$(\delta_{i,1},\ldots,\delta_{i,j},\ldots,\delta_{i,p})$	Differences $(\boldsymbol{x_i}-\boldsymbol{\mu})$
$\boldsymbol{z}$	$(z_i,\ldots,z_n)$	Class membership
$\boldsymbol{\eta}$	$(\eta_0,\eta_1,\ldots,\eta_z,\ldots,\eta_{c-1})$	Probabilities for class membership
$\boldsymbol{\theta}$	$(\theta_1,\ldots,\theta_j,\ldots,\theta_p)$	Parameters
Functions		
g, g_j	$Pr(\boldsymbol{x}\vert\cdot), Pr(x_j\vert\cdot)$	Cond. distr. of $\boldsymbol{x}$ or x_j (given $\cdot$)
Auxiliary		
$\hat{\square}$		Estimated value
$\square^T$		Transpose of a matrix
$\square^{-1}$		Inverse of a matrix
$\square^{(z)}$		Parameters for group z